The role of PNPLA3 in the development and progression of chronic liver disease

William Albert McCabe

University College London

This report is submitted to the degree of Doctor of Philosophy

March 05, 2020

Declaration of authorship

I, William Albert McCabe confirm that the work presented in this report is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed …… Date ……

Two roads diverged in a yellow wood,

And sorry I could not travel both

And be one traveller, long I stood

And looked down one as far as I could

To where it bent in the undergrowth;

Then took the other, as just as fair,

And having perhaps the better claim,

Because it was grassy and wanted wear;

Though as for that the passing there

Had worn them really about the same,

And both that morning equally lay

In leaves no step had trodden black.

Oh, I kept the first for another day!

Yet knowing how way leads on to way,

I doubted if I should ever come back.

I shall be telling this with a sigh

Somewhere ages and ages hence:

Two roads diverged in a wood, and I—

I took the one less travelled by,

And that has made all the difference.

Robert Frost

Acknowledgements

First, I would like to express my gratitude to my supervisors for their invaluable advice and support throughout this project. Dr Morgan, I thank you for sharing your vast knowledge of liver disease, your friendship, support and the freedom to explore and grow as a researcher; I truly learned so much from your guidance and outstanding example. Dr Coker, thank you for sharing your unwavering passion for crystallography, technical guidance and an ever-critical eye to spot every opportunity to improve my research.

My gratitude extends to all my colleagues and students from the X-ray Crystallography laboratory, who helped me throughout this work too numerous to mention and to all the other researchers who offered their valuable advice. A special thanks to Dr De Simone and Dr Sanz Hernandez from Imperial College London, whose insight into molecular modelling helped to shape some of the seminal findings of this work.

I am of course eternally grateful to the financial support received from the Rosetrees Trust and The Royal Free Charity; it is your generous aid which made this research physically possible.

I would also like to thank all my family and dear friends who have provided me with encouragement and support for many years, even when I strayed far from the right path.

Most of all I am grateful to my love Marika, who taught me what it means to be happy. Thank you for being my unwavering support and my strength through the great ups and downs that come from research. You are the guiding light that drives me every single day and my inspiration to dream big and to work hard. Without you there is no doubt I would never have made it this far. Marika, this thesis is dedicated to you.

Abstract

Chronic liver disease is now of great international concern due to rapidly increasing morbidity and mortality associated with the disease. There is significant evidence that carriage of the patatin-like phospholipase domain containing protein 3 (*PNPLA3)* risk allele, rs738409:G, plays a key role in determining risk for the development of chronic liver disease from a variety of causes. rs738409 is a common single nucleotide polymorphism which results in substitution of an isoleucine residue for methionine at position 148 of PNPLA3 (Ile148Met; I148M). However, the physiological role of PNPLA3 and the functionality of its I148M variant, are currently largely unknown. The central aim of this thesis was to investigate the biological function of PNPLA3 and elucidate the complex role that the I148M variant plays in the development and progression of liver disease. Investigation into the primary sequence of PNPLA3 was undertaken to characterise the protein and inform latter experimental design. Phylogenetic investigation revealed human PNPLA5 to share the highest homology with PNPLA3, and revealed more distant, previously undescribed relationships with the bacterial protein ExoU. A combination of expression trials and subsequent *in vitro* investigation into the behaviour of PNPLA3 was attempted. Despite attempts with numerous constructs, PNPLA3 remained unstable when expressed using an *E. coli* expression system and was not able to be produced in sufficient quantity to facilitate structural analysis. In the latter half of the thesis, both variants of PNPLA3 are investigated through *in silico* structural modelling and subsequent molecular dynamic simulation. The first simulation of full-length PNPLA3 is reported, revealing a more detailed description of the domain architecture of PNPLA3 and the local impact of the I148M variation. A novel disease mechanism is proposed, in which methionine at residue 148 effects the conformational stability of the PNPLA3 active site, resulting in a loss of lipase activity.

Impact statement

Context:

Liver disease is currently of great international concern due to the massive burden of disease and rapidly increasing mortality rates attributable to the disease. As the global levels of obesity and alcohol consumption, the two major environmental causes of the disease, are also on the rise, this trend is likely to continue. However, studies have shown that liver disease also has a large genetic proponent; in particular, the PNPLA3 I148M variant, has been shown to have an astounding correlation with increased risk of hepatic steatosis, non-alcoholic steatohepatitis, advanced fibrosis, cirrhosis, non-alcoholic liver disease, alcoholic liver disease, chronic hepatitis C related cirrhosis and hepatocellular carcinoma. These associations have been replicated over many studies, and indeed are one of the most well replicated genetic associations in clinical research. Despite this, investigations into the underlying mechanism of action and biochemistry of the protein have remained inconclusive. In this research, for the first time we have been able to model the full length PNPLA3 protein and used this model as a foundation to perform the most in-depth molecular dynamic simulations of the protein variants to date. This has allowed a leap in the understanding of the PNPLA3 structure and formed a hypothesis linking the I148M variant to the pathogenesis of the disease.

Impact on public health:

The most critical impact of this research is the potential for the results to be used to assist in the development of a treatment for liver disease on a broad international scale. Due to the significant association with the development and progression of liver disease, the PNPLA3 protein is of great interest as a drug target for the development of compounds which may prevent the progression of chronic liver damage in individuals at risk. In order to facilitate drug discovery processes and further assess the viability of the protein as a drug candidate, a 3-dimensional protein structure is often required. To date, the PNPLA3 protein has proved to be elusive to *in vitro* characterisation despite a number of studies on the protein. By successfully modelling a structure of PNPLA3, this research offers a small yet valuable first step in facilitating the evaluation of the protein as a drug target. This will both inform further expression protocols based on a better understanding of the domain architecture of the protein and could even be used directly as an early guide for *in silico* assessment of ligands and drug compounds.

Impact within academia:

There are many interesting proteins highlighted in genetic association studies which are not amenable to protein production and purification. An additional impact of this research is to provide an example of using a bioinformatic approach to enhance the understanding of biochemical data and the mechanism of genetic diseases even when the information regarding the protein is limited. As more examples of this approach are shared within the community, it is hoped this can encourage more frequent *in silico* exploration of proteins which are limited by experimental incompatibility.

Contents

Page

Contributions

General:

The design and conceptualization of this project was undertaken with Dr Marsha Morgan and Dr Alun Coker.

Chapter 3:

The 96 PNPLA3 expression vectors were created and initial small-scale expression trials performed by Michael way, in collaboration with staff at the Oxford Protein Production facility, Rebecca Boys and Shu-Fen Coker, Division of Medicine, University College London.

The pCold expression vector containing an mPNPLA3 insert was generously donated by Institute of Molecular Biosciences, University of Graz.

Mass spectroscopy was undertaken with assistance from Dr Graham Taylor and Dr Nigel Rendell, Centre for Amyloidosis, University College London.

The in-vitro activity assays of PNPLA3 were performed in collaboration with Rebecca Jeyaraj, Division of Medicine, University College London.

Chapter 4:

Model 9 was generated using modified I-TASSER scripts in collaboration with Maximo Sanz Hernandez, Department of Life Sciences, Imperial College London.

List of tables

Page

List of figures

Page

List of abbreviations and symbols

ADPN: adiponutrin
AhR: aryl hydrocarbon receptor
AMP-K: adenosine monophosphate activated kinase
AP: alkaline phosphatase
ALD: alcoholic liver disease
ALT: alanine aminotransferase
AmpR: ampicillin resistance gene
AST: aspartate aminotransferase
ATGL: adipose triglyceride lipase
ATP: adenosine triphosphate
B-factor: temperature factor
BEL: bromo-enol lactone
BLAST: basic local alignment search tool
BLOSUM: blocks substitution matrix
BMI: body mass index
C-score: confidence score
CASP: critical assessment of techniques for protein structure prediction
CD: circular dichroism
CDART: conserved domain architecture retrieval tool
CDD: conserved domain database
CPLA2: cytosolic phospholipase A2
ChREBP: carbohydrate regulatory binding protein
cryo-EM: cryo-electron microscopy
DAG: diacylglycerol
DEEPCNF: deep convolutional neural fields
DGAT: diacylglycerol acyltransferase
DGGR: 1,2-o-dilauryl-rac-glycero-glutaric acid-(6'methylresorufin) ester
EELD: endogenous ethanol liver disease
ER: endoplasmic reticulum
EP: expert pool
ETC: electron transport chain
G0S2: G0/G1 Switch Gene 2
GGT: γ-glutamyltransferase
GLP-1: glucagon like peptide 1
GMQE: global model quality estimation
GPU: graphics processing unit
GS2: gene sequence 2
GS2-like: Gene sequence 2 like
GWAS: genome wide association study

HCC: hepatocellular carcinoma
hg19: human genome build 19
HKII: hexokinase II
HMM: hidden Markov model
HSC: hepatic stellate cell
I148M: Isoleucine for methionine at amino acid position 148
IMAC: immobilised metal-ion affinity chromatography
iPLA$_2$γ: calcium independent phospholipase 2 gamma
JNK: c-jun terminal kinase
KC: Kupffer cell
LDAH: lipid droplet associated hydrolase
LFT: liver function test
LIPC: hepatic lipase
LPA: lysophosphatidic acid
LPL: lipoprotein lipase
LPS: Lipopolysaccharide
MAG: monoacylglycerol
MAPK: mitogen activated protein kinase activity
MBP: maltose binding protein
MELD: model of end stage liver disease
MM: molecular mechanics
MyD88: myeloid differentiation factor 88
NAFLD: non-alcoholic fatty liver disease
NASH: non-alcoholic steatohepatitis
NFY: nuclear transcription factor Y
NMR: nuclear magnetic resonance
NRE: neuropathy target esterase related esterase
NTE: neuropathy target esterase
OPPF-UK: Oxford protein production facility UK
P13K: phosphatidylinositol-3 kinase
PA: phosphatidic acid
PASH: PNPLA3-associated steatohepatitis
PDF: probability density function
PLA2G6: phospholipase 2 group iv
PNPLA: patatin-like phospholipase domain containing protein
PNPLA3: human patatin like phospholipase domain containing protein 3
pnpla3: murine patatin like phospholipase domain containing protein 3
PROSESS: protein Structure Evaluation Suite and Server
PSI-BLAST: position specific iterative basic local alignment search tool
PPAR: peroxisome proliferator activated receptors
PPARγ: Peroxisome proliferator-activated receptor-γ

QM: quantum mechanics
QM/MM: quantum mechanics/ molecular mechanics hybrid
RF: random forest
RMSD: root mean square deviations
RMSF: root mean square fluctuations
ROS: reactive oxygen species
SCAP: SREBP-cleavage activating protein
SDS-PAGE: sodium dodecyl sulfate poly acrylamide gel electrophoresis
SEC: size exclusion chromatography
SFA: saturated fatty acid
SMART: simple modular architecture research tool
SMTL: SWISS-MODEL template library
SNP: single nucleotide polymorphism
SOCS3: suppressor of cytokine signalling 3
SPARCLE: subfamily protein architecture labelling engine
SREBP: sterol regulatory element binding proteins
STAT3: signal transducer and activator of transcription 3
T3: tri-iodothyronine
TAG: triacylglycerol
TBST: tris-buffered saline and tween-20
TCEP: tris-(2-carboxyethyl)-phosphine
TG: triglycerol
TF: Trigger factor
TIRAP: toll/IL-1 receptor domain containing adaptor protein
TLR4: toll-like receptor 4
TM-Score: template modelling score
TROSY: transverse relaxation optimised spectroscopy
UCNP: up converting nanoparticles
UPR: unfolded protein response
UK: United Kingdom
vLCPUFA: very-long-chain polyunsaturated fatty acids
VLDL: very low-density lipoproteins
VMD: visual molecular dynamics
WESA: weighted ensemble solvent accessibility predictor
XBP1: x-box binding protein 1

Chapter 1

Introduction to PNPLA3: The role in chronic liver disease

"seize this very minute–

What you can do, or dream you can, begin it,

Boldness has genius, power, and magic in it,

Only engage, and then the mind grows heated—

Begin it, and the work will be completed!"

John Anster

(Inspired by Johann Wolfgang von Goethe)

1.1 Overview

Liver disease is currently of great international concern, not only because of the huge burden of disease and disability, but also increasing death rates attributable to the disease; indeed in 2014, the standardised mortality rate attributable to liver disease was five times greater in the United Kingdom (UK) than in 1970.[1]

Like many of the major diseases, liver disease is a classic example of a complex polygenic disorder, involving the interaction of a number of environmental and genetic factors. Teasing apart these factors is extremely challenging and so the pathogenesis of the disease remains poorly understood, and treatment options for liver disease remain limited.[2]

A wide range of genetic association studies have attempted to elucidate genetic variations which may contribute to the development and pathogenesis of the disease. Across these studies, the association between PNPLA3 and liver injury has been consistently and robustly replicated in a large number of populations from different ethnicities and geographical populations.[3]

These associations have been mostly attributed to the single nucleotide polymorphism rs738409 which leads to the substitution of Isoleucine for methionine at amino acid position 148 (I148M).[4,5]

Despite the clear importance of the I148M variant of PNPLA3 within liver disease, the role of the protein both within the context of the healthy liver and the disease state remains poorly understood; with no clearly defined role in cellular metabolism, and no key structural information.[6,7]

In this opening chapter, a background on the function of the liver, the current understanding of the pathogenesis of liver disease and a review of the current state of the art on PNPLA3 will be presented. This will act as a foundation from which to further evaluate the potential impact of the I148M variant on the normal function of liver.

1.2 The structure and function of the liver

The liver is the largest organ in the human body, weighing between 1200-1500g and making up one fiftieth of the average adult body weight. It is one of the most complex organs within the body and performs a vast number of vital functions required to maintain adequate homeostasis within the organism (Figure 1.1).[8]

In total, the liver is currently believed to play a role in over 500 biological functions within the body. These include: the synthesis of many essential biomolecules; the extraction and metabolism of various nutrients and toxins both ingested and released by apoptotic cells; the storage and excretion of metabolic products; the neutralisation of numerous foreign antigens and toxins from the gut.[9]

In order to carry out these diverse functional roles the liver has evolved into a structurally complex, multicellular tissue with a unique angioarchitecture.

1.2.1 The structure of the liver

The liver is primarily composed of four lobes: the left lobe, right lobe, quadrate lobe and caudate lobe. Each lobe within the liver can be further dissected into multiple hepatic lobules, which form hexagonal plates of hepatocytes between six portal triads around a central vein. These lobules comprise of the functional subunits of the liver (Figure 1.2).

The liver receives blood supply through connection with both the hepatic portal vein and hepatic artery; whereby around 25% of total liver blood flow is derived from the hepatic artery, and 75% from the portal vein. The blood circulates through the liver via afferent and efferent blood vessels of all sizes, uniformly distributed throughout the tissue, connected almost exclusively by the smallest capillary sized vessels, the sinusoids. The blood vessels, together with connective tissue form the major skeleton for the liver, causing it's soft and spongy character.[10]

The blood circulates from the portal triad, along each live sinusoid, and empties through the central vein and into the vena cava. Bile canaliculi are distributed through the tissue, which drain into bile ducts around the lobule perimeter (Figure 1.3).

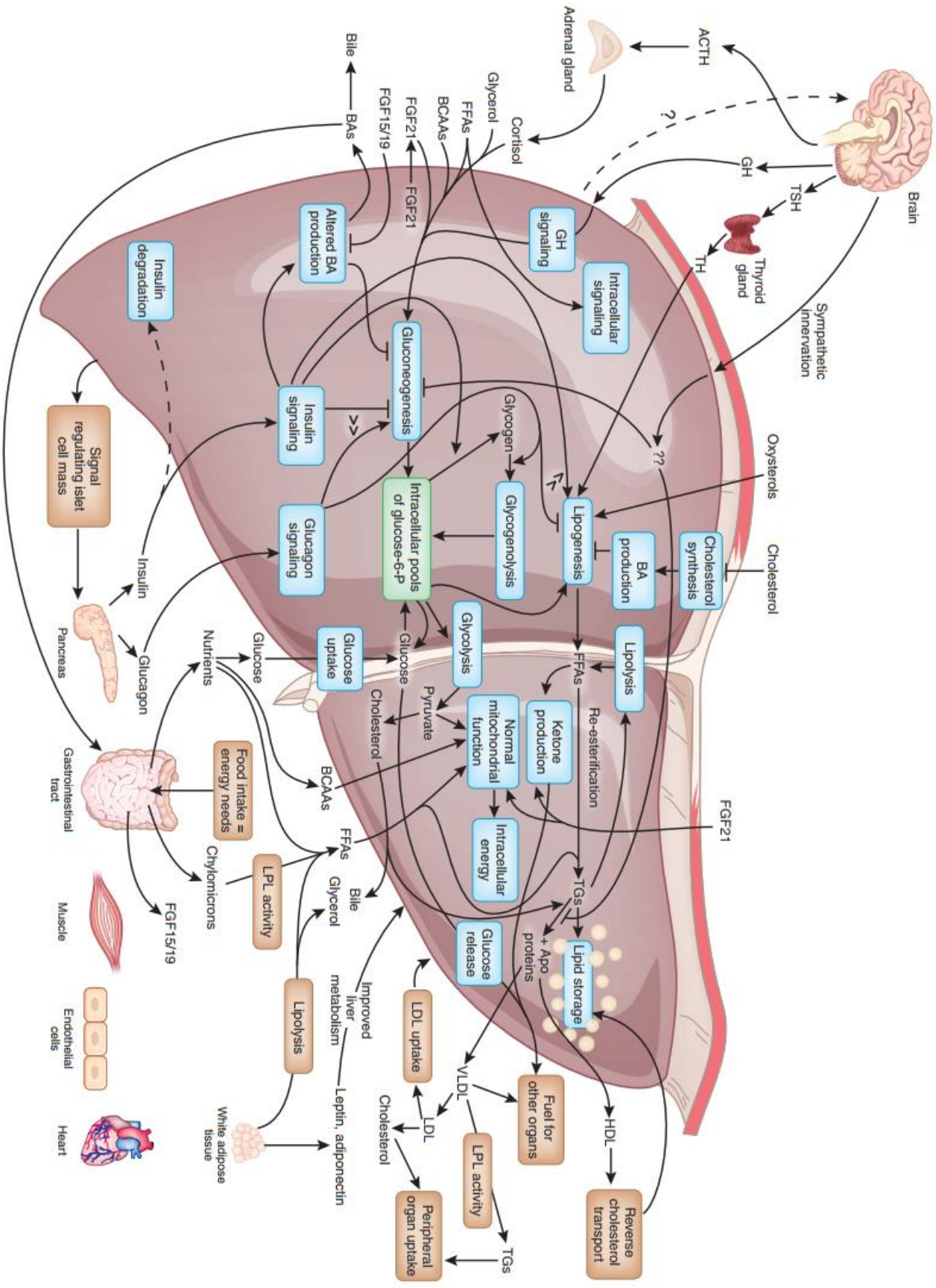
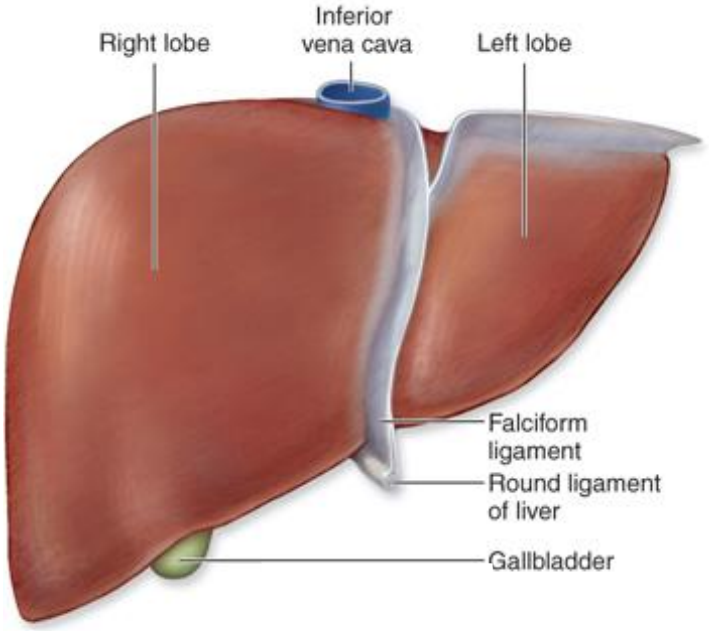


Figure 1.1 Complex hepatic metabolism at homeostasis

(Adapted from Nature).[11–21]

A)



B)

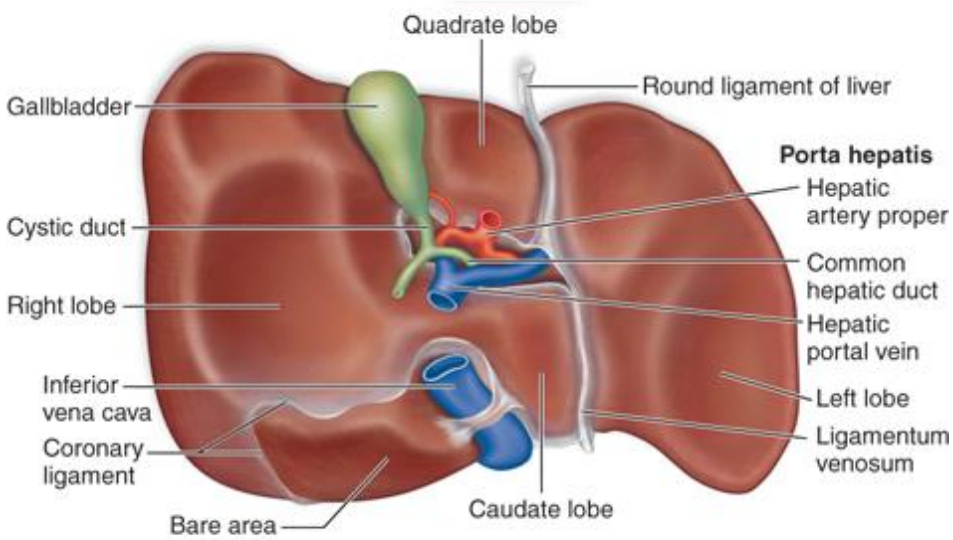


Figure 1.2 External structural view of the liver

A) Anterior view of the liver.
B) Posterior view of the liver (Adapted from Reisner 2014).[22]

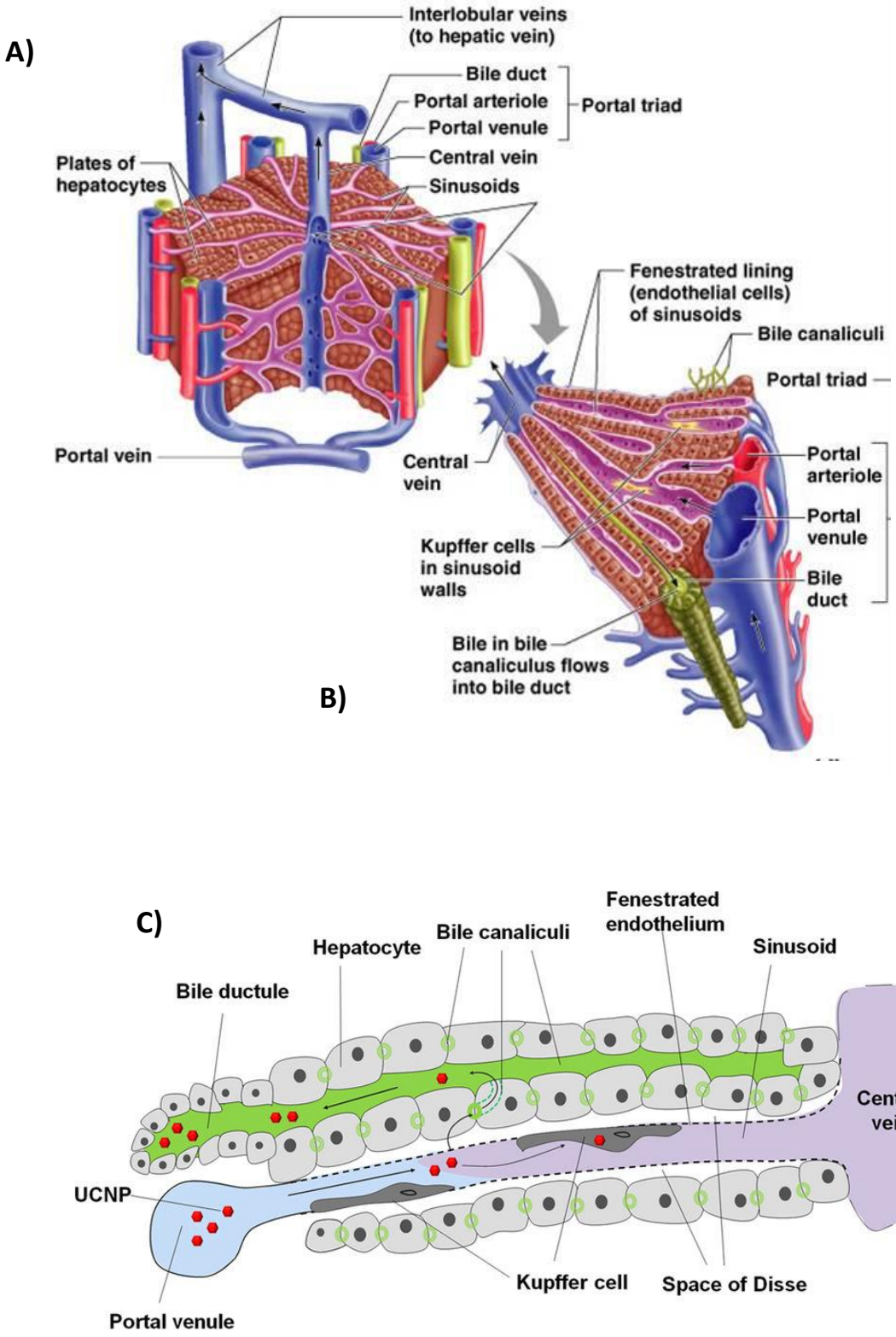


Figure 1.3 Three-dimensional representation of parenchymal liver tissue composition

A) Hexagonal architecture of liver lobule.
B) Zoomed section of the lobule (Adapted from austincc.edu).[23]
C) Schematic representation of cellular composition of a portal lobule segment. UCNP: Up converting nanoparticles (Adapted from Seo et al. 2015).[24]

1.2.2 Cellular composition of the liver

The liver comprises of a network of highly specialised cell types, which can function as an integrated community through signalling networks of cytokines and chemokines to perform the multiplicity of hepatic functions. The predominant cells present in the liver are hepatocytes, hepatic stellate cells, endothelial cells, Kupffer cells, pit cells and cholangiocytes.[8,9]

1.2.2.1 Hepatocytes

The main parenchymal tissue of the liver is made up of between 70-85% hepatocytes. These cells are thought to be responsible for the majority of the metabolic roles of the liver. To serve their highly metabolic roles, they are equipped with large quantities of mitochondria, Golgi complexes and endoplasmic reticulum.[25]

Hepatocytes form distinct cell surfaces which have different roles based on the orientation of the cell. Part of the hepatocyte will face sinusoids and form microvilli to extend into the space of Disse, while part of the membrane is in contact with other hepatocytes, other than in regions which are withdrawn to form bile canaliculi.[26]

Each surface of the cell is highly specialised, with canalicular facing membrane optimised for bile exertion, sinusoidal surface optimised for extraction and secretion of molecules into the blood and hepatocyte facing regions optimised for communication with gap junctions.[26]

1.2.2.2 Endothelial cells

The hepatic endothelial cells make up 3% of the hepatic mass and are located at the periphery of the sinusoidal space. Endothelial cells form flattened plates with fenestrations along the surface, which form sieve like plates.[25]

The unique interaction between the specialised endothelial cells and the complex molecular lining of the sinusoidal space allows for free escape of fluid components of the blood.[27]

The endothelial cells have a range of scavenger receptors on the cell surface and provide the main route for clearance of effete molecules and colloids from circulation.[28]

1.2.2.3 Hepatic stellate Cells

Hepatic stellate cells comprise of around 1.5% hepatic mass and are located in the space of Disse. Hepatic stellate cells can undertake several different functional roles. They play a key role with hepatocytes in the storage of vitamin A as retinyl esters and can synthesise or degrade elements of the perisinusoidal extracellular matrix.

After cytokine signalling Hepatic stellate cells develop a myofibroblast phenotype, producing collagens and laminin, and are believed to play an important role in the fibrotic response to liver injury.[29]

1.2.2.4 Kupffer cells

Kupffer cells are a type of specialised macrophage, which make up 2% of hepatic mass. They are located in the lumen of sinusoids, attached loosely to the endothelial cells.[28]

The Kupffer cells are phagocytic and together with the endothelial cells form a major system for clearing effete cells and molecules from circulation.

When activated, a number of chemokines and cytokines are released which help to coordinate the livers acute phase reaction to injury. They can become activated by a large range of agents including endotoxins, shock interferons, tumour necrosis factor and arachidonic acid.[30]

1.2.2.5 Pit cells

Pit cells are liver specific lymphocytes which while attached to the sinusoidal surface of the endothelium remain mobile. Pit cells differentiate in the sinusoid from circulating large granular lymphocytes.[31]

Pit cells are cytotoxic to virus infected and tumorous hepatocytes, and therefore involved in virus elimination as well as the development of antigen tolerance.[32]

1.2.2.6 Cholangiocytes

Cholangiocytes make up a minor portion of the liver and locate to bile ducts and portal tracts. The luminal face consists of microvilli, which respond to bile flow signalling. Despite being few in number, the cholangiocytes can modify the water and solute composition of bile.[33]

1.3 Liver disease

Liver disease is a description for a broad spectrum of diseases which all result in liver injury, which can either be acute or chronic in nature. The most prevalent types of liver disease in the western world are alcoholic liver disease (ALD) and non-alcoholic fatty liver disease (NAFLD). However, there are lower yet still significant incidence of other types of liver disease such as hepatocellular carcinoma and viral hepatitis.

Liver disease is currently of great international concern, not only because of the huge burden of disease and disability, but also because of increasing death rates attributable to the disease worldwide. In 2014, five times more people died of liver disease in the United Kingdom (UK) than in 1970. This rise in deaths from liver disease comes at a time when the death rate of all other major diseases in the UK are falling (Figure 1.4).[1]

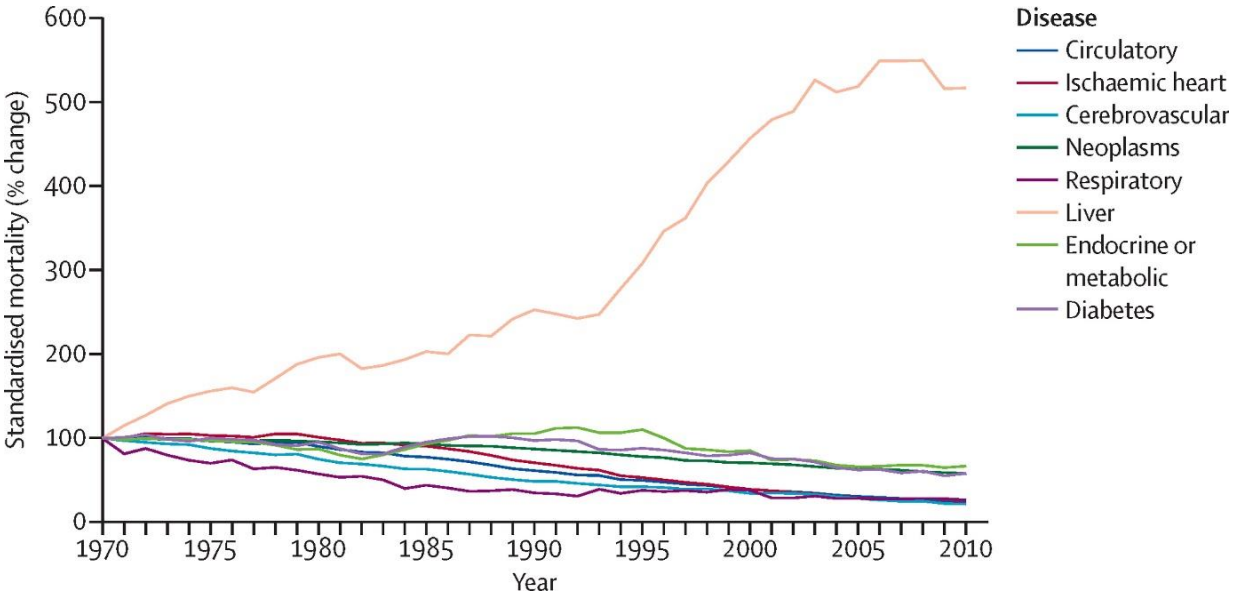


Figure 1.4 The standardised death rates caused by common diseases in the United Kingdom between 1970 and 2010

Liver disease is highlighted in orange showing a significant increase in death rate (Adapted from Williams *et al.* 2014).[1]

This trend is mirrored in the US,[34] where liver related deaths have increased by over 30% in the last 15 years alone,[35] leading to over 3.9 million people now diagnosed with the disease (Figure 1.5).[36]

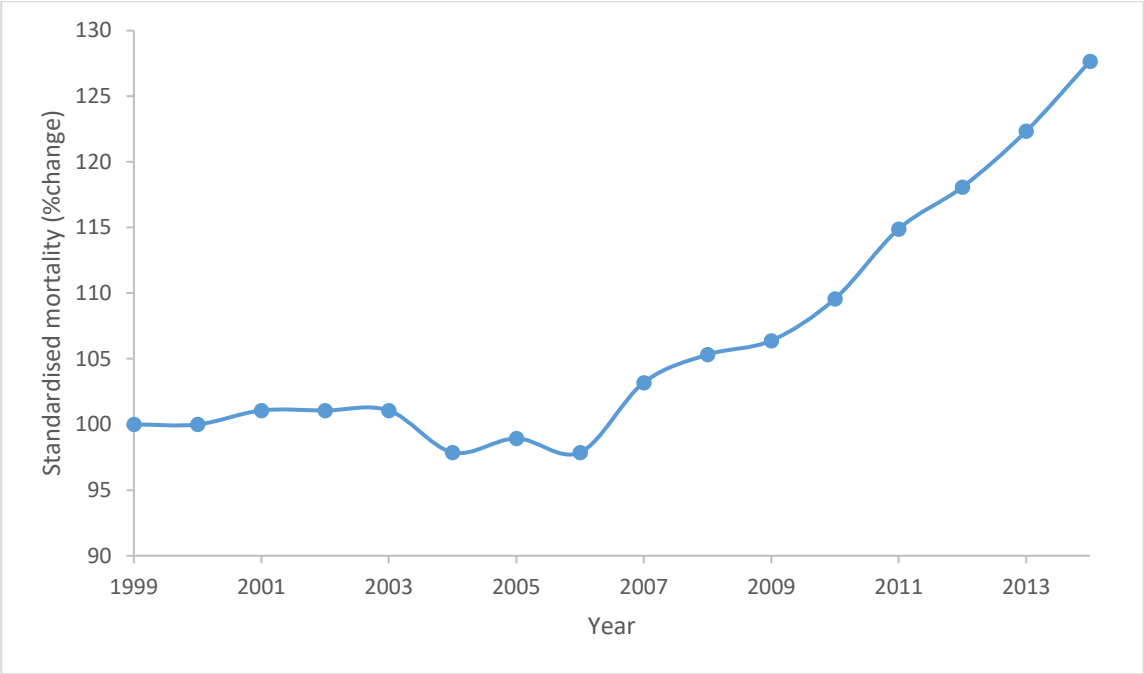


Figure 1.5 The standardised death rate caused by liver disease in the United States of America between 1999 and 2014

Liver disease is highlighted in blue (Adapted from Kochanek *et al.* 2016).[35]

Like many other major diseases, liver disease is a classic example of a complex polygenic disorder, involving the interaction of a number of environmental and genetic factors. Teasing apart these factors is extremely challenging and so the pathogenesis of the disease remains poorly understood, and treatment options for liver disease remain limited.[2]

The two most significant risk factors known to influence the development of liver disease are alcohol and obesity. Currently around 75 percent of liver related deaths are linked to excessive alcohol consumption in the UK. However, with obesity levels in the western world rising, and around 25 percent of people in the UK alone now classified as obese, non-alcoholic fatty liver disease is an ever growing concern.[1]

The best treatment for non-alcoholic steatohepatitis (NASH) or ALD remains abstinence from alcohol and improved diet and exercise. There are several repurposed drugs which can be prescribed in the case of early liver disease presentations, however none of these drugs directly target the cause of disease; rather interfering with the resulting symptoms to mediate the liver injury (Figure 1.6). These medications are not specific to liver disease and all come with serious risks and side effects implicit in general drug treatments, highlighting the need for better treatment options. Once cirrhosis of the liver begins there is no effective drug treatment and continued drinking will likely lead to total loss of liver function.[2]

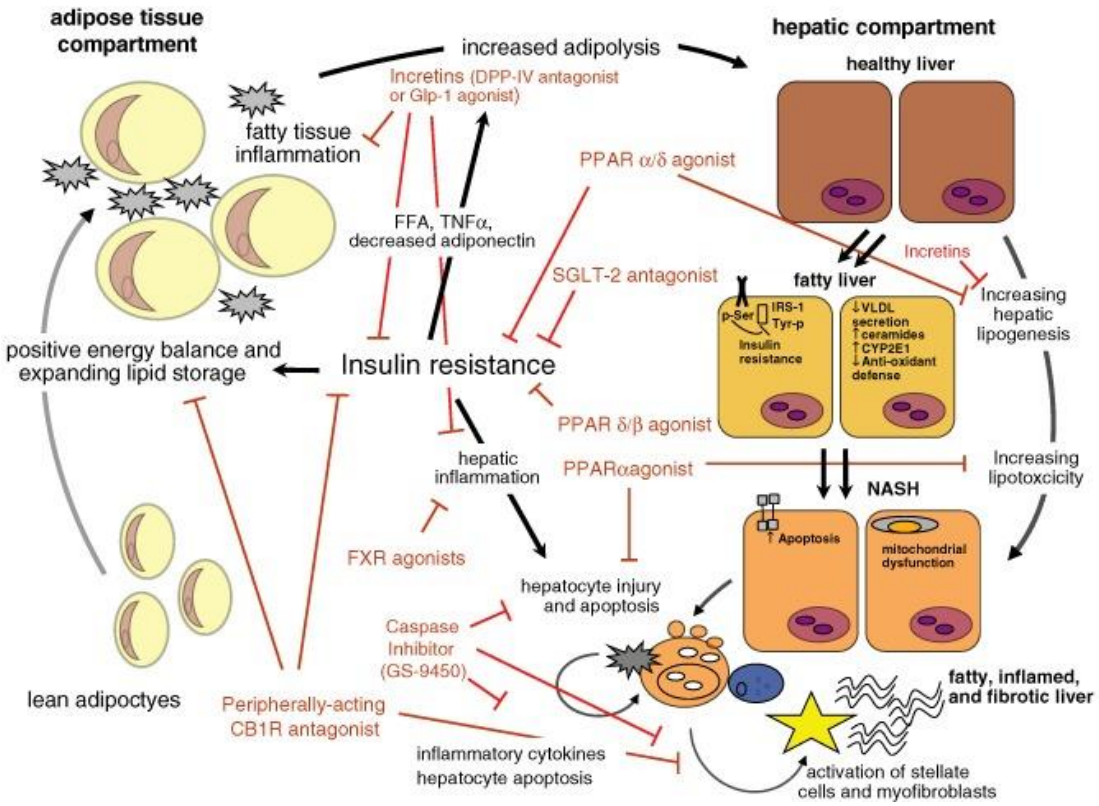


Figure 1.6 Diagrammatic representation of the cellular progression of liver disease

Cellular processes thought to contribute to the next stage are represented on black arrows, and potential therapeutic strategies to each point are represented in red (Adapted from Schuppan & Schattenberg, 2013).[37]

Due to the vast complexity of the liver, to date there remains no way to artificially compensate for loss of liver function. Some short-term assistance can be offered through artificial extracorporeal liver support. For example, the molecular adsorbent recirculation system, which is primarily used to aid in the removal of both water soluble and albumin-bound toxins, which would otherwise accumulate to cause encephalopathy and dysfunction of other organs. However, even with these techniques, long term survival is not possible and the only long term solution to treat complete liver failure remains orthotopic liver transplantation.[38] As the impact of extracorporeal liver support on survival of acute liver injury is minimal, mortality rates of patients on transplant lists remain high, even with liver support.[39,40]

1.3.1 Histopathology of liver disease

While alcoholic and non-alcoholic liver diseases have different aetiologies, they are both diseases which progress through a similar histopathological morphologic spectrum.[41]

This spectrum includes: (i) steatosis, characterised by the accumulation of lipid droplets in the cytosol of hepatocytes; (ii) steatohepatitis, characterised by further inflammation and cell death; (iii) cirrhosis, characterised by disruption of the cellular architecture by fibrous bands; and (iv) hepatocellular carcinoma, a cancer which often develops within the cirrhotic liver (Figure 1.7).

1.3.1.1 Steatosis

The lipid content of a normal healthy liver is ~0.5-1.5%. Alcohol consumption and obesity result in the deposition of lipids in the liver. In patients with severe steatosis the fat content of the liver may be as high as 50%. The majority of individuals that consume more than 60 g of alcohol daily will develop steatosis[42,43] and steatosis is even more prevalent in individuals misusing alcohol who are also obese.[44] Steatosis reverses rapidly when alcohol consumption is reduced. However, the presence of steatosis impacts the progression to more severe liver disease.[45]

1.3.1.2 Hepatosteatosis

In a subset of individuals, steatosis may also be associated with inflammation. This is identified as steatohepatitis, which is characterised by: ballooning degeneration, necrosis and apoptosis of hepatocytes in association with neutrophilic infiltration, cholestasis and the formation of Mallory bodies.

Ballooning degeneration occurs when lipids, proteins and water are retained within the hepatocytes. Mallory bodies are proteinaceous deposits, which may be histologically evident in the damaged hepatocytes as red material. The recurrent generation and resolution of inflammation leads to the accumulation of fibrous tissue and, over time, this may result in the development of cirrhosis.

1.3.1.3 Cirrhosis

Cirrhosis is defined histologically by characteristic features including the disruption of the liver's normal architecture, its replacement with fibrous bands of collagen, and the presence of regenerative nodules. At a histological level, these features are graded into three types based on the size and homogeneity of the nodules: micronodular, macronodular and mixed-nodular cirrhosis.

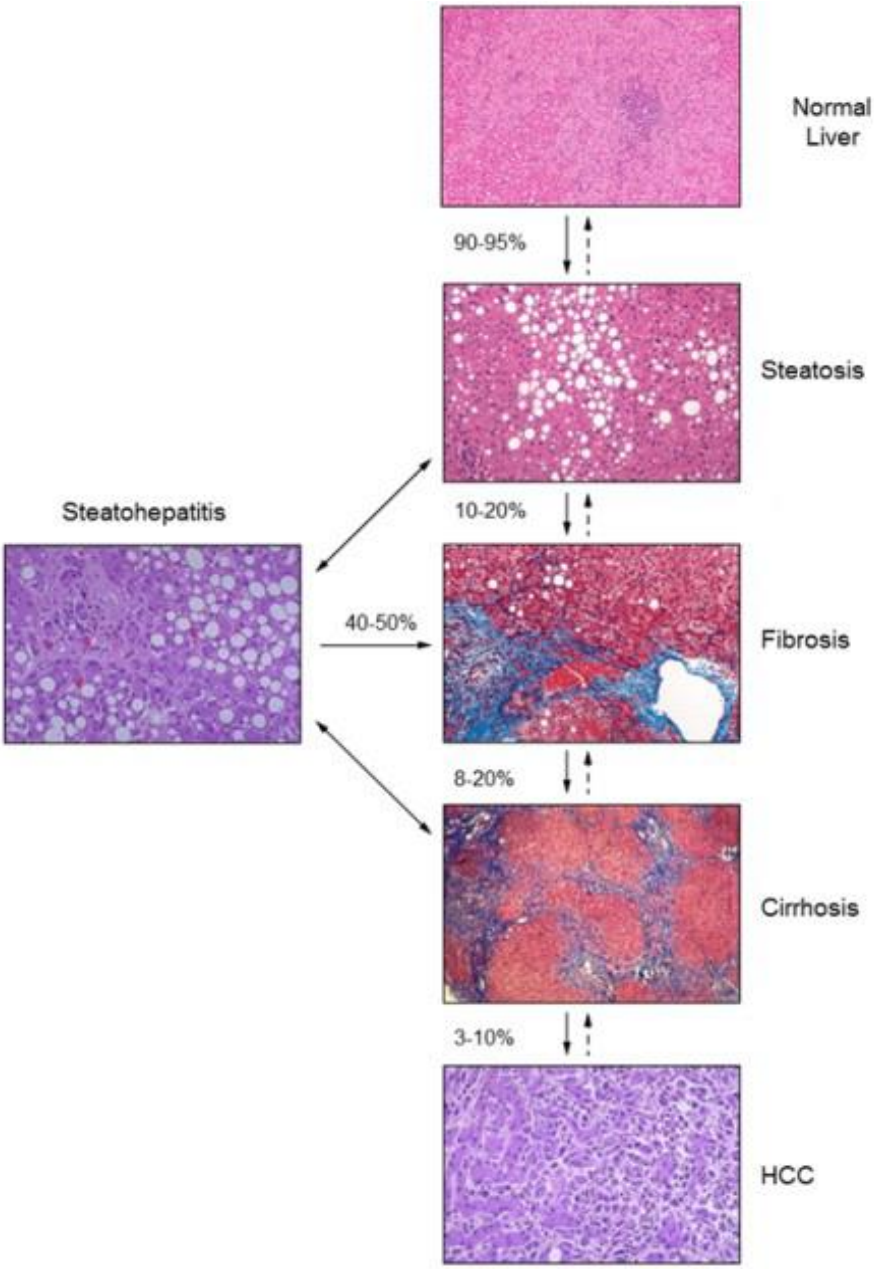


Figure 1.7 The stages of alcohol-related liver disease. Each image represents typical microscopic features seen on biopsy

The arrows represent an estimate of the number of individuals that will progress through each stage of alcohol-relate liver disease to more advanced forms of the disease. (Adapted from Mathurin et al. 2012).[46]

1.3.1.4 Hepatocellular carcinoma

The development of hepatocellular carcinoma (HCC) is considered part of the natural history of alcohol-related liver injury. As such it only develops in patients with long-standing alcohol-related cirrhosis and in this respect differs from other liver disease such as that caused by hepatitis B/C and non-alcohol-related fatty liver disease (NAFLD) where tumours may arise in

non-cirrhotic livers. The tumours arise in discrete nodules most likely around dysplastic rests; where most tumours are likely to be multifocal.[47]

1.3.2 Histological differences between alcoholic and non-alcoholic liver disease

Because alcoholic liver disease (ALD) and non-alcoholic fatty liver disease (NAFLD) both share similar morphological presentations differentiating between the disorders without alcohol consumption data can be challenging. This is made only more challenging by prevalent overlapping clinical features caused by high levels of alcohol use and obesity in western societies. Despite this there are key differences between the diseases beyond their pathogenesis.

In general, alcoholic liver disease will display a more severe disease histology at the time of the biopsy and some features are more rarely, or never observed in the NAFLD disease spectrum (Table 1.1).[41]

Table 1.1 Differences in histological features between non-alcoholic and alcoholic liver disease

NAFLD	ALD
Usually mild	Varying severity
Bridging necrosis rare	Bridging necrosis common
Poorly formed Mallory hyaline	Well-formed Mallory hyaline
Very rare	Sclerosing hyaline necrosis
Very rare	Phlebosclerosis
Very rare	Canalicular cholestasis
Not described	Foamy degeneration
Nuclear vacuolation – more common	Nuclear vacuolation – less common
Presence of iron/ hemosiderin less frequent	Presence of iron / hemosiderin more frequent
Ductular reaction less frequent/prominent	Ductular reaction more frequent/prominent
Fibrosis/cirrhosis less common	Fibrosis/cirrhosis more common

(Adapted from Sakhuja 2014).[41]

Alcoholic foamy degeneration for example, in which there is a diffuse prominent micro vesicular fatty change (may be more in perivenular zone) with minimal inflammation or Mallory-Denk bodies, has not been described in any cases of NAFLD.[48]

Sclerosing hyaline necrosis, characterized by perivenular liver cell necrosis with fibrosis in the same region, is extremely rare in NAFLD and may result in occlusion of the terminal hepatic venules and pre-cirrhotic portal hypertension.[49]

There are also three types of vascular lesions which rarely occur in NAFLD. These include lymphocytic phlebitis, phlebosclerosis (narrowing of the hepatic vein lumen) and veno-occlusive lesions. Whereas phlebosclerosis occurs in all cases of ASH.[50]

1.3.3 Physical symptoms of liver disease

The early stages of liver disease are generally considered asymptomatic, having few clinically evident manifestations. The disease can initially manifest as weight loss, fatigue, and osteoporosis as a result of vitamin D malabsorption followed by calcium deficiencies, and early diagnosis is often based on investigation based on related risk factors.

When liver injury is detected and symptoms present, this is generally after the liver is significantly injured and there is a severe decline in liver function; as this occurs, a large number of physical symptoms may be observed (Table 1.2).[51]

Table 1.2 Common physical findings in patients with late stage liver disease

Physical symptoms
Abdominal wall vascular collaterals (caput medusa)
Ascites
Asterixis
Clubbing and hypertrophic osteoarthropathy
Constitutional symptoms, including anorexia, fatigue, weakness, and weight loss
Cruveilhier-Baumgarten murmur—a venous hum in patients with portal hypertension
Dupuytren's contracture
Fetor hepaticus—a sweet, pungent breath odour
Gynecomastia
Hepatomegaly

Jaundice
Kayser-Fleischer ring—brown-green ring of copper deposit around the cornea, pathognomonic for Wilson's disease
Muehrcke's nails: paired horizontal white bands separated by normal colour
Terry's nails: proximal two thirds of nail plate appear white, whereas the distal one third is red
Palmar erythema
Scleral icterus
Vascular spiders (spider telangiectasias, spider angiomata)
Splenomegaly
Testicular atrophy

(Adapted from Sleisenger *et al.* 2010).[51]

In the decompensated liver, patients may develop jaundice, pruritus, gastrointestinal bleeding, coagulopathy, and heightened sensitivity and medication toxicity due to the loss of basic liver function. When there is sufficient physical obstruction due to secondary cirrhosis, additional symptoms may occur such as ascites, hepatic encephalopathy, variceal bleeding and oedema.[52]

1.3.4 Biomarkers of liver disease

Since the early stages of liver disease are mostly asymptomatic, detection is usually based on results of liver function tests (LFTs), or because patients present with other alcohol related problems.

LFT's examine an array of biochemical markers which may appear abnormal in the presence of liver injury. They are non-specific, and many factors can impact the levels of these biomarkers, however, they are frequently used as a key diagnostic tool in identifying potential liver damage.

The markers measured in LFTs fall into one of three categories: (i) Markers of hepatocellular injury; (ii) markers of cholestasis; and (iii) indicators of liver function.

Markers of hepatocellular injury: The main markers which indicate hepatocellular injury are circulating levels of aspartate aminotransferase (AST) and alanine aminotransferase (ALT), which are primarily expressed in hepatocytes and are released during cell death. These markers are also present in skeletal muscle and can be released during muscle strain.

Markers of cholestasis: The main markers of cholestasis are alkaline phosphatase (AP) and γ-glutamyltransferase (GGT). AP levels can rise during pregnancy or bone injury, while GGT levels can be elevated by both alcohol consumption and aromatic medications.

Indicators of liver function: The main markers Indicating impaired liver function are bilirubin, albumin and prothrombin time.[53] Bilirubin results from enzymatic breakdown of heme, which is normally rapidly transported into bile. Once the liver has lost approximately half of its excretory capacity it begins to be filtered into urine, but this is elevated by some genetic diseases. Albumin synthesised in the liver can provide a marker of liver synthetic capability, however with a long half-life, this is slow to be detected, and levels can be decreased in a huge range of conditions including burns, trauma, sepsis, malnutrition and pregnancy. Prothrombin time is abnormal once around 80% of liver synthetic function becomes lost and provides a good test of severe liver damage however it is also heavily influenced by vitamin K deficiency.

Since all of these biomarkers can have other causes and natural fluctuations, they are not an ideal marker of liver function. Further to this while some analysis of the liver function can be made it is not possible to determine the type of liver disease with confidence.[54]

Complex calculations have been devised which take into account the range of liver function results to provide the best possible data on liver function to prioritise patient treatment and surgery. Two of the most used examples are the Child-Pugh score [55] and more recently model of end stage liver disease (MELD) score;[56] However these still remain limited and a liver biopsy is the only way to definitively test the stage of liver damage present. For this reason, any additional markers of the disease which could increase the accuracy of non-invasion diagnoses are highly desirable.

1.3.5 Pathogenesis of non-alcoholic fatty liver disease

While both ALD and NAFLD pathologies result from complex interactions between the host genetics and numerous environmental factors to mediate the ultimate liver injury, the largest distinction between them is in the pathogenesis of the disease.

While the exact mechanisms are not understood, ALD is primarily driven by alcohol as the causative agent, while NAFLD is more heavily based on other dietary and genetic factors.

1.3.5.1 Two-hit hypothesis

Traditionally, a "two-hit" hypothesis has been used to describe the pathogenesis of NAFLD, in which the first hit was the accumulation of lipids within the liver caused by insulin resistance, a high fat diet and obesity. This increase in lipid content then weakens the liver to the second hit, which is an increase in inflammatory cascades mainly caused by reactive oxygen species through the lipotoxicity of peroxidised triglycerides; which then leads to hepatitis and more severe liver injury.[57]

The two-hit hypothesis provides a simple explanation for the development of NALFD. However, increasing evidence, including the observation of cases of inflammation preceding the steatosis within the liver, has confirmed that this does not adequately describe the pathogenesis of the disease.[58] It is now accepted that a more complex multiple hit hypothesis better represents the true pathogenic model of disease (Figure 1.6).

The multiple hit hypothesis attempts to connect a large range of potential factors which can trigger the development and progression of liver disease whereby the hits do not necessarily occur in any set sequence and can combine in numerous ways; each contributing different relative effect sizes. This means steatosis can both proceed and further cause inflammation, generating a feedback loop which exacerbates the disease. The specific mechanisms and relative effects of each mechanism are still not well understood, with novel mechanisms regularly being reported.

1.3.5.2 Steatosis preceding inflammation

Steatosis in response to dietary factors, obesity and insulin resistance was previously the first hit in the two-hit hypothesis. This is a well catalogued phenomenon, which can occur via a range of mechanisms. Namely; a direct excess lipid supply from diet and adipose lipolysis; an increase in de novo lipogenesis; a decrease in export of very low density lipoprotein-triglyceride; or reduced β-oxidation of free fatty acids.[59]

The most common cause of steatosis is due to a surplus of dietary carbohydrates. This surplus leads to the increase in expression of lipogenic transcription factors SREBP and ChREBP, which in turn lead to the upregulation of lipogenic genes in the cell; as well as being substrates for de novo lipogenesis themselves. This is a way for the body to store excess calories and to protect from excess free fatty acids as opposed to a dysfunction, and the amount of carbohydrate will directly influence the amount of de novo lipogenesis which occurs in the liver.

Since simple sugars are converted into fatty acids more readily than complex carbohydrates, these place increased risk of steatosis[60] and diets which are enriched for both saturated fat and simple sugars carry the highest risk.[61]

Another mechanism which may lead to steatosis is the change in cell type within the liver which occurs during obesity. Specifically, there is hepatic recruitment of a myeloid cell population which may be responsible for further promoting lipid storage. While the importance of this mechanism is unclear it does provide an additional link between obesity and the observed steatosis.[62]

Finally, the impact of other less prevalent micronutrients and the interplay between the gut microbiota may also have a crucial role in steatosis. For example, dietary choline is required for very low-density lipoprotein synthesis export and diets which are deficient in choline can lead to increased hepatic steatosis. There are a range of intestinal microbiota, which are able to convert choline into methylamine and overrepresentation of these microbiota in the gut can lead to deficiency in choline as well as acting as an inflammatory agent.[63]

1.3.5.3 Inflammation preceding steatosis

Although previously not thought to be a major proponent of the pathogenesis of the disease, inflammation which precedes any steatosis has become a large area of research.

The expression of two key proinflammatory cytokines, IL-6 and TNFα, are increased in obese patients and patients with insulin resistance. This is thought to occur due to adipose tissue becoming inflamed due to hypoxia and death of rapidly expanding adipocytes. This directly links obesity and insulin resistance to potential inflammation, and the enhanced expression of these cytokines decreases after weight loss, showing how closely the factors are linked.[64]

Free fatty acids and cholesterol lead to an increase in reactive oxygen species within the cell, which is particularly damaging within the mitochondria and again an increase in TNFα. While this is predominantly a concern for the steatotic liver, both lipids can occur in a healthy liver acting as an early inflammatory hit.[65]

Lipopolysaccharides (LPS), integral parts of gram-negative bacteria outer membranes, act as an immunomodulator and are affected by changes in our diet and gut microbiota. Intake of a high-fat or high-carbohydrate diet in humans over only 3 days, can lead to a significant increase in circulating LPS concentrations. This increase in LPS could lead to increased systemic inflammation and play a role in observed hepatitis.[66]

There are several other less well understood pathways which may be able to interfere with normal liver metabolism and induce inflammation prior to steatosis based on changes in gut microbiota. Some examples of this are the loss of short chain fatty acid production from dietary fibre sources and dietary triggering of the aryl hydrocarbon receptor (AhR); further highlighting how the intestinal microbiota and its products might directly regulate host gene expression and affect systemic inflammation.[67,68]

1.3.5.4 Steatosis due to inflammation

In addition to steatosis caused directly from external factors, some degree of steatosis can also be triggered by inflammation within the liver itself and systemic inflammation.

Serum levels of proinflammatory cytokines IL-6 and TNF- α correlate remarkably well with the presence of insulin resistance, and adipose tissue-derived TNF-α and IL-6 have been shown to regulate hepatic insulin resistance via upregulation of suppressor of cytokine signalling 3 (SOCS3). When combined with rapid expansion and death of adipocytes, this has been shown to accentuate fat loss from adipocytes and promote ectopic fat accumulation. TNF- α also leads to an upregulation of SREBP, further influencing steatosis.[69]

X-box binding protein 1 (XBP1) has long been known as a key regulator of the unfolded protein response (UPR) secondary to ER stress and inflammatory signals. This has recently been shown to also trigger hepatic lipogenesis, leading to increased steatosis.[70]

Finally, the aryl hydrocarbon receptor (AhR) is a ligand-activated transcription factor sensing xenotoxicants such as dioxin. The activation of the AhR has also been shown to trigger lipid accumulation within hepatocytes.[71]

1.3.5.5 Inflammation due to steatosis

The impact of steatosis leading to inflammation has been well explored as the second hit in the original two hit hypothesis of NAFLD.[72] This is believed to be primarily through lipotoxicity caused by an increase in fatty acids, leading to an increase in peroxidised triglycerides and reactive oxygen species.

This may in part lead to mitochondrial dysfunction and reduced enzymatic activities of mitochondrial electron transport chain (ETC) complexes, leading to further generation of

reactive oxygen species (ROS) as a result of ETC leakage during mitochondrial β-oxidation in energy production.[73]

Furthermore, long-chain saturated fatty acids (SFAs) such as palmitate and stearate are transported to mitochondria for β-oxidation or esterified for either excretion in the form of VLDL (very low density lipoproteins) or storage as lipid droplets.[59] However, overloading of the mitochondria with these species can lead to the activation of a number of intracellular responses, such as JNK1 and a mitochondrial death pathway.

While lipotoxicity provides a reasonable argument for liver injury, how this lipotoxicity is produced remains unclear. Increased peroxidation of triglycerides would certainly produce increased oxidative stress for the cell, however triglycerides are generally more inert within the cell and it is possible that triglyceride synthesis is in fact protective against free fatty acids which are known to be more highly lipotoxic.[74]

Finally, toll-like receptor 4 (TLR4) activates proinflammatory signalling in response to excessive SFAs. This pathway is initiated by toll/IL-1 receptor domain containing adaptor protein (TIRAP) and myeloid differentiation factor 88 (MyD88) that ultimately leads to activation of nuclear factor κB with production of TNF-α .[75] This pathway can also be activated in response to LPS, and contribute to inflammation preceding steatosis.[76]

1.3.5.6 Link between NASH and cirrhosis

The primary theory linking NASH to cirrhosis is the deposition of fibrotic tissue in direct response to liver injury, which can be induced through several different mechanisms.

Mitochondrial dysfunction is triggered by lipotoxicity, which in turn contributes to a lack of ATP within the cell and increased endoplasmic reticulum (ER) stress, by which dysfunction of the ER from reactive oxygen species or lack of adenosine triphosphate (ATP) leads to a build-up of unfolded protein. Prolonged activation of this unfolded protein response leads to activation of c-jun terminal kinase (JNK) which can activate inflammation and apoptosis.

Furthermore, direct peroxidation of plasma and intracellular membranes may cause cell necrosis/apoptosis and megamitochondria, while ROS-induced expression of Fas-ligand on hepatocytes may induce fratricidal cell death.[59] Carbohydrate feeding, which results in steatosis as well as features of the human metabolic syndrome, also up-regulates Fas expression in hepatocytes resulting in increased sensitivity to Fas-mediated apoptosis and liver injury (Figure 1.8).[77]

The primary cell-type responsible for the deposition of fibrotic tissue is the hepatic stellate cell (HSC). Free cholesterol can lead to liver injury through the activation of intracellular signalling pathways in Kupffer cells (KCs), hepatic stellate cells (HSCs), and hepatocytes. The activation of KCs and HSCs in particular promotes inflammation and fibrogenesis. In general, liver injury promotes the transition of HSCs to a myofibroblast-like state, where collagen synthesis is upregulated and, matrix degradation is downregulated resulting in fibrosis.

Further to this, adipose tissue dysfunction also releases a range of pro-inflammatory cytokines as well as hormones such as leptin which can activate hepatic stellate cells through the hedgehog or mTOR pathways. Several other changes also occur following HSC activation including HSC proliferation and the increased migration of HSCs to the site of liver injury.[78]

1.3.5.7 Endogenous alcohol theory

More recently, it has been proposed that NAFLD may in fact be an endogenous ethanol liver disease (EELD) whereby ALD and NAFLD would actually have a very similar mechanistic background. In this hypothesis endogenous alcohol production is thought to provide alcohol as a causative agent.

There is significant weight behind this theory, which was hypothesised based on an increase in alcohol in the breath of obese mice, when compared to lean animals. Further to this, an increase in alcohol producing bacteria in the microbiome of NASH compared to healthy individuals was detected. This aligns well with the fact that both diseases are extremely similar in their morphology, and that genetic variants such as PNPLA3 were associated well with both diseases.[79,80]

This hypothesis has been brought into question, because gut microbiota are thought to produce insufficient endogenous alcohol for hepatoxic effects, and a lack of breath alcohol concentration differences between NASH patients and healthy controls.[79] Instead it was suggested the increased alcohol detected previously may in fact result from insulin dependent impairments of alcohol dehydrogenase.[81]

This has been explained by the causative agent in fact being endogenous acetaldehyde, which is a far more potent causative agent of liver damage when produced extrahepatically, and several theoretical predictions suggest potent levels can be reached endogenously, however this needs further research to support these claims.[82]

1.3.5.8 Genetic factors contributing to NAFLD

It is clear from twin and adoption studies that there is a large genetic component to liver disease, and early studies into alcohol related cirrhosis, provided heritability estimates ranging from 21 to 67%.[83,84]

A number of genetic association studies, controlling for addiction, alcohol misuse, and obesity have now identified several common genetic variants, which are likely to play a role in the development and progression of the disease.

The strongest association and most commonly reported, has been found with a variant in *PNPLA3* gene, which is the focus of this thesis and will be discussed in detail later in this chapter. This association has been detected with a huge range of liver related pathologies in a range of cohorts and ethnicities, making it one of the most well documented disease related associations to date.[3]

Associations between liver injury and variants in two additional genes, TM6SF2 and MBOAT7, have more recently been detected, which are less definitive. Currently TM6SF2 is thought to be more closely linked with steatosis and MBOAT7 with fibrosis.[85]

These associations have been confidently replicated cementing a role for these variants within the progression of liver disease, however the precise biological mechanism by which these variants may contribute to the pathogenesis of the disease currently remains unknown.

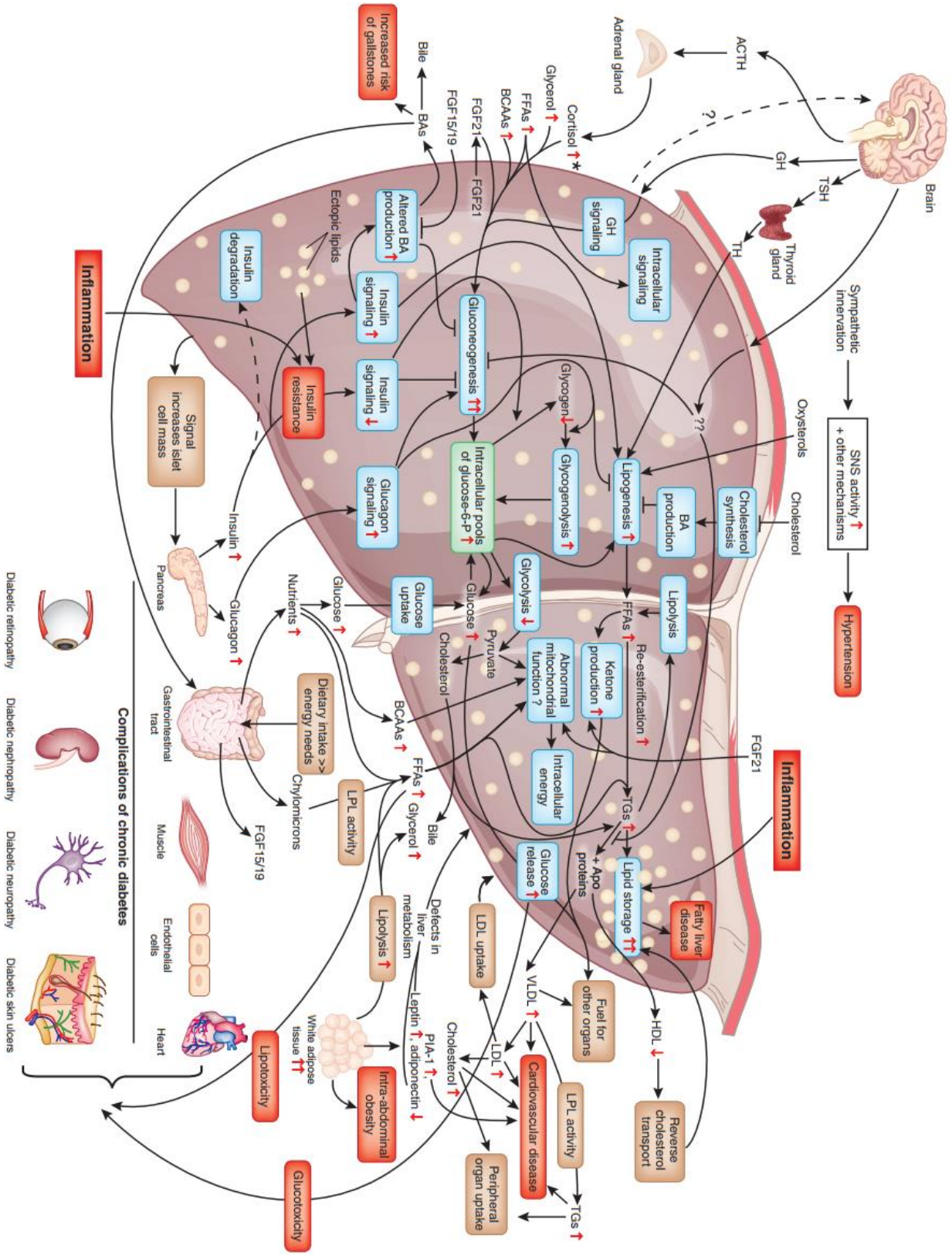


Figure 1.8 Hepatic metabolism with metabolic syndrome

(Adapted from Nature).11–20,86

1.3.6 Pathogenesis of alcoholic liver disease

The key driving factor of ALD is the consumption of large quantities of alcohol and most patients report heavy alcohol use (over 100g/day) for two or more decades.[87] Despite knowing alcohol is the major factor contributing to the pathogenesis of the disease, the biochemical mechanism by which alcohol triggers the disease remains unclear. There are several putative mechanisms which have been identified (Figure 1.9), however many of the factors discussed are based on mouse models of liver disease or cell-based assays; therefore, the relative impact of each factor as well as the true nature of the pathogenesis of the disease is still poorly understood.[88]

Alcohol causes direct damage to hepatocytes through the generation of oxidative stress which as discussed previously is thought to be a key factor liver injury.[89] Similarly, the primary metabolic product of alcohol, acetaldehyde, is highly reactive and can form various protein and DNA adducts. This generates more reactive oxygen species which can lead to lipid peroxidation, and further contribute to oxidative stress and liver injury.[90]

Under severe conditions, the damage from this oxidative stress can lead to hepatocyte death and the release of damage-associated molecular patterns, resulting in sterile inflammation through the production of proinflammatory cytokines, localisation of immune cells to the site of injury and assembly of the inflammasome.[90]

Alcohol has been shown to alter the ratio of NADH to NAD+ which inhibits β-oxidation of fatty acids in mitochondria. Furthermore, this alteration in the redox state, in combination with direct regulation by acetaldehyde and the ER stress response leads to the downregulation of peroxisome proliferator activated receptors (PPAR) and the upregulation of sterol regulatory element binding proteins (SREBP) which results in promotion of lipid accumulation pathways and the inhibition of lipid metabolism, thereby contributing to steatosis.[91]

Further to this, alcohol can inhibit adenosine monophosphate activated kinase (AMP-K), Sirtuin 1, adiponectin and signal transducer and activator of transcription 3 (STAT3), which are all negative regulators of SREBP-1c. This Inhibition results in a further overall increase in SREBP-1c and lipid accumulation.[91]

Alcoholic disrupts the intestinal tight junction integrity, leading to increased gut permeability and higher levels of plasma LPS than healthy control subjects.[76] As described above LPS interacts with toll like receptor 4 (TLR4), leading to proinflammatory cytokines. Alcohol also activates the complement system (C3, C4) which further interacts with Kupffer cells producing more

inflammatory cytokines and chemokines for neutrophil recruitment.[92] In animal models, the loss of TLR-4 results in insensitivity to endotoxin and also alcohol-induced liver injury.[93]

It is likely a complex interaction between other environmental factors and genotype dictate the progression of the disease phenotype, whereby the mechanisms deemed to play a role in NAFLD above, are also able to influence the outcome of ALD.

Presumably, under these conditions a certain level of each of the factors in unique combination are needed to overcome a threshold in which liver damage begins, the inflammatory response becomes more severe and ultimately lead to apoptotic response from cells, activation of HSCs and cirrhosis.

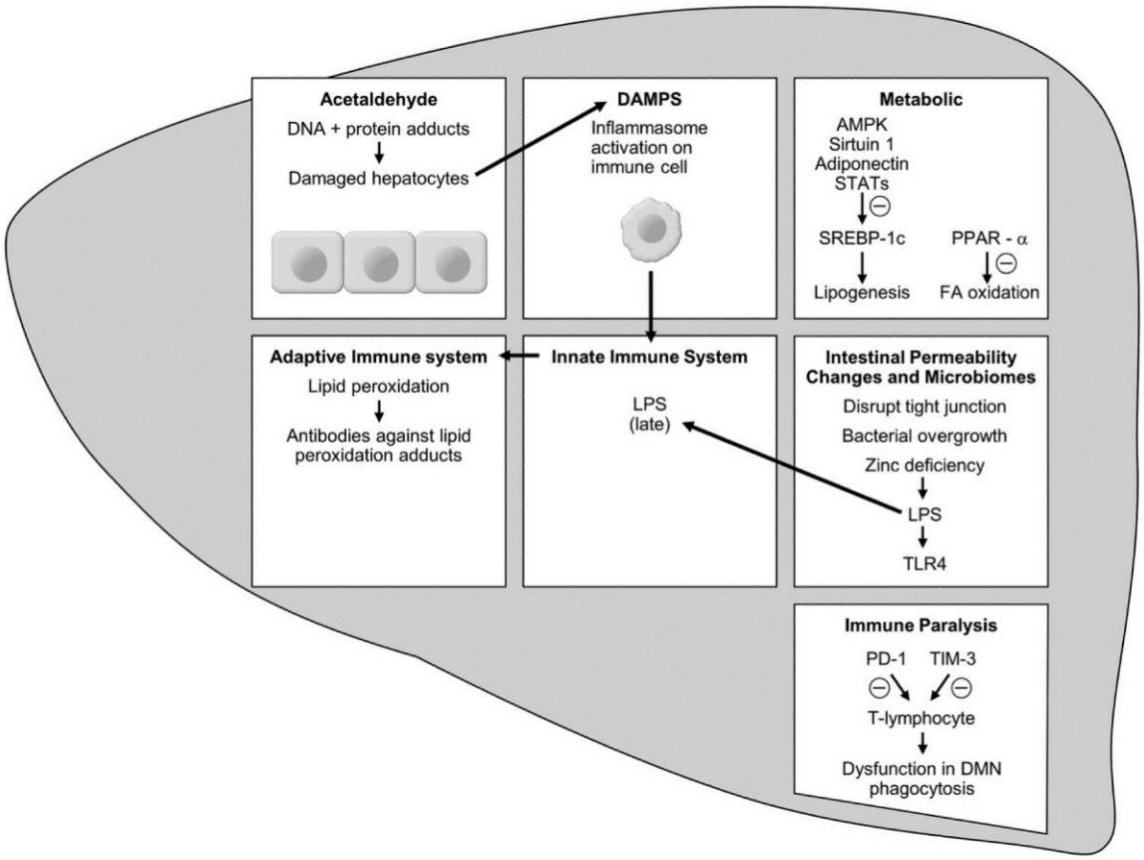


Figure 1.9 Mechanisms of alcoholic liver disease

(Adapted from Dunn and Shah).[88]

1.4 PNPLA3

1.4.1 The discovery of PNPLA3

The patatin like phospholipase domain containing protein 3 (*Pnpla3)* gene was initially discovered through the detection of nutritionally upregulated mRNA products in murine adipocytes.[94] Its expression profile was similar to other genes associated with energy homeostasis, such as leptin, supporting a role for *Pnpla3* within the lipid metabolism.[95] A potential role of the protein in diabetes increased the interest in *pnpla3,* when expression of the gene was observed to be downregulated by Thiazolidinedione compounds, which are used as treatment for type 2 diabetes.[96]

The human homologue to *Pnpla3* was discovered shortly thereafter, and the first report of *PNPLA3* in humans showed that the gene displayed similar expression profile to *pnpla3*.[7,97] However it wasn't until a single nucleotide polymorphism (rs738409) was detected in a GWAS investigating NAFLD, that PNPLA3 became a protein of significant interest.[98]

The association between *PNPLA3* variants and liver injury has been consistently and robustly replicated in many different geographic and ethnic populations (Table 1.3). The rs738409 variant in particular has been associated independently with higher risk not only of severity of steatosis, but all stages of NAFLD and ALD. More recently the same association has been detected with liver injury in Wilsons disease, chronic hepatitis C and hepatocellular carcinoma.[3]

Table 1.3 A number of key studies evaluating association between liver injury and the PNPLA3 I148M variant

Study	Population	N	Key findings
Romeo *et al.* [98]	American-European and African origin, Hispanics	2111	I148M was associated with increase in liver fat in all participants and increased ALT and AST in Hispanics. Genome-wide association study HS was measured by ^{1}H-MRS
Valenti *et al.* [99]	Italy and UK	432/321	Patients with NAFLD I148M were associated with the severity of steatosis and fibrosis and presence of NASH. Hepatic steatosis was diagnosed with LB
Hotta *et al.* [100]	Japanese NAFLD/Controls	253/578	I148M was susceptible to NAFLD. Associated with ALT, AST, ferritin and histological fibrosis stage
Tian *et al.* [101]	Mestizo subjects		The rs7,38,409 in PNPLA3 is strongly associated with alcoholic liver disease and clinically evident alcoholic cirrhosis
Stickel *et al.* [102]	German cohort	1043	The rs7,38,409 in PNPLA3 is associated with alcoholic liver cirrhosis and elevated aminotransferase levels in alcoholic Caucasians

Rotman *et al.* [103]	USA NAFLD	894	Association of I148M with steatosis, ALT and fibrosis. Liver biopsy proved NAFLD
Sookoian and Pirola. [104]	Meta-analysis	16 studies	rs7,38,409 G allele had strong influence on liver fat accumulation plus susceptibility of more aggressive disease. Increase in serum ALT
Valenti *et al.* [105]	Italians	819	rs7,38,409 influenced steatosis in CHC. Independently associated with cirrhosis and HCC
Trepo *et al.* [106]	Caucasian CHC	537	PNPLA3 rs7,38,409 C>G polymorphism favoured steatosis and fibrosis in CHC
Zain *et al.* [107]	Chinese, Indian and Malay	144/198	The G allele was positively correlated with susceptibility to NASH, NASH severity and presence of fibrosis
Burza *et al.* [108]	Swedish obese subjects cohort	4047	Association of the PNPLA3 I148M and HCC in obese individuals who had not undergone bariatric surgery
Vigano *et al.* [109]	CHB patients	235	In CHB patients the PNPLA3 I148M influences susceptibility to steatosis and, in particular, when associated with severe overweight and alcohol intake, severe steatosis
Liu *et al.* [110]	European Caucasians with NAFLD-related HCC	100	PNPLA3 I148M is associated with greater risk of progressive steatohepatitis and fibrosis but also of HCC
Trepo *et al.* [111]	Meta-analysis	2503	rs7,38,409 was strongly associated with HCC in cirrhotic patients, particularly pronounced in ALD than in CHC etiology
Singal *et al.* [112]	Europeans, Asians/NAFLD, ALD, CHC, CHB	9915	PNPLA3 associates to advance fibrosis, NAFLD and ALD
De Nicola *et al.* [113]	Europeans CHC	247	Interaction between homozygosity for the PNPLA3 I148M variant in determining fibrosis progression
Brouwer *et al.* [114]	CHB patients	531	PNPLA3 was independently associated with steatosis, steatohepatitis, lobular inflammation and iron depositions
Stattermayer *et al.* [115]	Wilson disease	98	The PNPLA3 I148M predisposes to increased steatosis development in patients with Wilson disease
Krawczyk *et al.* [116]	Meta-analysis	5100	The I148M PNPLA3 genetic variant is associated with increased risk of developing HCC in NAFLD patients
Luukkonen *et al.* [117]	Finnish cohort	125	Metabolically harmful saturated, ceramide-enriched liver lipidome in “Metabolic NAFLD” but not in “PNPLA3 NAFLD”
Mancina *et al.* [118]	IBD (Italy)	158	PNPLA3 I148M carriers with IBD have higher susceptibility to hepatic steatosis and liver damage
Atkinson *et al.* [119]	UK/Irish	1188	Carriers of the genetic variant of PNPLA3 are more at risk of developing severe alcoholic hepatitis and have less survival even after stopping drinking

(Adapted from Bruschi *et al.* 2017).[3]

1.4.2 PNPLA3 gene

PNPLA3 has been reported using multiple aliases as identified by a database search of Uniprot KB and NCBI Entrez. These names have been used interchangeably to refer to either the gene or the protein product, although *PNPLA3* is now the accepted nomenclature (Table 1.4).[120,121]

Table 1.4 Alternative names for *PNPLA3* as listed in Uniprot KB and NCBI Entrez [120,121]

Alternative names
Patatin-Like Phospholipase Domain-Containing Protein 3
Patatin-Like Phospholipase Domain Containing 3
Adiponutrin
ADPN
Calcium-Independent Phospholipase A2-Epsilon
IPLA2-Epsilon
IPLA (2) Epsilon
IPLA2epsilon
Acylglycerol O-Acyltransferase

PNPLA3 is roughly 24 kb in length, containing nine exons, located chromosome 22. Five transcripts have been detected, 4 of which are believed to be non-protein-coding variants. The transcript ENST00000216180 encodes the reference protein sequence.[122]

The key variant shown to be associated with liver disease, rs738409, encodes a cytosine to guanine substitution on the forward strand of chromosome 22 at the position 43,928,847 on human genome build 19 (hg19). This is a common variant, with minor allele (rs738409[G]) frequencies ranging from around 10-50% in global populations (Table 1.5).[123,124]

There is a high level of linkage disequilibrium across the PNPLA3 gene. However, the robust association with the rs738409 has been supported by a range of biological investigations, supporting the fact that this variation is likely the major functional driver of the association.

Table 1.5 Minor allele frequency across populations tested in the 1000 genomes project and UK10K [123,124]

Genome Project	Population	MAF (G)
1000 genome	ALL	0.26
	African	0.12
	American	0.48
	East Asian	0.35
	European	0.23
	South Asian	0.25
UK10K	All	0.22

1.4.3 Sites of *PNPLA3* expression

PNPLA3 expression has been extensively catalogued in both human and murine subjects, where expression was detected in multiple metabolically active tissues (Table 1.6).

In rodents, *pnpla3* is predominantly expressed in brown adipose tissue and cardiac tissue, with very low levels in the liver and pituitary gland.[125] Upon feeding of a western diet, expression of *pnpla3* in the liver specifically is highly upregulated (26 fold) raising levels closer to those of cardiac tissue.[126]

In human subjects, *PNPLA3* is mainly expressed in the liver, followed by followed by skin and adipose tissue with expression levels one third that of the liver. Further characterisation within the liver show hepatocytes express four times more *PNPLA3* than hepatic stellate cells.[6]

The different expression profiles in mice and humans likely reflects the metabolic differences between the two species rather than there being large functional difference between the protein products.

The fact *PNPLA3* is mainly expressed in the liver supports a functional role for *PNPLA3* within the liver. It would, therefore, be reasonable to suppose that *PNPLA3* polymorphisms could disrupt liver function and play a role within liver disease.

This is supported by a higher prevalence of steatosis and exclusively NASH cases developed in liver transplant recipients who received livers homozygous for the PNPLA3 I148M allele. Also suggesting PNPLA3 exerts its key physiological action in the liver.[127]

To date, attention to the potential role of *PNPLA3* in other tissues in the body have taken a backseat. However, fasting conditions have been shown to increase expression levels of *pnpla3* in the pituitary gland 1.7 fold.[128] This could suggest local regulatory circuit that may fine tune the feedback effects of adipose hormones in the control of energy balance within the brain. So, while the effects of the I148M variant on liver disease are probably derived within the liver tissues, it is likely that the role of PNPLA3 extends beyond that of a liver enzyme.

Table 1.6 Tissues expressing PNPLA3 in mice, rats and humans, ranked from the highest expression to the lowest (1-4) in each species

Species	Tissue	Ranked expression
Human (*Homo sapiens*)	Liver (Hepatocytes, Hepatic stellate cells)	1
	Skin (epithelial cells)	2
	Adipose Tissue (adipocytes)	3
Mouse (*Mus musculus*)	Adipose tissue (adipocytes)	1
	Cardiac tissue	2
	Liver (Hepatocytes)	3
	Brain (Pituitary gland)	4
Rat (*Rattus norvegicus*) *	Adipose tissue (adipocytes)	1

*Other tissues were not tested in this species

1.4.4 Regulation of *PNPLA3* expression

Both *pnpla3* and *PNPLA3* are primarily nutritionally regulated through insulin mediated responses, however, are influenced by multiple regulatory pathways allowing complex fine-tuned responses for tight control of protein expression. There is no evidence to date that rs738409 has an impact on the expression levels of *PNPLA3*.

1.4.4.1 Nutritional regulation

Pnpla3 and *PNPLA3* are upregulated when fed a high carbohydrate diet, in particular with simple sugars such as sucrose and fructose. This effect has been shown to be magnified in obese mice, however is not effected by high fat diets.[7,94,97,129]

It is believed this is mediated almost entirely through a response to insulin. Expression is increased upon insulin injection,[130] and the expression levels of *PNPLA3* correlate with those of phosphatidylinositol-3 kinase (PI3K) and hexokinase II (HKII) in the basal state and upon insulin infusion.

Induction of transcription after feeding is rapid within 3 hours and the mRNA product half-life has been observed to be around 5 hours, which is similar to other genes associated with energy homeostasis such as leptin.[95]

1.4.4.2 Transcription factor binding sites

Promoter sites for both SREBP factors and CCAAT binding nuclear transcription factor Y (NFY), which is known to interact with SREBP are predicted near the *PNPLA3* site.[131] This same combination of promotors is also found for lipoprotein lipase (LPL) and hepatic lipase (LIPC) which could suggest a functional link with these enzymes.

Both the SRE motif and NFY motif were shown to promote gene expression in a synergistic fashion, although the primary activation of the *PNPLA3* promoter was through SREBP.[132]

1.4.4.3 *PNPLA3* promotion by sterol regulatory binding proteins

In both mice and humans, SREBP-1c promoted *PNPLA3* expression and SREBP-cleavage activating protein (SCAP) murine knockouts lost most of this response.[133] In murine subjects

carbohydrate regulatory binding protein (ChREBP) also increased expression of *PNPLA3* transcripts, however human cell lines do not respond to ChREBP, supported by no detectable binding site in the vicinity of the human gene.[134–136]

All of the three main SREBPs, SREBP-1a SREBP-1c and SREBP-2 cause an increase in *PNPLA3* gene expression, however in mice SREBP-1a shows the highest increase causing a 100-fold increase from basal expression levels, followed by SREBP-1c with 12-fold increase; This suggests SREBP-1a is the predominant promotor of *pnpla3* in mice.

Despite this observation, SREBP-1c is more commonly produced in the liver, where it would not be expected that SREBP-1a occur at high levels, leading to the hypothesis that SREBP-1c is the most important promotor of *PNPLA3* in man. Generally, SREBP-1a is only produced in cells with a high rate of proliferation and is activated by the mitogen activated protein kinase activity (MAPK) pathway (Figure 1.10). Since this transcription factor has not been well characterised in disease state cells, this should not be ruled out as an important additional promotor.[137]

1.4.4.4 Alternative regulation of *PNPLA3*

While SREBP upregulation is the best classified example of *PNPLA3* promotion, multiple other pathways facilitating regulation of PNPLA3 have been identified.

Thiazolidinedione compounds, which are commonly used as diabetic treatments, are agonists to the Peroxisome proliferator-activated receptor-γ (PPARγ) have been shown to downregulate *PNPLA3* expression.[96,138]

Pnpla3 has been shown to be downregulated by the agonist of the glucagon like peptide 1 (GLP-1) receptor agonist Exenatide.[139] This process is not well characterised, however may operate through a pathway involving SIRT1 a NAD+ dependent protein deacetylase; which through the deacetylation of lysine residues on various histones and transcriptional factors, is known to among other things, cause a downregulation of SREBPs.

Pnpla3 was shown to be downregulated in response to cold,[140] while expression of *pnpla3* in rats and *PNPLA3* in cultured human cells was decreased by the hormone Tri-iodothyronine (T3).[141] Since T3 levels are reduced under cold adaptation this could provide one potential mechanism by which cold may downregulate *pnpla3* expression.[142]

After incubation with retinol and palmitic acid, there is a time-dependent downregulation of *PNPLA3*, which is accompanied by a parallel increase in lipid droplet accumulation.[143]

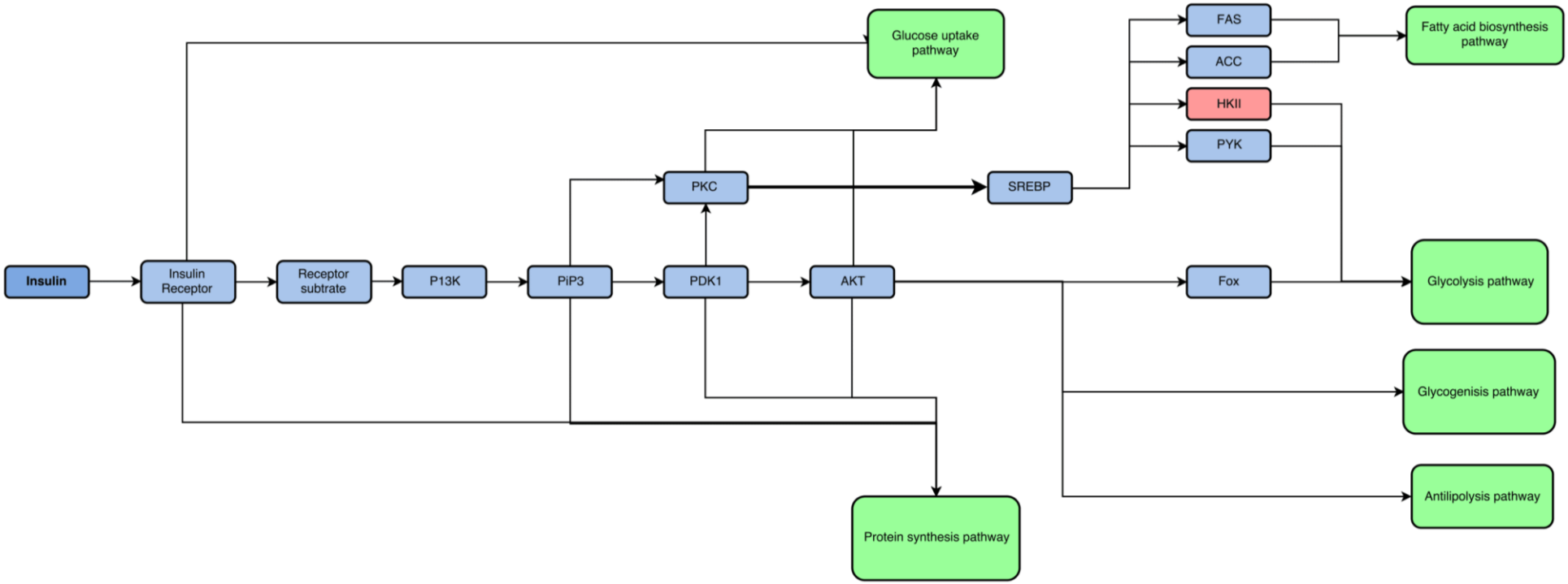


Figure 1.10 A simplified version of the downstream insulin activated pathway

The path to HKII is highlighted with a pink box. The blue boxes represent an upregulated protein. The green boxes represent the downstream effect of the previous protein state.

Ethanol can increase gene expression in hepatic cells cultured from rat livers.[144] This could once again be linked to the SREBP pathway which is modified by alcohol through the effect on the NADH/NAD+ ratio within the cell, and is lent support by the fact that levels of *PNPLA3* in studies of ALD were higher than in NAFLD across studies.[98,101]

In HSCs PNPLA3 expression is downregulated in response to incubation with retinol, coupled with an increase in lipid droplets within the cell.[143]

Small saturated fatty acids are the downstream result of SREBP-1c expressed proteins and were shown to increase the half-life of PNPLA3 from 2.4 to 6.7 hours; The greatest effect observed from palmitate and oleic acid. The combined effect of SREBP induced expression with a longer protein half-life creates a feed forward loop greatly increasing the amount of PNPLA3 present within the cell.[6]

PNPLA3 is clearly implicated in a number of important metabolic functions, through a variety of careful regulatory pathways. The most obvious role, based on strong evidence of SREBP modulation, would be that of a lipogenic agent; however, it is likely to have a more complex role in overall lipid homeostasis.

1.4.5 Homology

PNPLA3 is a protein encoding gene, which is translated into the protein also referred to as PNPLA3. Like the gene, the protein has also been described under alternative names, however PNPLA3 is now the commonly accepted naming convention.

PNPLA3 is a member of the patatin-like phospholipase domain containing protein (PNPLA) gene family and shares homology with a broad range of proteins across the tree of life, including over three thousand species of prokaryotes and three hundred species of eukaryotes.[145]

The domain which is conserved between the PNPLA3 superfamily is the patatin domain. In PNPLA3, this spans approximately 170 amino acids from residue 10 to 179, however the C-terminal residues share little to no conservation with any other known proteins.

1.4.5.1 Patatin

The domains namesake, patatin, is the major protein component of potatoes, where it acts as both a lipase and a storage molecule.[146] Remarkably this can make up to 45% of the soluble

protein content of the potato and is a member of the broader super family of serine hydrolases.[147,148]

The structure and function of patatin is well characterised and its unique sequence motifs produce a characteristic SER-ASP catalytic dyad and classical $\alpha/\beta/\alpha$ hydrolase fold.[147,148]

1.4.5.2 Human patatin-like phospholipase domain containing proteins

Nine members of the PNPLAs family are found in man,[149] which display a wide range of lipase and acyltransferase activity, consistent with the broad activity of patatin (Table 1.7).[150] However alignment of the sequences for comparison is challenging because of minimal homology between the proteins.[149]

Table 1.7 The human patatin like phospholipase domain containing proteins (PNPLAs) with alternate nomenclature and functions, where known [150]

PNPLA nomenclature	Alternate nomenclature	Suggested function
PNPLA1	None	Unknown
PNPLA2	Adipose triglyceride lipase (ATGL)	Triacylglycerol hydrolysis, transacylase activity
PNPLA3	Adiponutrin (ADPN)	Unknown
PNPLA4	Gene sequence 2 (GS2)	Retinol ester hydrolase, acylglycerol and retinol transacylase, triacylglycerol hydrolase
PNPLA5	Gene sequence 2 like (GS2-like)	Triacylglycerol hydrolase
PNPLA6	Neuropathy target esterase (NTE)	Lysophospholipid esterase and hydrolase
PNPLA7	Neuropathy target esterase related esterase (NRE)	Lysophospholipid esterase
PNPLA8	Calcium independent phospholipase 2 gamma ($iPLA_2\gamma$)	Lysophospholipid hydrolase
PNPLA9	Phospholipase 2 group iv (PLA2G6)	Glycerophospholipid hydrolase, acetyl hydrolase

These nine proteins can broadly be separated into two groups, those with phospholipid specificity (PNPLA1, PNPLA2, PNPLA3, PNPLA4, PNPLA5) and those without (PNPLA6, PNPLA7, PNPLA8, PNPLA9); this separation can be clearly seen in phylogenetic analysis (Figure 1.11).[151]

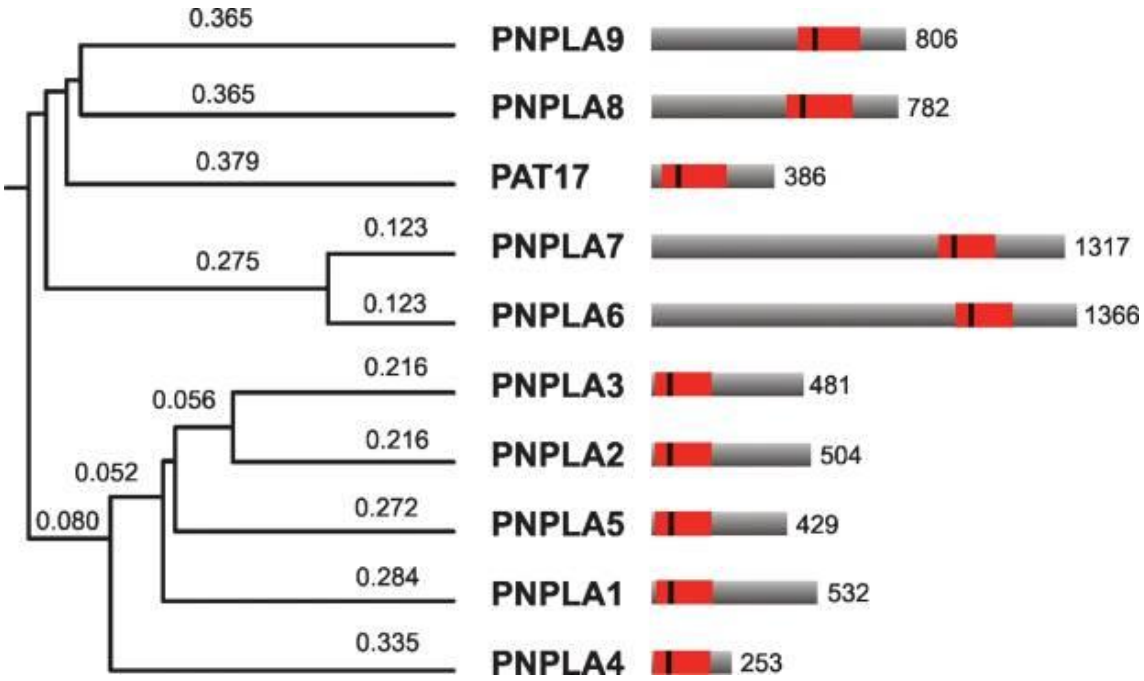


Figure 1.11 Relationship between human patatin like phospholipid protein homologues

On the left-hand side is a distance based phylogenetic tree representing the relationship between human PNPLA proteins and an isoenzyme of patatin, PAT17. On the right-hand side are the structural alignments of these proteins. The patatin domain of each is highlighted with a red box and the predicted active site serine as a black bar. The numbers on the right denote the full amino acid length of each protein (Adapted from Kienesberger *et al.* 2009).[150]

PNPLA3 shares the highest degree of homology with PNPLA2, which is its closest common ancestor. PNPLA2, commonly referred to as adipose triglyceride lipase (ATGL), possesses both phospholipase activity and transacylase activity and like PNPLA3 also localises to lipid droplets.

PNPLA2 is responsible for the initial step in TG hydrolysis[152] and loss of function cannot be fully compensated for, causing rapid accumulation of lipids within the cell.[153] This is supported by a range of clinical pathologies brought about by mutation in this gene, for example Jordan's anomaly and neutral lipid storage disease with myopathy.[154,155]

Close homology between PNPLA2 and PNPLA3 may lead us to predict the proteins have a similar biochemical function, However, *PNPLA2* is regulated in an almost opposite fashion to *PNPLA3;* being upregulated during fasting and suppressed by feeding.[151] In addition peroxisome

proliferator activated receptors (PPARγ) agonists upregulate *PNPLA2* but downregulate *PNPLA3*.[96,138] This instead suggests a potentially contrasting complimentary function for *PNPLA3*.

1.4.5.3 PNPLA3 homologues

There are a range of PNPLA3 homologues which are found primarily in mammals, although none of these proteins are well characterised. The orthologues of PNPLA3 form related clusters, showing more recent common ancestors than between the other PNPLA proteins. Frog and chicken PNPLA3s form the exception to this rule being more closely related to their PNPLA2 homologues

The most closely related proteins to human PNPLA3 are found in the rodent species, followed by pig, cow and fish respectively (figure 1.12).[156]

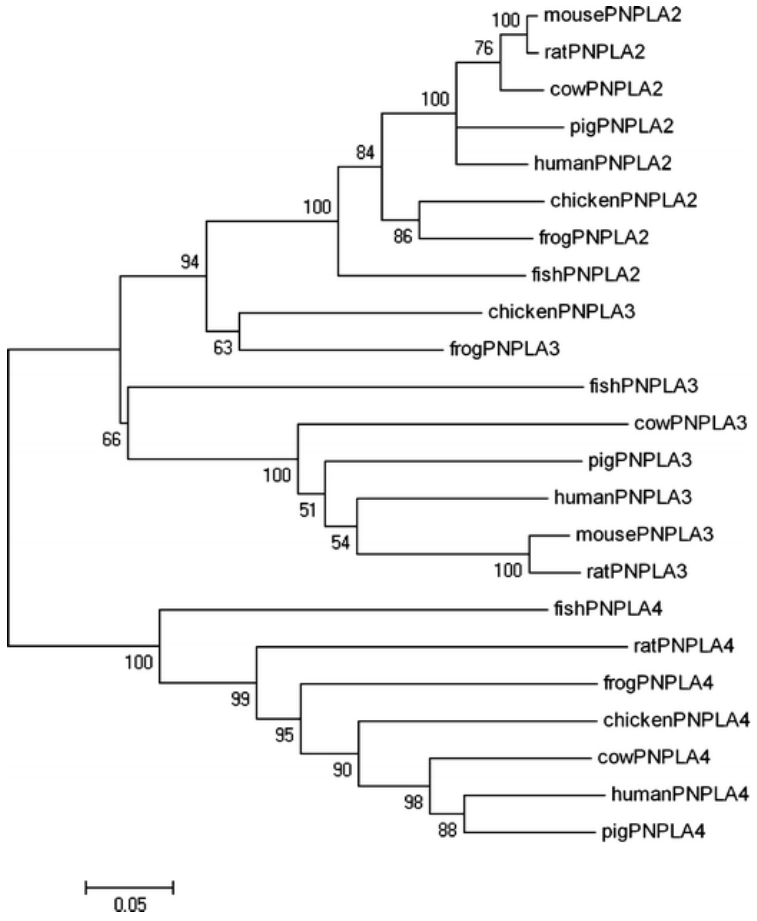


Figure 1.12 The evolutionary link between patatin like phospholipase proteins across several species

Distance based phylogenetic tree constructed on homologous sequences (Adapted from Chen et al. 2011).[156]

1.4.6 PNPLA3 protein

PNPLA3 is 481 amino acids in length and has a molecular weight of 52,865 da. There are two potential glycosylation sites at amino acid 89 and 280, but no other known post translational modifications apart from ubiquitination.

The rs738409 variant corresponds to an isoleucine to methionine change at residue 148 (I148M). The change from isoleucine to methionine results in little change in the local environment as both amino acids have non-polar side chains with similar chemical properties (Figure 13).

The largest difference between these amino acids is the size and shape of the side chains, as Isoleucine has a branched chain with a long chain 3 carbons in length, while methionine has a straight chain 4 bonds in length. This will result in different steric hindrances and potential interactions around the residue.

Additionally, methionine contains a sulphur atom, which can be polarised and in principle could be oxidised to form salt bridges, however in practise this is very unlikely. Overall, the extent that this might affect the structural integrity and in turn the enzymatic activity is less clear without further examination of the local environment in the protein.

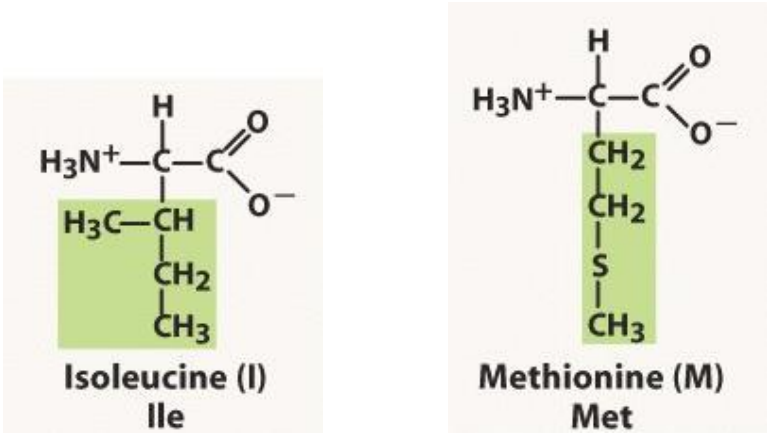


Figure 1.13 The chemical structure of amino acids Isoleucine and methionine [157]

1.4.7 Subcellular localisation

PNPLA3 was predicted to be a transmembrane protein computationally, supported by early studies in which PNPLA3 was isolated from the membrane fraction of the cell.[125] It has now been shown to only associate to membranes and lipid droplets equally in Hepatoma (HuH-7) cells, and imaging studies of *ex-vivo* cell lines observed the protein almost solely associating to lipid droplets (Figure 1.14).[158,159]

The association to lipid droplets is maintained by truncated constructs of PNPLA3 expressing only the patatin domain, whereas C-terminal variants have been shown to have modified sub-cellular localisation. This suggests the patatin domain may confer the tendency to associate to the lipid droplets, while the C-terminal domain which varies more widely between homologues plays a more intricate role in intracellular trafficking.[158,160]

The propensity to associate to lipid droplets and cell membranes does not differ between the wild type and I148M variant, or a small construct expressing only the patatin domain.[158]

PNPLA3 has also been found in human blood plasma as a multimeric complex (0.8-2.5µg/ml).[161] While this level of circulating protein is very low, it is comparable to other low abundance proteins known to have circulatory function, such as adiponectin which circulates at 5-10 µg/ml.[162] Plasma levels of each variant have not been compared. It is possible these levels are an artefact of proteins released after cell death during normal liver damage.

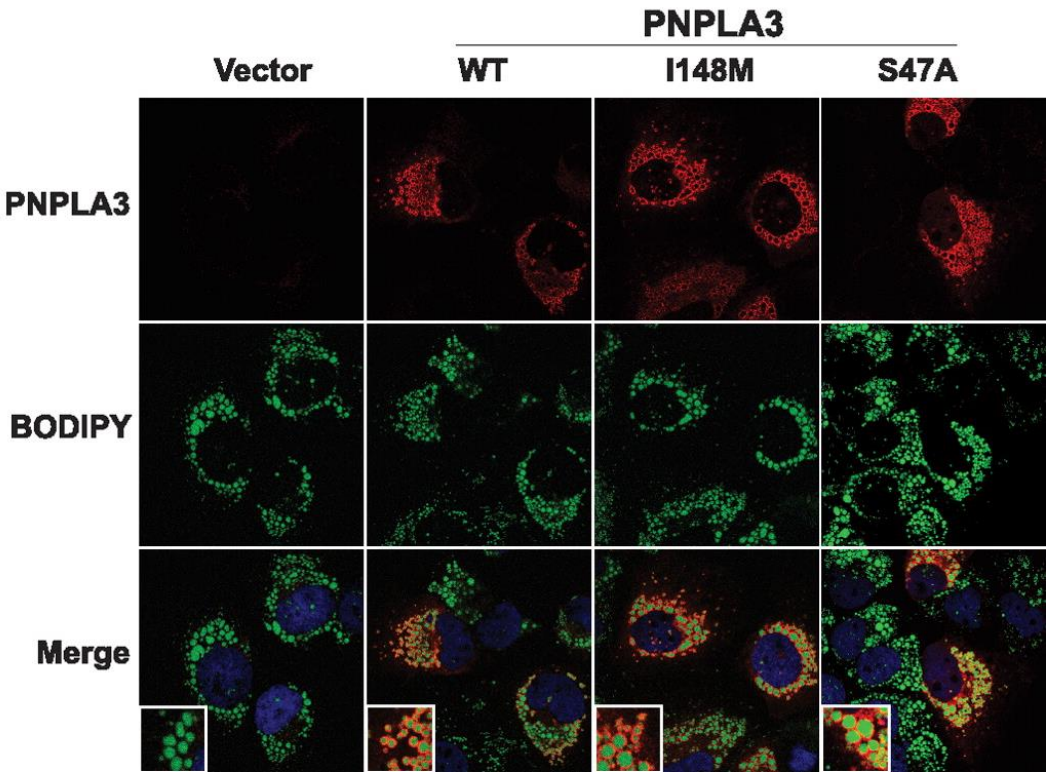


Figure 1.14 Immunolocalization of recombinant human PNPLA3 to lipid droplets

Human hepatoma cell line (HuH-7) cells infected with a control recombinant adenovirus (vector) or with adenoviruses encoding either wild type *PNPLA3, PNPLA3-I148M* or *PNPLA3-S47A*. PNPLA3 is visualised in red, boron-dipyrromethene (biodipy) used to stain lipid droplets in green and areas of co-localisation are shown in yellow. (Adapted from He *et al.* 2010).[158]

1.4.8 Biological activity

To elucidate the biological activity of PNPLA3, a combination of *in vitro* biochemical assays, *ex vivo* cell lines and *in vivo* animal models have been used across the literature. A range of often contrasting functions have been detected across these studies, meaning the nature of the activity of PNPLA3 remains unknown.

1.4.8.1 *In vitro* functional studies

PNPLA3, has been shown to have a broad range of activities *in vitro*, including hydrolase activity towards a broad range of substrates, *viz.*: The hydrolysis of 1-palmitoyl-2-[1-14C]linoleoyl-sn-glycerol-3-phosphocholine to linoleic acid and 1-palmitoyl-2-[1-14C]arachidonyl-sn-glycerol-3-phosphocholine to arachidonic acid.[163] Hydrolysis of the colorimetric lipase substrate DGGR,[151] triolein,[158,164] triacylglycerol (TAG), diacylglycerol (DAG),[165] and most recently as a retinol esterase.[143]

In an opposing fashion, lipogenic functions have also been observed, namely transacylase activity forming triolein from monoolein[163] and acyl-CoA-dependent acylation of LPA to generate phosphatidic acid (PA).[166]

While several studies have found PNPLA3 to behave in a predominantly lipogenic fashion,[163,166] most observe a preference for lipolytic activity and find little to no lipogenic activity at all.[164,165] The general consensus from *in vitro* analysis supports the main function of PNPLA3 being lipolytic in nature.

In these studies, there is a preference to cleave at the sn1 and sn3 fatty acid positions in TAG, which leads to a relative accumulation of 1,2-DAG. Across all lipase assays PNPLA3 showed a strong preference for glycerolipids at all fatty acid chain lengths, in which the acyl group was oleic acid (triolein, diolein, monolein).[165] Additionally, substrate specificity in lipogenic assays was for oleoyl-CoA, linoleoyl-CoA, and arachidonoyl-CoA as acyl donors.[166]

1.4.8.2 *Ex-vivo* functional studies

The activity of PNPLA3 in *ex-vivo* functional studies has been observed in proxy through changes in the cellular phenotype. In general, PNPLA3 has been observed to affect the levels of cellular TAGs, but like *in vitro* studies, contrasting responses have been observed across experiments.

PNPLA3 was overexpressed in Hepatoma (Huh-7) cells showed no change in cellular triglycerides compared to expression of an empty vector.[158] Following this, overexpression of PNPLA3 in Hek293 cells only showed a non-significant increase in cellular TAG.[151,167]

PNPLA3 expression in primary hepatocytes from mice caused an Increase in intracellular triglyceride accumulation.[136] PNPLA3 and pnpla3 overexpression in Cos-7 cells caused around a 2-fold increase in LPAAT activity.[166]

PNPLA3 has no retinyl-esterase activity in hepatocytes but does in HSCs. In HSCs, there is an increase in lipid droplet accumulation paralleled by a downregulation of PNPLA3 in response to retinol availability.[143]

In contrast to other studies, overexpression of PNPLA3 in McA-RH 777 cells caused a reduction in intracellular lipid content and higher ApoB secretion, used as a measure of secreted very low density lipid (VLDL) sized, ApoB containing, triglyceride rich particles.[168]

Most recently overexpression of PNPLA3 in Huh-7 cells Induced an increase relative amount of TAG compared with saturated fatty acids and MUFA and caused an increase in the PUFA within the membrane. Despite this, there was no change in the inclusion of labelled glycerolipids compared with controls, suggesting a complex role in intracellular lipid remodelling.[169]

1.4.8.3 Animal model functional studies

Initial animal studies in mice, have shown that overexpression or silencing of *PNPLA3* has no large physiological effect on the animal. Indeed overexpression of *pnpla3* in wild type or obese mice showed no significant physiological changes or in cellular TAG content.[158] Similarly basal and overexpression levels of *PNPLA3* resulted in no significant differences when fed a range of fatty or western diets.[170]

Pnpla3 deficient mice backgrounds appeared to have no changes in hepatic lipid content, on a range of Chow or high fat or high carbohydrate diets. Although an increase of *pnpla5* mRNA expression in adipocytes could be performing a compensatory role.[171,172]

More recently, more subtle effects of PNPLA3 expression within animal models have been detected, whereby PNPLA3 expression correlated to a depletion of very-long-chain polyunsaturated fatty acids (vLCPUFA) in the hepatic lipid droplets and an enrichment in phospholipids. Once again suggesting a more complex role in the remodelling of triglycerides and phospholipids in LDs.[173]

1.4.8.4 PNPLA3 Inhibitors

The only demonstrated inhibitor of PNPLA3 is bromo-enol lactone (BEL) (figure 1.15) which was shown to inhibit lipase activity at low concentration of 0.1μM.[163] BEL is however a broadly acting irreversible lipase inhibitor, and is not useful for biological *in vivo* inhibition experiments.

Inhibition with BEL further supports the mechanism of a catalytic dyad similar to patatin, and it is likely many broad-spectrum lipase inhibitors will also inhibit PNPLA3.

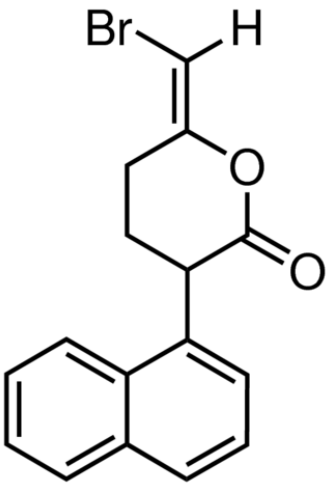


Figure 1.15 The chemical structure of the bromo-enol lactone (BEL).

1.4.9 The effect of the I148M variant on biological activity

1.4.9.1 *In vitro* functional studies

In vitro functional studies have shown that the purified I148M variant PNPLA3 most commonly reduces triacylglycerol lipase and LPAAT to almost baseline levels of activity[158,164,165] although it has also been observed to cause a 2-fold increase in LPAAT activity.[166]

1.4.9.2 *Ex-vivo* functional studies

Overexpression of the PNPLA3 I148M variant in cell lines consistently resulted in an increase in cellular triglycerides. However this has been attributed to either a loss of lipase activity,[135,158,170] or a gain of LPAAT activity.[166] In one study, the higher intracellular lipid content when expressing

the I148M variant was attributed to a decrease of secreted very low density lipid (VLDL) sized, ApoB containing, triglyceride rich particles.[168]

Rather than a total increase in cellular lipids, more recently PNPLA3 expression has been shown to play a role in complex remodelling of the LD lipidome. Expression of the I148M PNPLA3 variant caused net increase of unlabelled TAG within these experiments, suggesting a loss of this activity.[169]

1.4.9.3 Animal model functional studies

While the wild type enzyme has no apparent effect on hepatic lipid accumulation, overexpression of the I148M variant in murine models results in an increase in lipid droplet quantity and size, as well as increased tissue levels of triglycerides and cholesterol esters. A similar result was obtained with another catalytically inactive S47A variant of PNPLA3, lending support for the I148M variant being catalytically inactive.[158,170,174]

A recent investigation found a complex role in the remodelling of triglycerides and phospholipids in LDs of the I148M variant. In the LDs of I148M knock-in mice, vLCPUFAs were depleted from triglycerides and enriched in phospholipids. Unlike previous studies, this was not attributed to a simple loss of activity because an inactivated S47A variant caused vLCPUFAs to be enriched in triglycerides and depleted from the phospholipids.[173]

1.4.10 Structure of PNPLA3

To date no experimentally derived 3-dimensional structure of PNPLA3 has been reported. The first 179 residues of the patatin domain have been modelled *in silico*, but further modelling has been limited by the lack of homology with other proteins of known structure.

The current model of the patatin domain suggests that PNPLA3 maintains a conserved α/β, patatin-like fold, similar to that of patatin; with the same conserved serine-aspartate catalytic dyad. This supports functional studies that PNPLA3 maintains the broad lipase activity of patatin, likely *via* a similar reaction mechanism (Figure 1.16).[149]

Further modelling of the I148M variant places residue 148, in the active site pocket close to the catalytic restudies. This has been used to hypothesise that the Ile to Met substitution may result in steric hindrance to the active site, limiting lipase activity (Figure 1.17).[158]

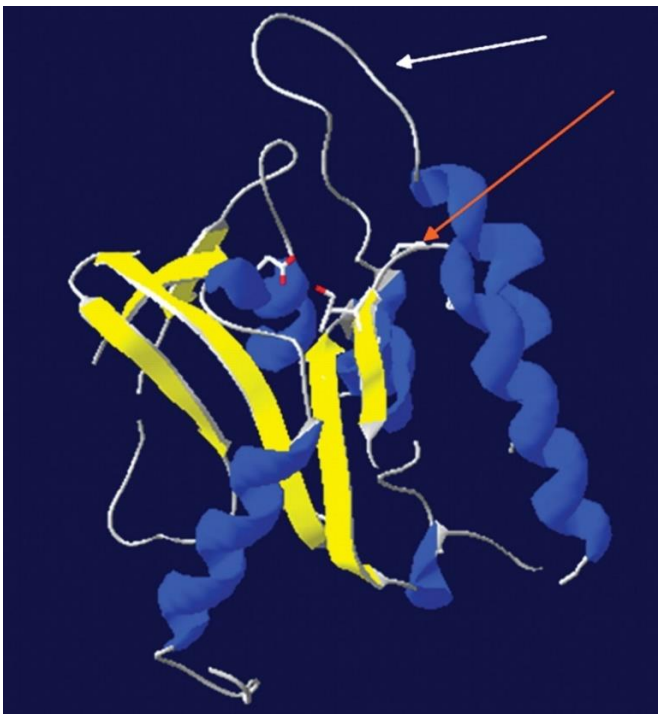


Figure 1.16 Homology model of the human PNPLA3 patatin-like domain

Image generated using the DeepView/Swiss-PDB Viewer. Regions predicted to fold as β-sheets and α-helices are shaded yellow and blue, respectively. Side chains of the catalytic aspartate and serine residues are rendered in stick format and the atoms coloured using standard Corey, Pauling & Kultun (CPK). The glycine-rich region is indicated by the red arrow and the putative "lid" region by the white arrow (Adapted from Wilson *et al.* 2006).[149]

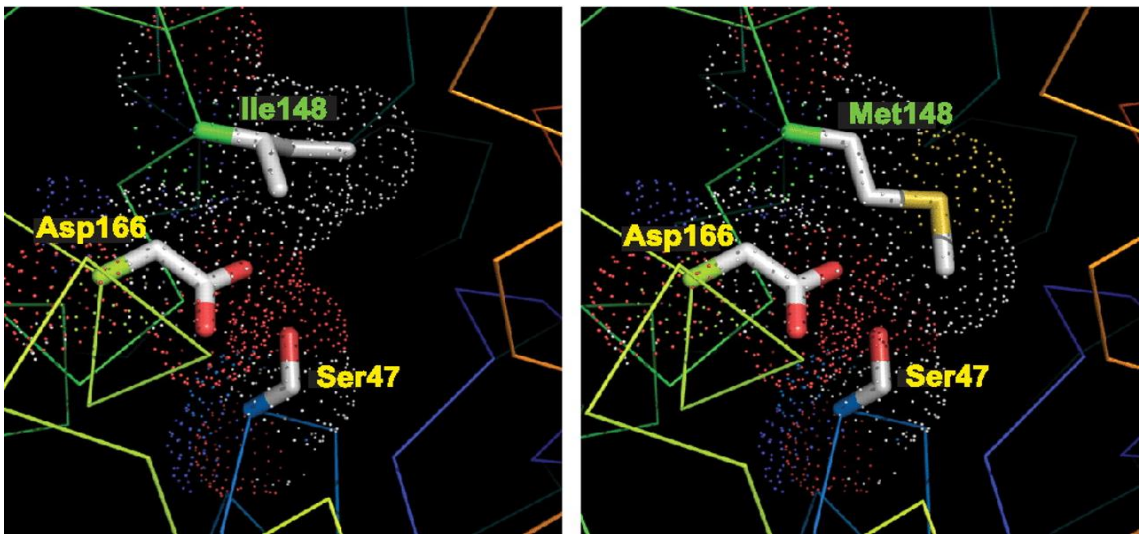


Figure 1.17 Homology model of wild type and mutant (I148M) PNPLA3

The domain structure of PNPLA3, showing the patatin-like domain (black) and locations of the catalytic dyad (Ser47 and Asp166) and the I148M substitution is shown. Protein traces are rainbow-colored from N to C terminus (blue to red) with side chains of catalytic dyad residues (positions 47 and 166) shown. The dots indicate a space-filling model corresponding to van der Waals atomic radii with oxygen and sulphur atoms are coloured red and yellow, respectively (Adapted from He *et al.* 2010).[158]
Left panel; Models normal (Ile148) PNPLA3 variant.
Right panel; Models mutant (Met148) PNPLA3 variant.

1.5 The role of PNPLA3 I148M in the pathogenesis of liver disease

With our current understanding of both PNPLA3 and the pathogenesis of liver disease, the precise nature of the role of the I148M variation in the progression of the disease is unclear.

Lipids are known to play a key role in the pathogenesis of liver disease; as PNPLA3 has been shown to localise to lipid droplets within the cell, and the I148M variant is associated with increased intracellular TAGs and lipid droplet size, it is possible the I148M variant plays a role in the disease through a mechanism which leads to increased levels of steatosis within the liver.[175]

Studies observing the increase of cellular lipids in *ex vivo* cell lines expressing the I148M variant, have led to several theories as to how this effect may come about:

(i) PNPLA3 is lipolytic in its wild type form and the I148M variant is associated with lose of that activity resulting in lipid accumulation in hepatocytes.[169]

(ii) PNPLA3 is lipogenic in its wild type form and the I148M variant is associated with an increase in function leading again to lipid accumulation in hepatocytes.[166]

(iii) A functional change *per se* is not the cause of steatosis, but rather the accumulation of inactive PNPLA3 on lipid droplets causes accumulation of TG by restricting access to the lipid droplet or sequestering a factor required for hydrolysis.[174]

(iv) The I148M causes an accumulation on lipid droplets due to reduced ubiquitination and leads to increased net activity. This impacts the homeostatic lipid balance and has complex downstream effects leading to a pathogenic response.[173]

While it is easy to assume that it is simply through the increasing size of lipid droplets that PNPLA3 exerts its effects, complex interactions with other cellular processes may also play a pivotal role.

Retinyl esters are hydrolysed to retinol in hepatocytes and then transferred to hepatic stellate cells (HSCs) and stored as retinyl palmitate in lipid droplets. When needed, the retinol stored in lipid droplets can be released to reach the specific sites where it exerts its physiological functions.[143] Cells expressing PNPLA3 have been shown to reduce levels of retinol within hepatic stellate cells, and contribute to an activated profibrogenic phenotype.[176]

In addition, circulating levels of PNPLA3 imply there may be additional signalling functionality of the protein which requires further investigation.[161]

1.6 Summary

The I148M variant of PNPLA3 represents the most significant and consistent genetic risk factor for liver injury; Associations have been detected between this genotype and risk of liver injury in NAFLD, ALD, Wilsons disease, chronic hepatitis C and hepatocellular carcinoma.

The precise biological function of PNPLA3 is not well understood because of conflicting evidence from a range of investigations. Most evidence supports lipase functionality for PNPLA3, with hydrolytic activity against a broad range of substances; however, lipogenic functions have also been detected.

The I148M variant is known to cause an increase in lipid droplet size and number *in vivo*, however the biochemical mechanism by which the variant causes this phenotype is unknown. Similarly, the role of the variant in liver injury above the associated risk has yet to be elucidated.

Because of the sheer range of disease related phenotypes that the I148M variant is associated with, it provides an exciting drug target, which may open the door to developing the first ever genetics based broad spectrum treatments for liver disease.

Gaining an understanding of the true function of PNPLA3 and the effect of the key I148M mutation is required to validate PNPLA3 as a drug target and leverage any further clinical opportunities.

1.7 Thesis aims

The overarching aim of this thesis is to elucidate the role of the I148M PNPLA3 variant in the pathogenesis and progression of liver disease. As the genetic association with liver disease has been extensively explored, this work focuses on answering this question through structural and functional characterisation of both common PNPLA3 variants.

The specific aims of this thesis are:

Chapter 2: To use a range of bioinformatic tools to predict the biochemical properties of PNPLA3 based on the amino acid sequence, and inform further *in vivo* and *in silico* experimental design.

Chapter 3: To develop a system for the robust expression and purification of PNPLA3 in quantities amenable for structural and functional characterisation.

Chapter 4: To create a detailed 3-dimensional model of PNPLA3 to gain additional insight into the structure of PNPLA3 and the position of the I148M variant.

Chapter 5: To perform detailed molecular dynamic simulations of both PNPLA3 variants to predict the mechanistic implications of the I148M variant on the protein structure and function.

Chapter 2

Primary-sequence based investigation of PNPLA3

"The important thing is not to stop questioning. Curiosity has its own reason for existence. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvellous structure of reality. It is enough if one tries merely to comprehend a little of this mystery each day."

Albert Einstein

2.1 Overview

A significant amount of information can be gleaned from the primary sequence of a protein through phylogenetic comparison with homologous proteins and consideration of the innate properties of the constituent amino acids.

Sequence-based studies have provided information on PNPLA3 by reference to the putative function and structure of other PNPLA proteins within the Patatin like phospholipase family. However, these investigations have not addressed the fundamental character of the protein.

In this chapter, bioinformatic tools are used to gain an improved understanding of PNPLA3 based on its phylogenetic relationships with related proteins whose structure and function has been defined, and the properties inferred through the primary amino acid sequence.

The phylogenetic investigations better define the position of PNPLA3 within the larger framework of the Patatin and cPLA2 superfamily of proteins. Clustering of the protein subfamilies suggest that the closest structurally related homologue to PNPLA3 is PNPLA5 rather than PNPLA2 as was previously described.

Bioinformatic prediction tools suggest that PNPLA3 is unlikely to be amenable to *in vitro* purification or downstream structural studies. However, the prediction software identified a range of PNPLA3 fragments, which are better candidates for purification and could be taken forward for further structural and functional exploration.

2.2 Introduction

Four distinct levels of increasing structural complexity are used to describe proteins, *viz:*

(1) primary structure, consisting of the amino acid sequence;
(2) secondary structure, consisting of regular local substructures along the protein backbone;
(3) tertiary structure, comprised of the three-dimensional structure of the whole protein subunit; and
(4) quaternary structure, comprised of the complex spatial arrangement between multiple independent polypeptide chains.

The primary amino acid sequence of a protein determines its structure and function, because all levels of protein structure and function in their essential nature arise from the information contained within the primary sequence. Thus, a protein can theoretically be fully characterised if its primary sequence is known.

In practice, our ability to predict structural information from the sequence is limited due to the complexity of the protein system, hence bioinformatic approaches to understand a protein based on its sequence alone rely on: (i) interrogating the structure and function of homologous proteins identified via phylogenetic-based approaches; and (ii) determining the structural implications of the amino acid sequence.

The phylogenetic approach relies on the conservation of both structure and function between ancestrally-related proteins. It allows information about both the structure and function of an unknown protein to be inferred based on comparisons between proteins with similar ancestry.

The structural approach is an iterative process by which information gleaned from successive stages of the structural determination is used to inform the next stage. This process leverages the known mechanisms by which the basic properties of the primary structure interact to generate the quaternary structure.

By combining these mechanistic and phylogenetic approaches it is possible to obtain substantial information about the structure and function of a target protein from the primary sequence data alone. The sequences of thousands of proteins are now readily available thus considerably facilitating this approach. The low costs of implementation and the ability to work with proteins which are not amenable to structural investigation make it a valuable alternate approach to traditional experimentation.[177]

2.2.1 Primary structure

The primary structure of a protein is defined by the linear sequence of amino acids which comprise the polypeptide chain. This sequence is directly encoded within the translated gene.

In general, all proteins are created from a combination of 20 common naturally occurring amino acids (Figure 2.1).

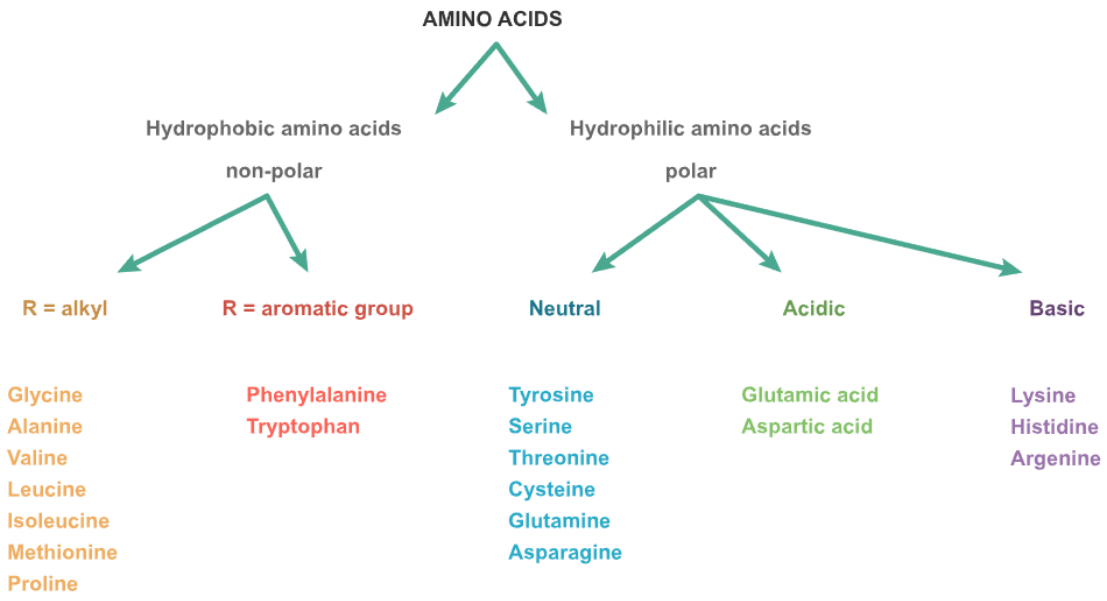


Figure 2.1 Flow chart separating amino acids based on chemical properties

(Adapted from www.khanacademy.org).[178]

Each amino acid has a similar core structure, consisting of an amino group joined to a carboxyl group through a single carbon atom, denoted the α carbon. Individual amino acids only differ in the side chain, the R group, which extends from this α carbon (Figure 2.2).

The amino acid side chains are integral to the protein's properties; these are generally classified into five categories: non-polar alkyl, non-polar aromatics, polar uncharged, polar acidic and polar basic.

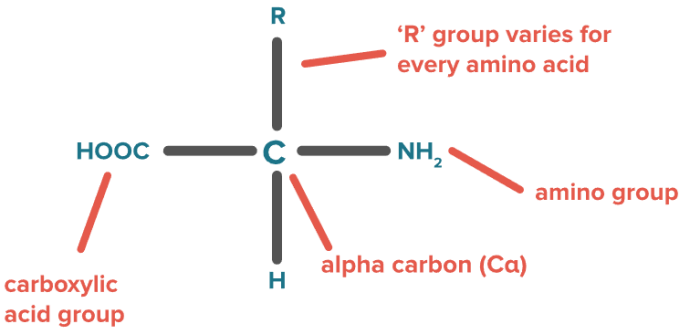


Figure 2.2 Structure of an amino acid

(Adapted from www.khanacademy.org).[178]

A protein is defined by a chain of **n** amino acid residues; the amino acids are joined by covalent peptide bonds between the amino group of residue **i**, and the carboxyl group of residue **i+1** (Figure 2.3).

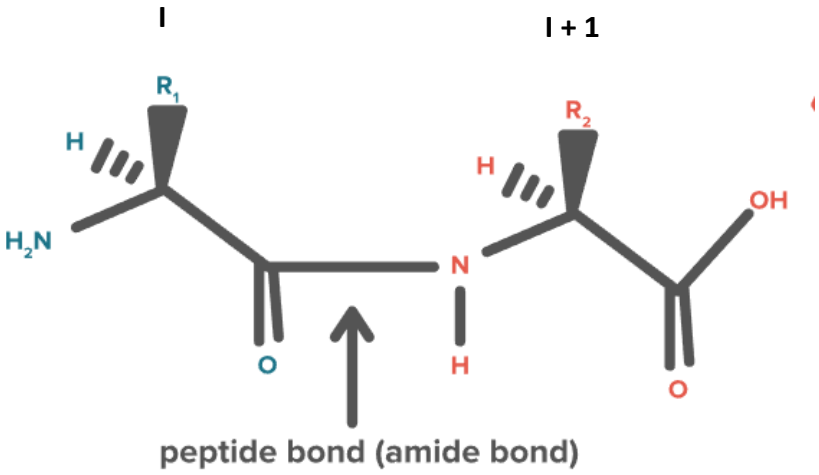


Figure 2.3 Peptide chain highlighting the peptide bond between amino acids

(Adapted from www.khanacademy.org).[178]

2.2.2 Secondary structure

The secondary structure of the protein consists of regular local substructures which are adopted by short chains of amino acids along the polypeptide backbone. These substructures are facilitated by the flexibility of the polypeptide chain.

The polypeptide remains intrinsically flexible because the majority of the bonds along the peptide backbone are rotationally permissible. The flexibility is only constrained by the peptide bond, which is rigid and planar. The rigid character of the peptide bond results from its ability to form multiple resonant structures in which electrons are differently distributed between atoms in the confirmation; this gives double bond character to the peptide bond.

The bonds before and after the peptide bond are free to rotate. This means that the backbone of the protein can be accurately described using two angles of rotation (Φ) and (Ψ) (Figure2.4).

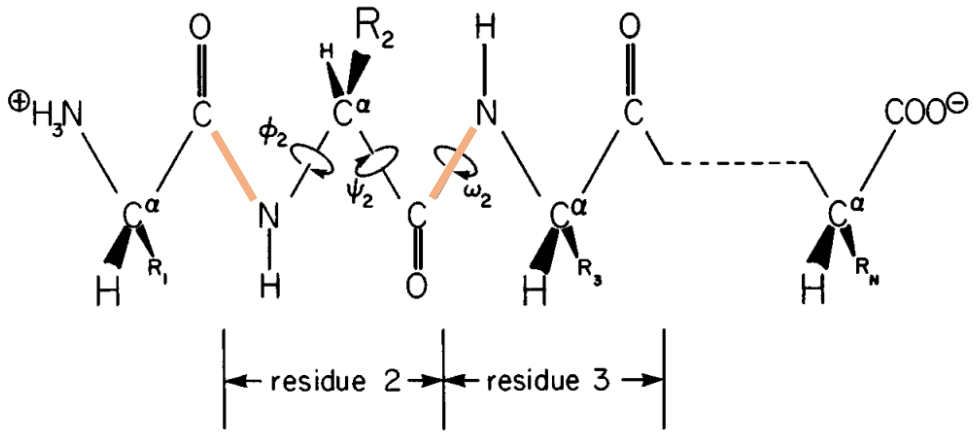


Figure 2.4 Structure of a polypeptide chain

Rotationally permissive bonds denoted (Φ and Ψ) and rigid peptide bonds coloured in orange; each amino acid residue is labelled below (Adapted from McCammon *et al.*).[179]

The second law of thermodynamics dictates that the energy within the system will always be minimised at equilibrium, as a low energy state is more energetically favourable. This means the protein will adopt confirmations which will minimise energy within the structure.[180]

The conformation adopted in the secondary structure is predominantly determined by the properties of the amino acid side chains and their interactions with the protein backbone. The protein adopts confirmations to minimise energy mainly through neutralising polar charges within the structure.

There are three main classes of secondary structure *viz.* α-helices, β-sheets (comprised of β-strands) and disordered loops (random coil). There are additional rare secondary structural elements, such as β-turns and β-barrels, but these are not frequently encountered and generally confined to specific protein classes.

α-helices are right-handed helical turns; where one turn is completed every 3.6 residues. The turn is created by hydrogen bonds which form between the carboxyl and amino group from residue I and I+4; meaning the helix must consist of at least 4 residues. The side chains of the amino acids all face outward along the turn (Figure 2.5).

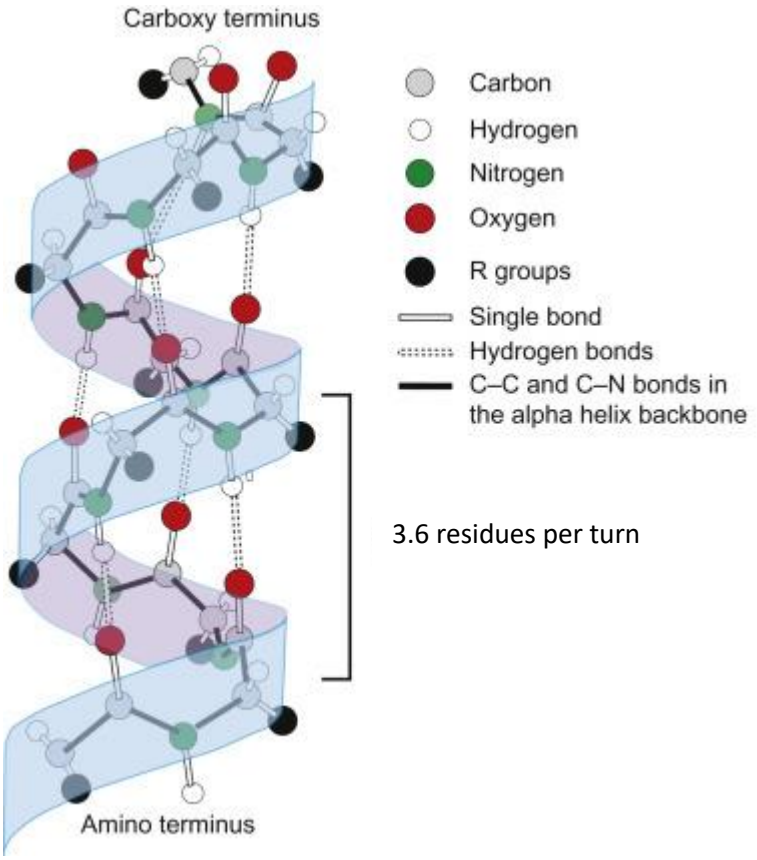


Figure 2.5. The conformation of the α helix

The dark bonds indicate those in the polypeptide backbone that are directly involved in the peptide bonds. The dashed lines indicate specific hydrogen bonds that bond successive turns of the helix with each other (Adapted from Feher *et. al*).[181]

B-sheets are formed from β-strands; short sections of the protein chain, usually between 5 and 10 amino acids in length, which adopt an extended straight conformation. Hydrogen bonds can form between the carboxyl and amino groups of aligning strands resulting in either a parallel conformation where the amino acids are positioned in the same orientation between strands and the hydrogen bonds are offset, or anti-parallel, where the amino acids are oriented in opposite directions and the hydrogen bonds form symmetrically (Figure 2.6).

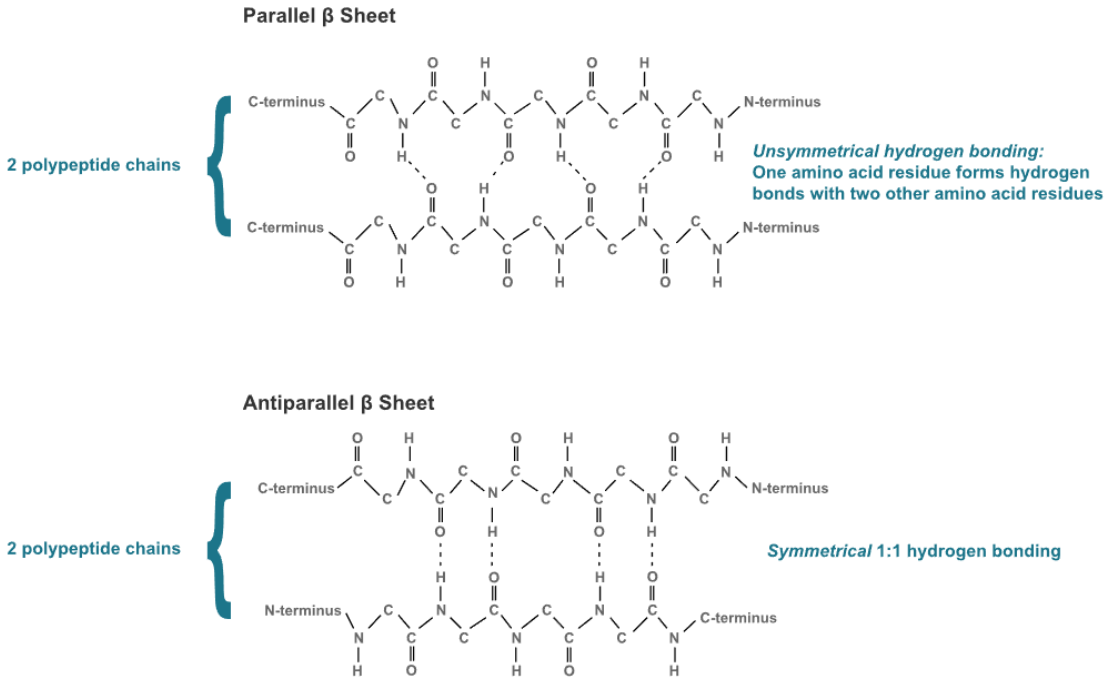


Figure 2.6 The conformation of parallel and anti-parallel β-sheets

(Adapted from www.khanacademy.org).[178]

Disordered loops do not have a specific structural definition, but rather are defined by the absence of any other definable secondary structure; they do not contain repeating backbone dihedral angles or regular patterns of hydrogen bonding. While helices and β-strands tend to remain structurally stable, loops remain flexible and adopt a multitude of potential confirmations. This is often facilitated by the presence of chains with a preponderance of short hydrophilic amino acids.

The characterisation of flexible loops (and other disordered regions) is of interest as it has become clear that native disorder can play important functional roles within the protein. In particular, they are believed to play a vital role in molecular recognition via the development of local metastable conformations around ligands, allowing the recognition of ligands with both high specificity and lower specificity.[4]

Disordered regions are less well understood than other secondary structures, primarily because their position within a protein crystal is variable. This means they are often difficult to visualise using crystallography; and indeed their presence may even hinder the crystallisation process itself.[182,183]

2.2.3 Tertiary structure

Protein folding involves the complex re-organisation of a protein into a three-dimensional tertiary structure (Figure 2.7). During this process, secondary structural components of proteins interact to form compact protein domains that remain independently stable and can fold in isolation.[184]

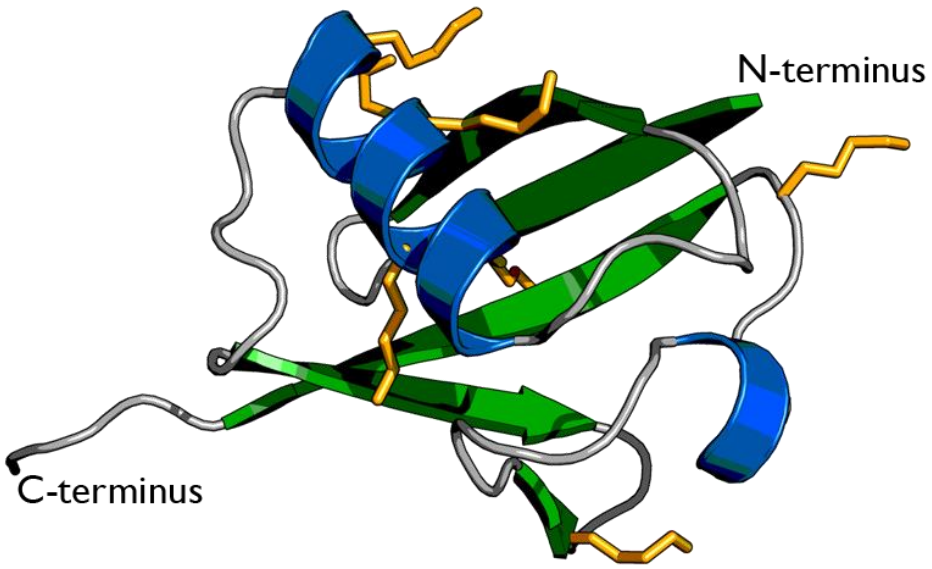


Figure 2.7 The tertiary structure of the protein

The protein shown is ubiquitin. Helices are shown in blue. Sheets are shown in green. The flexible loops are represented in grey. The lysine residues are highlighted in orange as a representation of the amino acid side chains.[185]

The native tertiary structure of a protein should be its most stable confirmation and energetically favourable in solution. However, there is an inherent flexibility to the structure, and it does not form a static rigid body.

The tertiary structure is maintained by a range of stabilising processes including hydrophobic and ionic interactions, Van der Waals forces, hydrogen bonds and disulphide bridges. Each of these interactions is due to the intrinsic properties of the amino acid side chains along the structure.

Hydrogen bonds and ionic interactions are facilitated by oppositely charges amino acid side chain groups. Wan der Waals forces occur as the weak electrical attraction between closely positioned atoms. Disulphide bridges can form between cysteine residues closely positioned in 3-dimensional space.

The hydrophobic residues in soluble proteins are maintained within the core of the protein, while the hydrophilic residues remain exposed to the surface. This increases the proteins' energetic favourability in solution, and is a key driver in the folding of proteins within a native solvent; which in biological systems is predominantly water. The folding of other classes of protein such as membrane proteins are also heavily influenced by hydrophobic interactions; however, within the membrane, hydrophobic residues would be exposed to the surface and hydrophilic residues protected within the protein core.[186]

The more of these stabilising factors present in the structure, the more ordered and more stable the protein.

2.2.4 Quaternary structure

Proteins do not exist in isolation, and many proteins such as COX-2, can form additional structures in combination with one another. The combined structure of interacting polypeptide chains is described by the quaternary structure.

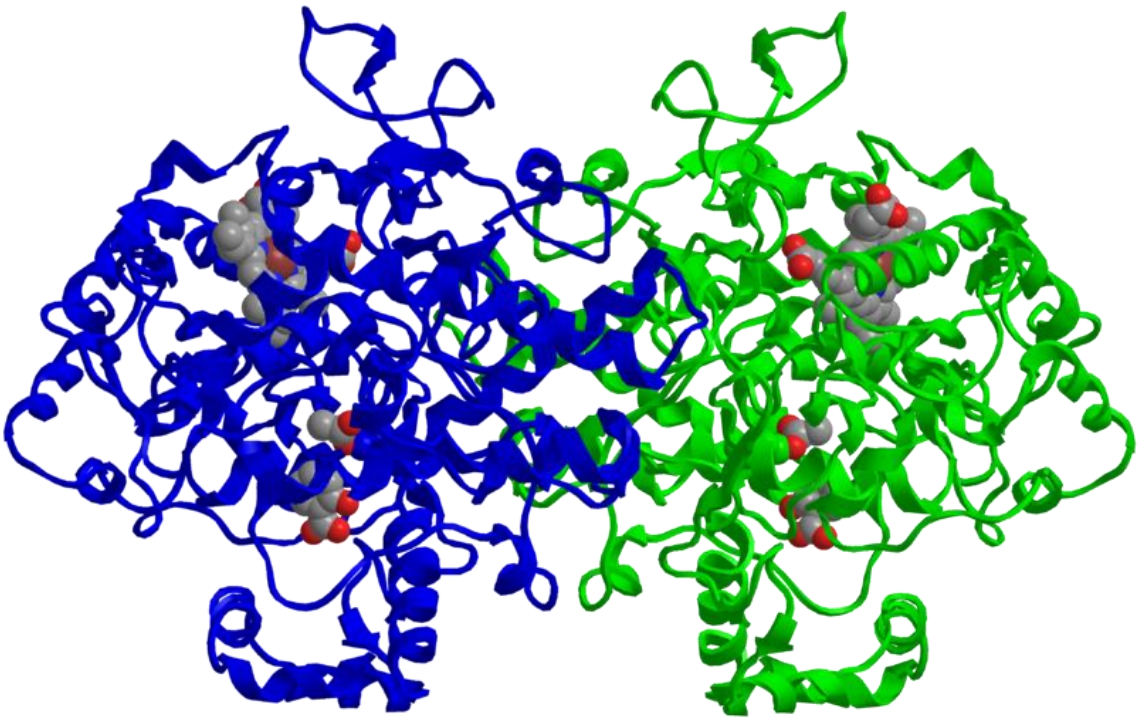


Figure 2.8. The quaternary structure of a protein

Structure of COX-2 (also called prostaglandin H synthase) inactivated by aspirin. The molecule is a dimer, the blue and green halves are identical. In each monomer, the active site serine has been acetylated, resulting in inactivation. Also visible in each monomer is the salicylic acid from which the acyl group came, and the heme cofactor (Adapted from Wikimedia.com).[187]

Quaternary structures are formed in the same manner as the tertiary structures and are governed by the same stabilising factors. The key difference being that the interactions occur between different polypeptide chains rather than within the same chain.

Quaternary structures can form between identical protein subunits to create homo-oligomers, or between different proteins to form hetero-oligomers. Quaternary structures could, in theory, comprise of any number of polypeptide subunits; many native proteins contain more than five polypeptide subunits. The additional level of complexity introduced by quaternary structures, make this level of structural detail the hardest to predict. Information on the quaternary structure of a protein and might provide insights into its function.

2.2.5 Phylogenetic relationship between proteins

The diversity of naturally occurring proteins is the result of cumulative modifications that have taken place over the history of evolution. Phylogenetics is the study of the evolutionary relationships between biological entities, in this instance proteins.[179]

When comparing evolutionary-related proteins, the spatial structure is more conserved than the sequence information. This means that in lieu of advanced sequence interrogation, inference about the structure and function of an unknown protein can be made based on sequence homology between proteins.[188]

In order to exploit the phylogenetic relationship between homologous proteins to characterise an unknown protein, the evolutionary history of the protein must first be explored. This is usually achieved by applying a blocks substitution matrix (BLOSUM) to the genetic sequence of the target protein and the genetic sequences of selected proteins. This allows a weighted evaluation of sequence changes to be undertaken, based on an assumption that certain substitutions are more readily accepted within a protein sequence and would thus occur more frequently and at faster rates over the evolutionary history.[189]

It is possible that proteins which share the same fold and overall architecture do not share a common ancestor and may have arisen independently. Thus, it could be argued that when undertaking this kind of structural and functional prediction, degrees of similarity are often more informative than true phylogenetic relationships.[190]

When sequence homology is low, using less restrictive search options can enable the identification of more structurally related homologues over long evolutionary distances.

However, care should be taken not to over annotate evolutionary related proteins which possess unique folds or significant architectural changes.[191]

While BLOSUM matrices are able to find both closely related and more distant homologues, for structural comparison between disparate homologues, the conserved domain architecture retrieval tool (CDART) is a better option; because it applies weighting to functionally important domains and folds within the protein.[192]

2.2.6 Interrogation of primary sequence information

Because of the propensity for each structural element within the protein to build from information contained in the previous level in the structural hierarchy, it is theoretically possible to obtain a full structural understanding of the protein solely from the primary sequence.

However, the complexity of these characteristics means that it is not possible to obtain a full description of the protein using first principles. Over recent years, the availability of computational resources and more extensive database of known protein structures has considerably facilitated this approach.

By using a combination of first principles and phylogenetic inferences, it is now possible to predict a wide range of protein characteristics with high accuracy.

2.2.6.1 Protein sequence databases:

The first step in determining the structure of an unknown protein is to obtain a primary sequence. Proteins generally exist as multiple isoforms containing a range of polymorphisms; thus, selecting a sequence can be challenging.

Protein sequences can be obtained from online databases, which maintain observed curated sequences that have been experimentally validated. The Uniprot database[121] for example, provides protein sequence information, together with other curated information such as structural annotations and functional information. This is an excellent, frequently-used resource for obtaining protein sequence information.

2.2.6.2 Whole protein qualities:

Information on some of the basic physicochemical parameters of a protein can be obtained from the primary structure alone using software packages such as ProtParam as part of the ExPASy webserver.[193] Examples of the information available include the summation of the amino acid properties for the complete protein; prediction of its isoelectric point based on the Henderson-Hasselbach relationship with average pKA values of ionisable side chains; its average molecular weight; the extinction co-efficient; and, a grand average hydrophobicity.[194,195]

In practice, these theoretical calculations will often differ slightly from subsequent experimental measurements, as they are based on average properties which do not account for additional intricate changes which can occur in solution or following post-translational modification. However, they are believed to be sufficiently accurate for practical application.[196]

2.2.6.3 Hydrophobicity:

The hydrophobicity of a protein can be determined simply by evaluating the hydrophobic and hydrophilic properties of the amino acid side chains. In order to identify regions of the protein which are more hydrophilic or hydrophobic, a moving window approach is implemented, to generate average hydropathy within given segments. By then plotting the grand average of hydropathy, it is possible to identify regions of the protein which are hydrophobic or hydrophilic. This would make it possible to predict which areas of the native protein are likely to be internal and which external.

There are a number of hydrophobic indices which can facilitate this process; most of these provide estimates of hydropathy which differ only marginally from those originally provided when it was pioneered by Kyte and Doolittle.[197]

2.2.6.4 Disorder prediction:

Disorder prediction is generally highly accurate and based on both the amino acid properties associated with the disordered regions but also the homology with structurally known disordered regions.

One of the most common disorder prediction tools, DISopred, uses a complex algorithm based on a matrix of known amino acid factors which lead to intrinsic disorder within a protein region.

For example, amino acids which are small, aromatic and hydrophilic are believed to allow greater flexibility and hence a greater likelihood of being disordered.

The algorithm also predicts disorder based on regions missing from known crystal structures of identified proteins, which are presumed to represent disordered segments. Homology with these regions is used to weight the probability of disorder in new proteins.[198] While this produces some bias in the data it has been shown to improve accuracy of disorder predictions by novel proteins and led to being one of the highest ranked disorder predictors in during CASP benchmarking.[199] PrDOS operates with similar underlying principles as DISopred, based on a unique protein training set.[200]

2.2.7 Secondary structure prediction

Secondary structure prediction is of particular importance, because it is the initial step in building an understanding of the overall protein structure. All additional inferences on tertiary and quaternary structures rely on the secondary structure prediction. This is evidenced by the number of structural and functional tools which input this information[201–204]

Secondary structure prediction is more challenging than predicting the properties of the whole protein (*vide supra*), as it requires inference based not only of total amino acid composition, but also of the local sequence environment. Although a number of methods have been developed to investigate the secondary structure of proteins, the most commonly applied is a software which uses a two-stage neural network approach, PSIPred[205], which is available to run using a webserver.[206]

PSIPred uses sequence profiles, derived from a specialised database of non-redundant sequences (intermediate PSI-BLAST sequence profiles), and applies feed-forward, back-propagation network processing to predict the secondary structure. Despite its relative simplicity to other packages, PSIPred has achieved an accuracy of around 80% accuracy in challenging performance benchmarking, which is very impressive; it outperforms most other approaches[205]

More advanced methods have been developed in an attempt to improve the prediction accuracy of PSIPred. One such method, based on machine learning *Dee*p *C*onvolutional *N*eural *F*ields (DEEPCNF). [207] This approach uses longer range sequence information and can model more complex sequence structure relationships with deep hierarchical architecture.

DEEPCNF achieved over 84% accuracy on average prediction, which outperforms PSIPRED. Despite this the data is heavily influenced by the training dataset, and in instances where sequence identity to other known proteins is low, the performance of the algorithm remains unpredictable.

Overall both PSIPred and DEEPCNF are valuable for determining proteins' secondary structure. If a protein shares over 25% homology with another known protein, DEEPCNF would now be the method of choice. However, when homology is low PSIPred arguably offers a more robust approach.

2.2.8 Transmembrane helix prediction:

Transmembrane helix prediction remains one of the most computationally challenging areas of computational prediction. This is, in part, due to paucity of known transmembrane protein structures available for use as a training set for machine learning algorithms.

There are a number of approaches available for predicting transmembrane helices. These generally rely very heavily on their small training sets and often perform poorly with novel structural proteins.

Three of the most common approaches to predicting transmembrane helices make use of neural networks, hidden Markov models or support vector machines. Neural networks are used by Octopus,[208] Spoctopus[209] and MEMSAT3.[210,211] Hidden Markov models by TMHMM,[212] Phobius[213] and support vector machines by MEMSAT-SVM.[214]

A combination of approaches should be used when working with novel protein in order to help identify false positive results.[215]

2.2.9 Domain boundaries

Proteins are often comprised of multiple autonomous domains but identification of the specific boundaries of each domain is challenging and often not absolute. The most accurate method for the correctly assigning domain boundaries would be the analysis of the three-dimensional structure of the protein. However, the necessary structural information is not available for a large number of proteins. Thus, other approaches have to be taken in order to identify domain boundaries.

DomPred uses both a homology and fold-recognition based approach. This applies a combination of overlapping alignments using the PSI-BLAST alignments and secondary structure alignments using the DOMSSEA method.[216,217]

Threadom uses a more advanced approach, relying on multiple threading alignments with known proteins. The underlying principle of this approach involves the assumption that true domain boundaries will only occur in regions where there are few overlapping threading alignments.[218]

2.2.10 Post translational modifications

Post-translational modifications are molecular alterations which do not alter the comprising amino acid chain but rather a specific residue. In most cases this involves the enzymatic addition of a biochemical entity. There are a wide array of post translational modifications; examples include, glycosylation, phosphorylation, acetylation, amidation, methylation, sulfation, sumoylation and ubiquitination.[219]

It is possible for different modifications to occur at the same site, but each amino acid residue is normally only subjected to a single posttranslational modification at any one time. It is very difficult to predict which residues might be amenable to modification because of the large number of potential modifications available.

However, in general, sites of posttranslational modification must: 1) lie in a recognition pattern to allow protein binding; and, 2) must be relatively exposed to allow addition of the molecule.[219] Some examples of predictive software for post translational modification based on these premises are SecretomeP 2.0,[220,221] Sumoplot,[222] FFPred 2.0.(NetPhos 3.1 and NetNGlyc 1.0c)[223,224], iUbiq-Lys,[225] UbPred[226] and PeptideCutter.[193]

2.2.11 Propensity to crystallise:

The bid to identify novel protein structures often fails at the crystallisation stage. A range of software has been developed in an attempt to predict the propensity of a given protein to crystallisation.[227] The two used in this thesis being XtalPred-RF[228] and Crysalis.[229]

XtalPred-RF uses a random forest machine learning algorithm to compare the predicted features of a target protein with a training set of other solved three-dimensional structures. The final algorithm is based upon 48 variables deemed most relevant to the crystallisation of a protein;

the most important of these were surface ruggedness, the length of the longest predicted disordered fragment and the overall percentage of serine residues in the protein sequence.[228]

Crysalis is also based on machine learning algorithms; however, it uses support vector regression to make predictions on success or failure at a range of experimental steps within the typical crystallography pipeline, including: sequence cloning failure, protein material production failure, purification failure, crystallization failure, and ultimately structure determination failure.[229]

The algorithm was trained using 4706 multifaceted sequenced-based features to investigate the most important characteristics, which were optimised to include 25, 78, 95, 68, and 65 variables for each stage of failure to minimise redundancy within the calculation.[229]

In addition to the characteristics discussed above, further prediction of the protein structure can be investigated using 3-Dimensional modelling of the protein and subsequent molecular dynamic simulations. These will be addressed in *Chapter 4: homology modelling of PNPLA3* and *Chapter 5: Molecular dynamics of PNPLA3*.

2.2.12 Functional prediction

In most cases, a protein is defined by its native function. Functional information allows the role of the protein in cellular processes and its wider role within the organism to be determined. To aid with functional classification, the Gene Ontology (GO) system provides well defined terms for the biological process, molecular function and cellular component of a protein; this is used to assist in consistent definition of function.[230]

An abundance of proteins have been generated by genomic sequencing but less than 8% of these have undergone functional assays.[231] This undoubtably reflects the high cost of directly determining function and hence processes to predict rather than measure activity have been developed.

In order to assess a protein's function, account must be taken of how the various levels of protein structure interact with one another and how changes in the sequence can alter function.

Platforms such as Gotcha[232] and ESG[233] can be used to predict the function of an unknown protein where there are a number of highly homologous proteins whose function is known. Under these circumstances these predictions tend to rely on knowledge of structural motifs and domains.[234] However, because function can diverge at as high as 70% sequence homology, the techniques are not applicable to proteins with low homology.[235]

To perform functional predictions of proteins with lesser homology, methods, such as those employed by FFPred 2.0 have been developed. These methods employ detailed feature-based comparisons with functionally annotated proteins, which facilitates a more accurate prediction of proteins with low homology. FFPred 2.0 for example draws comparisons based on amino acid composition, sequence features, transmembrane predictions, secondary structure, disordered regions and post translational modifications, which can all be more confidently predicted based on sequence information alone (*vide supra*).[223]

2.2.13 Sequence based investigations into PNPLA3

Several features of PNPLA3 have been predicted previously using homology-based comparisons. It was through sequence identity that the conserved patatin domain, which spans approximately 170 amino acids from residue 10 to 179 was classified, and the prediction that the protein may function as a lipase *in vitro*.[125]

PNPLA3 is an integral member of the Patatin like phospholipases; nine members of the PNPLA family are found in man,[149] Within the family PNPLA3 is said to be most closely related to the lipase PNPLA2 .[150]

Structural modelling of a partial patatin domain of PNPLA3 based on Pat17, suggest a conserved patatin fold within the protein architecture, in which a conserved α/β patatin-like fold and spatial orientation of the catalytic dyad are maintained.[149,158]

However, other than early studies into the potential for membrane interaction where PNPLA3 was computationally predicted to be a transmembrane protein, most studies have focused on achieving experimentally derived characteristics. This means the underlying properties of the protein and predicted behaviours have undergone little investigation.[125]

2.3 Aims

The overarching aim of this chapter is to gain insight into the structural and functional characteristics of PNPLA3 using bioinformatic tools to analyse its primary amino acid sequence.

This will involve two parallel approaches: 1) Using a phylogenetic analysis to derive knowledge from related proteins. 2) Using software to predict protein properties and further structural information.

Insights gained from this investigation can later be used to guide experimental conditions and enhance further *in vitro* investigation into the protein.

2.4 Methods

A number of bioinformatic tools were used to characterise PNPLA3 based on increasing levels of sequence complexity (Table 2.1). Most of these tools produce an output of a specific phenotypic descriptor along with a score to estimate the confidence of the result. While scores are generally presented as a probability ranging from 0 to 1. No attempt will be made to directly compare scores between programs, due to underlying differences in their production.

Table 2.1. Bioinformatic tools used to characterise PNPLA3

Tools	Function
Uniprot	Sequence database
CDART; SMART, CDD, SPARCLE	Conserved domain analysis
ExPASy; ProtParam,	Whole protein property prediction
ExPASy; ProtScale	Hydrophobicity prediction
DISopred	Disorder prediction
PrDOS	
DomPred; DOMSSEA	Domain boundary prediction
ThreaDom	
PSIPred; WESA	Secondary structure prediction
TMHMM	Membrane helix prediction
Phobius	
Octopus	
Spoctopus	
MEMSAT3	
MEMSAT-SVM	
Secretome P2.0	Post translational modification prediction
Sumoplot	
iUbiq-Lys	
UbPred	
ExPASy; PeptideCutter tool	Cleavage site prediction
Xtal-Pred-RF	Crystallisation potential prediction
Crysalis	
FFPred 2.0	Function and localisation prediction

Amino acid sequences:

The protein sequences of human PNPLA3 (Uniprot Accession: Q9NST1) and murine pnpla3 (Uniprot Accession: Q91WW7; Appendix I) were retrieved from the Uniprot database.[121] These sequences were used as the basis for the primary sequence-based investigations (Figure 2.9).

```
>sp|Q9NST1|PLPL3_HUMAN Patatin-like phospholipase domain-containing
protein 3 OS=Homo sapiens GN=PNPLA3 PE=1 SV=2
MYDAERGWSLSFAGCGFLGFYHVGATRCLSEHAPHLLRDARMLFGASAGALHCVGVLSGI
PLEQTLQVLSDLVRKARSRNIGIFHPSFNLSKFLRQGLCKCLPANVHQLISGKIGISLTR
VSDGENVLVSDFRSKDEVVDALVCSCFIPFYSGLIPPSFRGVRYVDGGVSDNVPFIDAKT
TITVSPFYGEYDICPKVKSTNFLHVDITKLSLRLCTGNLYLLSRAFVPPDLKVLGEICLR
GYLDAFRFLEEKGICNRPQPGLKSSSEGMDPEVAMPSWANMSLDSSPESAALAVRLEGDE
LLDHLRLSILPWDESILDTLSPRLATALSEEMKDKGGYMSKICNLLPIRIMSYVMLPCTL
PVESAIAIVQRLVTWLPDMPDDVLWLQWVTSQVFTRVLMCLLPASRSQMPVSSQQASPCT
PEQDWPCWTPCSPKGCPAETKAEATPRSILRSSLNFFLGNKVPAGAEGLSTFPSFSLEKS
L
```

Figure 2.9 Human PNPLA3 protein sequence retrieved from Uniprot database

The top line includes summary information relating to the sequence identity, the origin species, the gene name, evidence for protein existence and the sequence version. Each following letter corresponds to a single amino acid within the protein sequence.

2.4.1 Phylogenetic analysis

Similar proteins were searched and grouped by similarity using Conserved Domain Architecture Retrieval Tool (CDART),[192] and by searching the Simple Modular Architecture Research Tool (SMART)[236], Pfam[145,237] and the Conserved Domain Database (CDD).[238–241] Protein architecture was labelled using the Subfamily Protein Architecture Labelling Engine (SPARCLE).[239]

Phylogenetic trees were generated using CDTree.[241]

2.4.2 Protein property prediction

Whole protein properties were calculated using the protein identification and analysis tool ProtParam within the ExPASy webserver.[193]

Hydrophobicity was predicted using ProtScale on the ExPASy webserver. The Kyte and Doolittle hydrophobicity index was used with a 19 residue window size.[193]

Predictions of native protein disorder were determined with DISopred,[198] DOMSSEA[198] and PrDOS.[200]

Domain boundaries were predicted using DomPred[216] and ThreaDom.[218]

2.4.3 Secondary structure prediction

Secondary structure was predicted using PSIPred[205] using the online webserver.[206] Solvent exposed residues were predicted using the Weighted ensemble solvent accessibility predictor(WESA) module of the Pi2PE engine.[242]

Membrane helices and associations were predicted with TMHMM,[212] Phobius,[213] Octopus,[208] Spoctopus,[209] MEMSAT3,[210,211] MEMSAT-SVM,[214] and memtype-2L.[243]

Membrane topology and pore regions through the membrane were predicted with PoreWalker[244] and TMKink[245] (results not shown).

2.4.4 Post translational modifications

Both mammalian[220] and bacterial[221] non-normal secretion signals were predicted with SecretomeP 2.0.

Sites of potential sumoylation were predicted with Sumoplot.[222]

Phosphorylation and glycosylation sites were predicted using the NetPhos 3.1 and NetNGlyc 1.0c packages within FFPred 2.0.[223,224]

Ubiquitination sites were predicted using iUbiq-Lys[225] and UbPred.[226]

Cleavage sites were predicted using the PeptideCutter tool on the ExPASy webserver.[193]

2.4.5 Functional prediction

Potential protein functions were predicted based on sequence homology using FFPred 2.0.[223]

2.4.6 Propensity to crystallise

The probability of crystallisation was predicted using XtalPred-RF [228] and Crysalis.[229]

2.5 Results

2.5.1 Phylogenetic analysis

In the conserved domain database, PNPLA3 is predicted to contain one classified domain namely an N-terminal Patatin_and_cPLA2 domain. CDART detected 61,548 non-redundant protein sequence homologues in cellular organisms based on homology with this domain, as well as an additional 20,552 protein sequences based on similar related domains.

Two specific conserved regions of functional importance are observed between sequence alignments of diverse species from the patatin_and_CPLA2 superfamilies (Figure 2.10). The first, a catalytic dyad, is formed by a conserved serine and aspartate; in which the positions of the aspartate varies widely across sequences due to large insertions. The second, an oxyanion hole is formed by two glycine residues and one arginine residue and always occurs early in the sequence (Figures 2.11).

A phylogenetic tree was constructed from these alignments; the Pat_PNPLA_like branch (containing PNPLA1-5), is more closely related to the Pat_TGL3-4-5_SDP1 branch (containing proteins such as PLPL and SDP1), than the branch containing PNPLA6 and 7 (Figure 2.12).

Sequence alignments of the patatin branch show some increased areas of conservation across the proteins; thus, alignments appear longer and with less insertions and deletions present (Figures 2.13).

The Pat_PNPLA_like family of proteins is the shortest branch and contains Pat_iPLA2, Pat_like, Pat_PNPLA1, Pat_PNPLA2, Pat_PNPLA3, Pat_PNPLA4 and Pat_PNPLA5_mammals (Figure 2.14).

The mammalian PNPLA5 proteins, which lie on the most closely related branch, are the ones most similar to PNPLA3. Next in terms of similarity are PNPLA2, IPLA2, PNPLA1, PNPLA4 and Pat_like respectively.

Sequence alignments of members of these families show nearly complete conservation of sequences over the first 250 residues with very few insertions and deletions present (Figure 2.15).

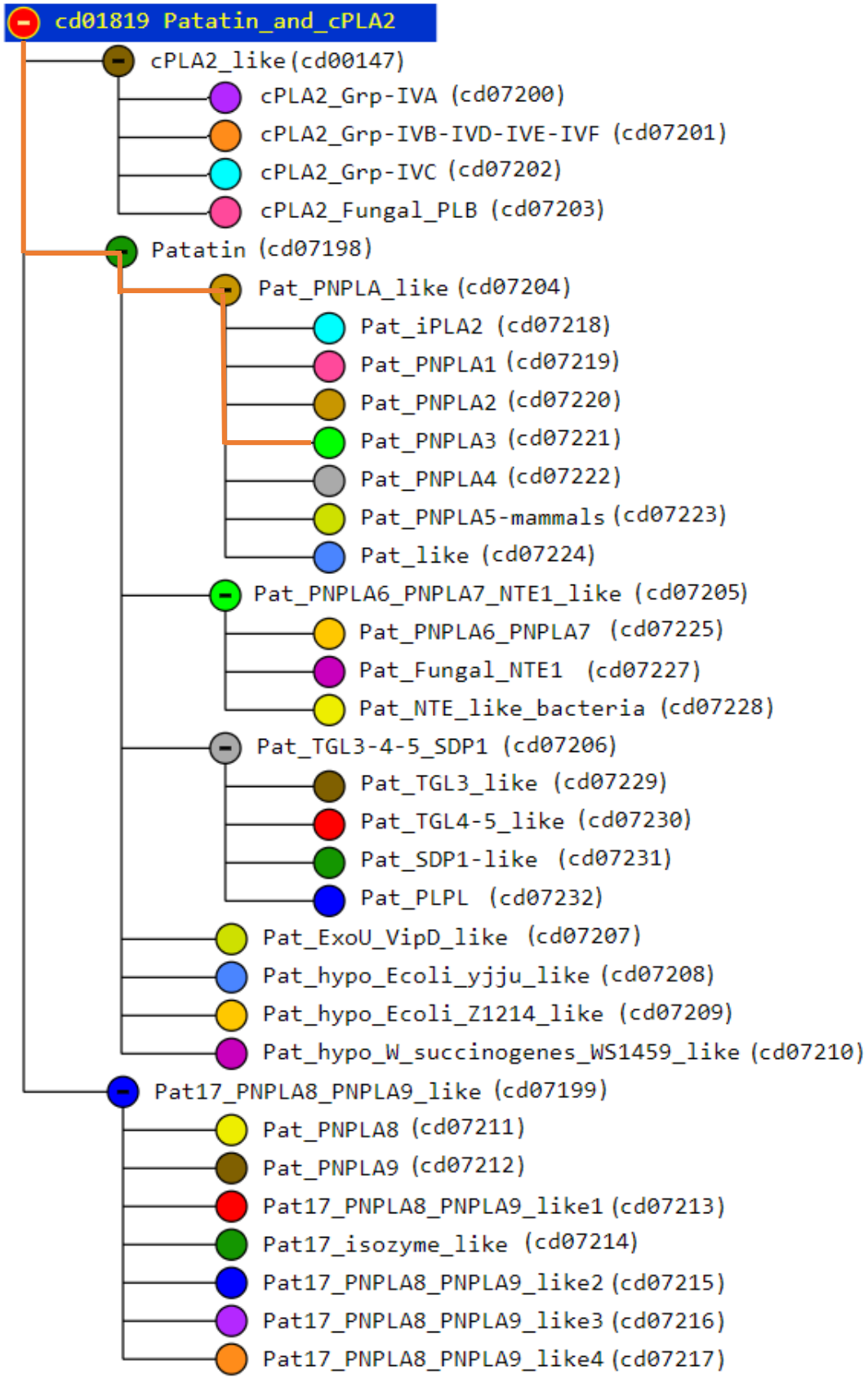


Figure 2.10 Phylogenetic sequence tree of the Patatin and CPLA2 superfamily

Alignments were generated from curated conserved domains from the CDD. The branches were grouped based on pair-wise alignment scores, in which aligned residue pairs are scored with the BLOSUM62 matrix and grouped into supergroups of proteins. The orange line highlights the path to the PNPLA3 branch, which will be explored in following figures.

Figure 2.11 Sequence alignment of the most diverse members from the cluster of Patatin and CPLA2 superfamily sequences (cont'd next page)

Aligned residues are shown in upper case, unaligned residues in lower case, and variation in sequence length shown as dashes. Red is used to indicate highly conserved and blue to indicate less conserved residues; unaligned (lower case) residues are shown in grey. The numbers at the beginning and end of each sequence row indicate the span of the sequence data imported from the complete protein sequence record. Hash-marks (#) in the top row of a multiple sequence alignment indicate residues involved in a conserved feature, such as a binding or catalytic site, that has been annotated on an NCBI-curated domain.

Figure 2.11 (Continued)

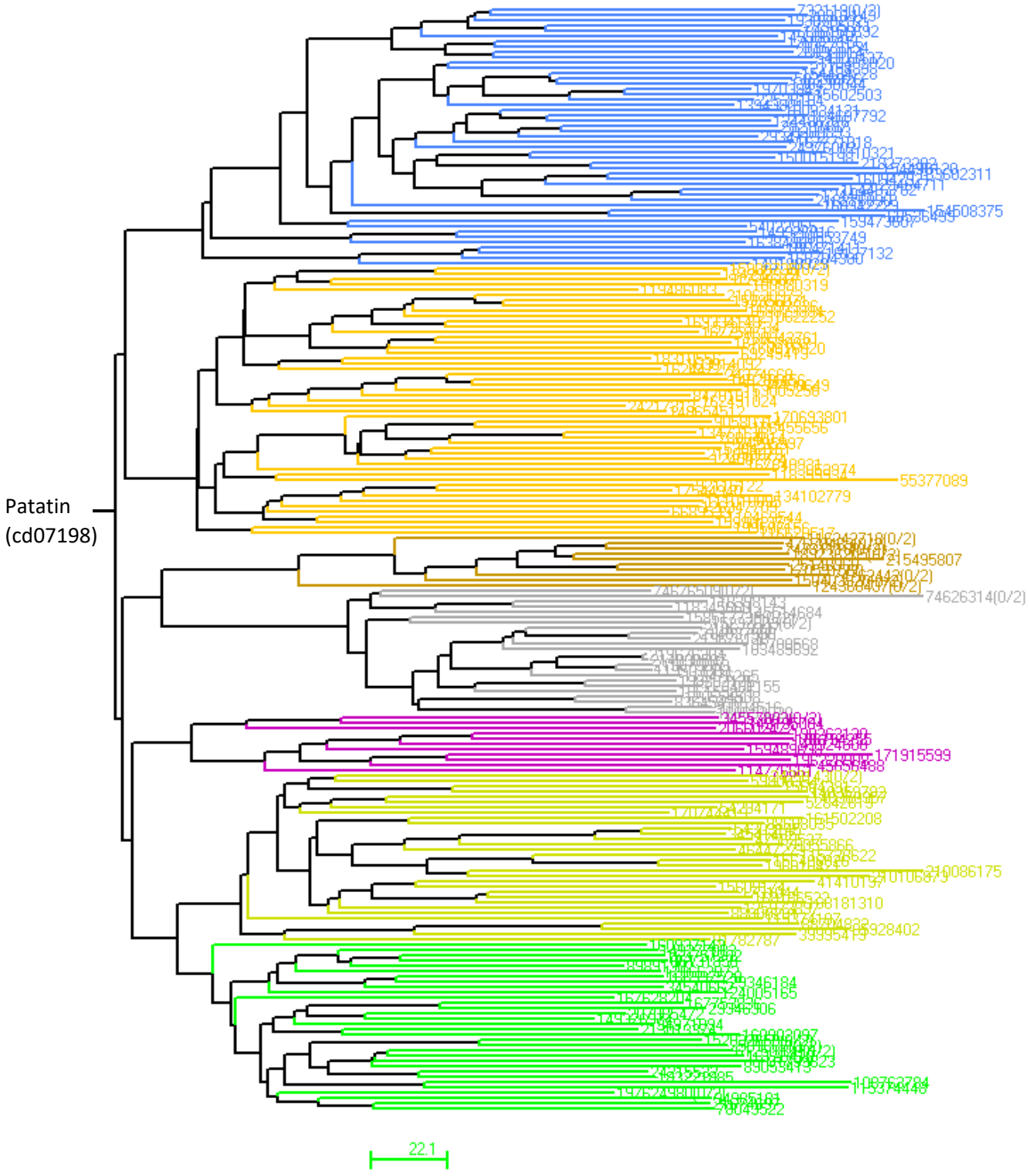


Figure 2.12 Distance based phylogenetic sequence tree of the Patatin (cd07204) family of proteins

The distance data have been obtained from pair-wise alignment scores, scored with the BLOSUM62 matrix. Groups of branches rendered in the same colour correspond to alignment rows assigned to a particular subgroup.

Blue: Pat_hypo_Ecoli_yjju_like_cd07208;
Gold: Pat_hypo_Ecoli_X1214_like_cd07209;
Brown: Pat_PNPLA_like_cd07204;
Grey: Pat_TGL3-4-5_SDP1_cd7206;
Pink: Pat_hypno_W_succinogene_cd07210;
Lime green: Pat_ExoU_CipD_like_cd07207;
Bright green: Pat_PNPLA6_PNPLA7_NTE_cd07205.

Figure 2.13 Sequence alignment of the most diverse members from the cluster of Patatin (cd07204) sequences (cont'd next page)

Aligned residues are shown in upper case, unaligned residues in lower case, and variation in sequence length shown as dashes. Red is used to indicate highly conserved and blue to indicate less conserved residues; unaligned (lower case) residues are shown in grey. The numbers at the beginning and end of each sequence row indicate the span of the sequence data imported from the complete protein sequence record. Hash-marks (#) in the top row of a multiple sequence alignment indicate residues involved in a conserved feature, such as a binding or catalytic site, that has been annotated on an NCBI-curated domain.

Figure 2.13 (continued).

Of the PNPLA3 proteins, the two which are closest and share the most recent common ancestor are those in *H. sapiens* and *M. musculus*. The next most similar are the PNPLA3 proteins in *E. caballus* and *C. lupus familiaris*.

Despite variation in the degree of similarity between the homologous proteins, the PNPLA3 sequence branch is highly conserved over the initial 250 residues (Figures 2.16 and 2.17). The terminal residues do not have a strong alignment between species.

All the proteins have a catalytic serine and aspartate, and the I148 is highly conserved, only observed in the diverse sequences in a hypothetical PNPLA3 of *E. caballus*.

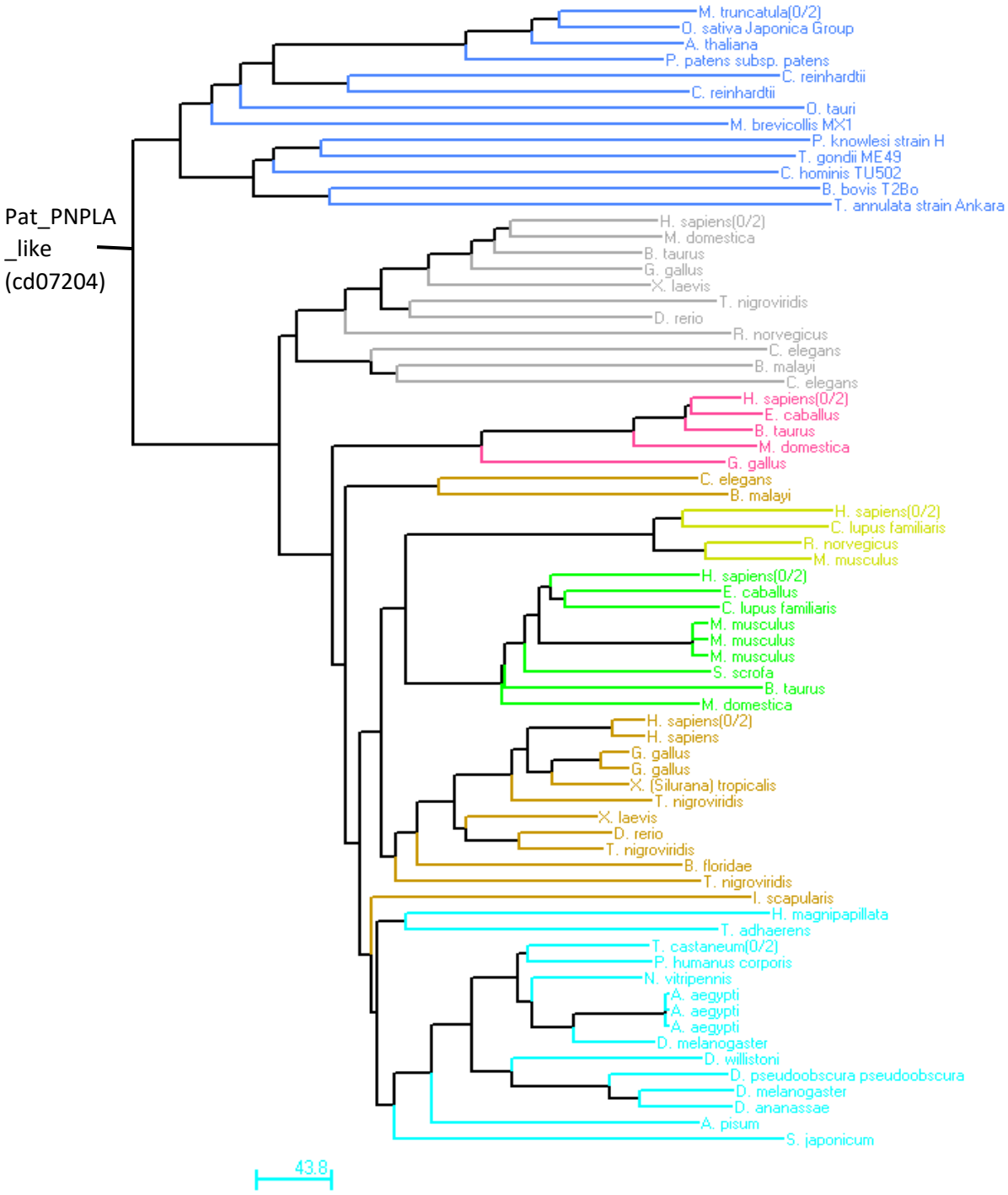


Figure 2.14 Distance based phylogenetic sequence tree of the Pat_PNPLA_like (cd07204) family of proteins

The distance data have been obtained from pair-wise alignment scores, scored with the BLOSUM62 matrix. Groups of branches rendered in the same colour correspond to alignment rows that have been assigned to a particular subgroup.

Dark blue: Pat_like_cd07224;
grey: Pat_PNPLA4_cd07222 ;
Pink: Pat_PNPLA1_cd07219;
orange: Pat_PNPLA2_cd07220;
lime green: Pat_PNPLA5_cd07223;
green: Pat_PNPLA3_cd07221;
light blue: Pat_iPLA2_cd07218.

Figure 2.15 Sequence alignment of the most diverse members from the cluster of Pat_PNPLA_like (cd07204) sequences

Aligned residues are shown in upper case, unaligned residues in lower case, and variation in sequence length shown as dashes. Red is used to indicate highly conserved and blue to indicate less conserved residues; unaligned (lower case) residues are shown in grey. The numbers at the beginning and end of each sequence row indicate the span of the sequence data imported from the complete protein sequence record. Hash-marks (#) in the top row of a multiple sequence alignment indicate residues

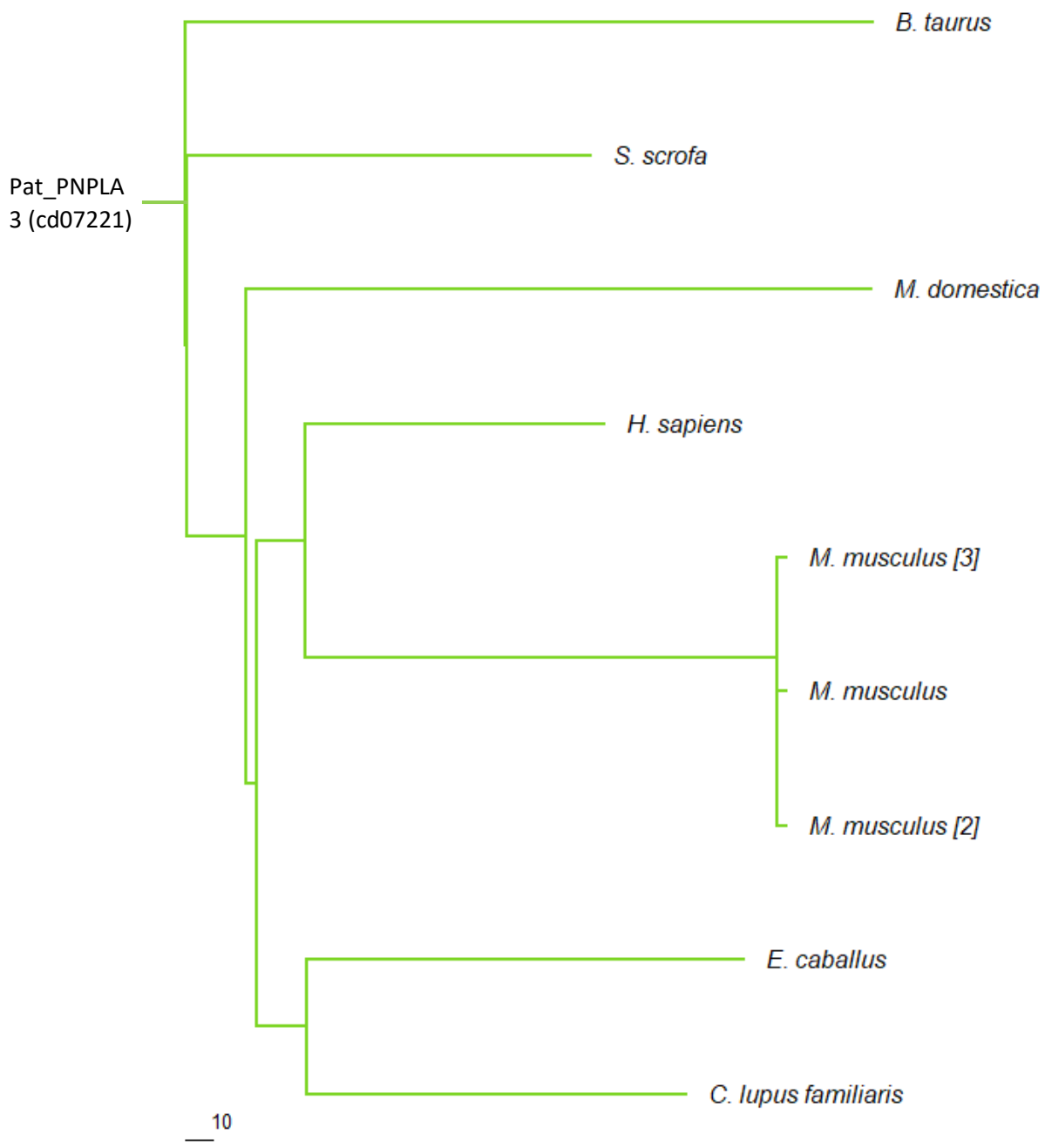


Figure 2.16 Distance based phylogenetic sequence tree of the PNPLA3 (cd07221) family of proteins

The distance data have been obtained from pair-wise alignment scores, scored with the BLOSUM62 matrix.

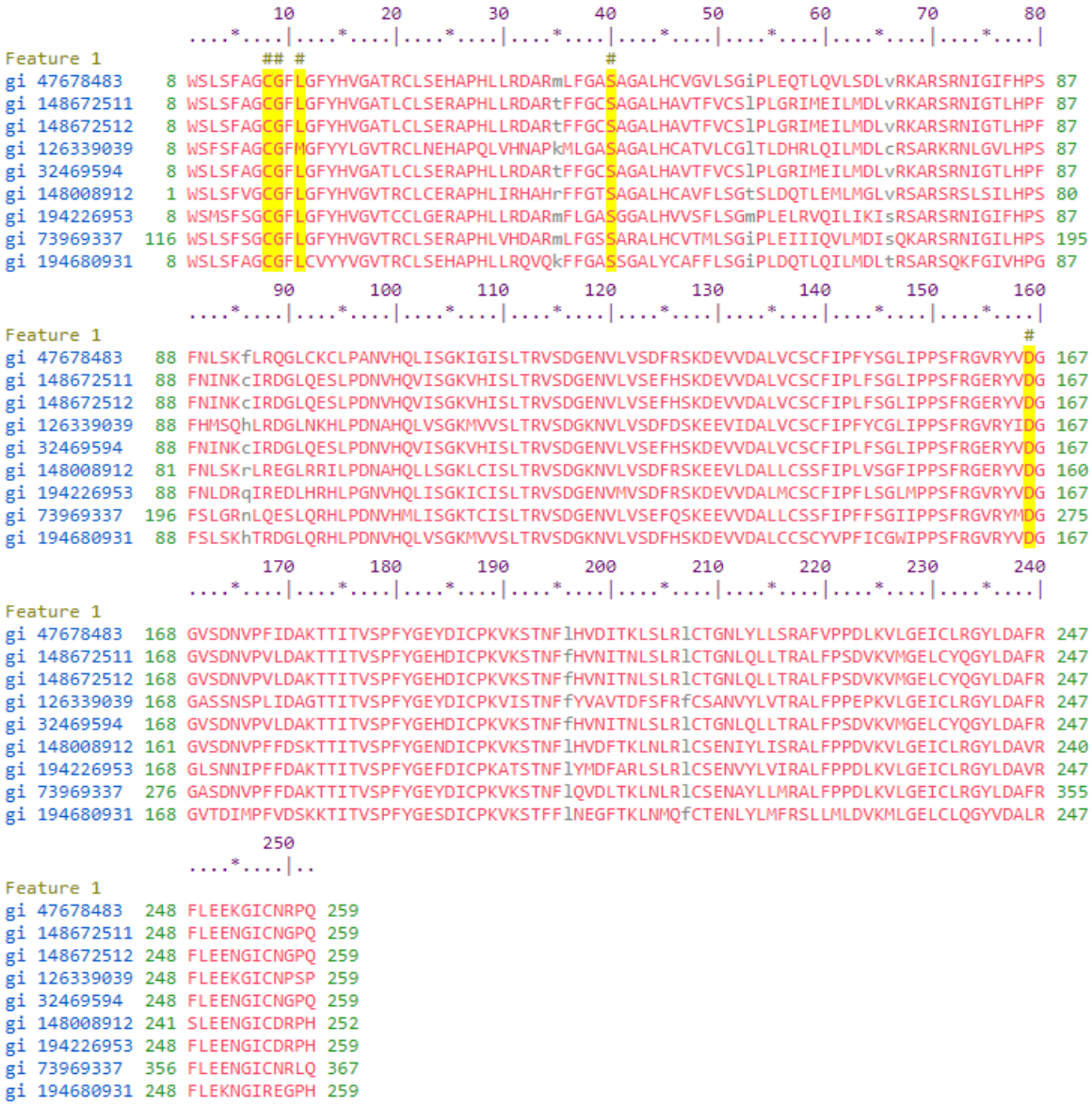


Figure 2.17 Sequence alignment of the most diverse members from the cluster of PNPLA3 (cd07221) sequences

Aligned residues are shown in upper case, unaligned residues in lower case, and variation in sequence length shown as dashes. Red is used to indicate highly conserved and blue to indicate less conserved residues; unaligned (lower case) residues are shown in grey. The numbers at the beginning and end of each sequence row indicate the span of the sequence data imported from the complete protein sequence record. Hash-marks (#) in the top row of a multiple sequence alignment indicate residues

2.5.2 Protein property prediction

2.5.2.1 Amino acid composition:

PNPLA3 has an even distribution of amino acids; none is above expected levels. The highest levels observed are leucine and serine with make up 13 and 12 percent of the protein respectively. Over 60% of the residues are hydrophobic, and 40% polar, with low numbers of aromatic residues (Figure 2.18).

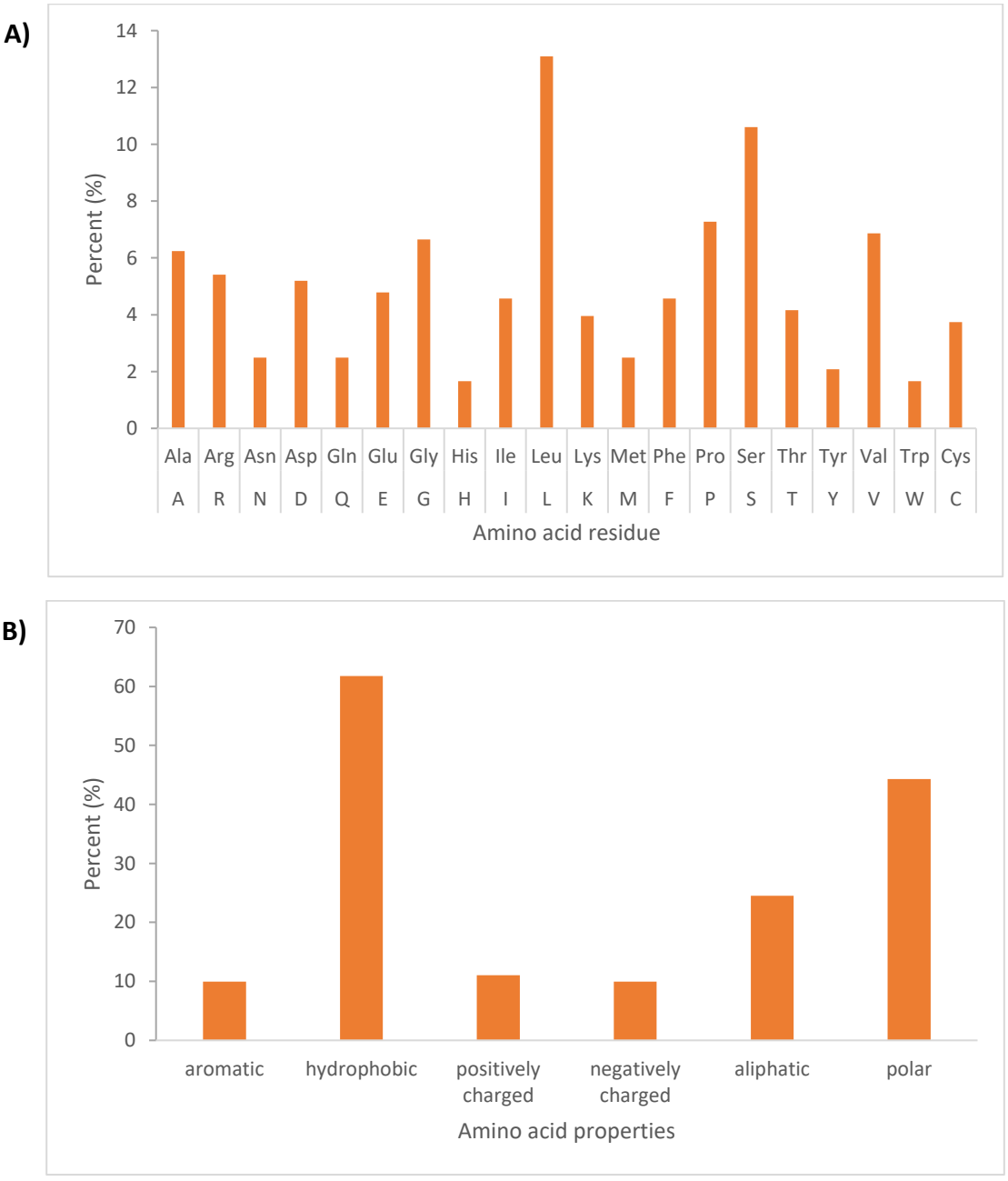


Figure 2.18 The amino acid profile of PNPLA3

A) The percentage abundance of each amino acid in the protein.
B) The percentage abundance of each amino acid property in the protein.

PNPLA3 has a molecular weight of 52,865.3 Da, consists of a total of 7,443 atoms and has a theoretical isoelectric point of 6.27.

Its theoretical extinction coefficient, assuming the cysteines are oxidised is 60025 M-1 cm-1, and the Abs 0.1%, is 1.135. Under fully reduced conditions the extinction coefficient and Abs 0.1% are 58900 and 1.114 respectively.

The estimated half-life of PNPLA3, is predicted to be 30 hours in mammalian cells, above 20 hours in yeast and above 10 hours in *E. coli*. The instability index is 54.94. The aliphatic index is 95.05 and the grand average of hydropathicity 0.100.

The hydropathicity of PNPLA3 is predicted to vary along the length of the protein. There are four particularly large hydrophobic peaks between residues 40-60, 135-160, 340-370 and 390 to 410. In addition, there are two large hydrophilic peaks between residues 245-270 and 410-440 (Figure 2.19). The whole C-terminal half of the protein is more hydrophilic than the N-terminal portion, apart from the large section between residues 350 and 400.

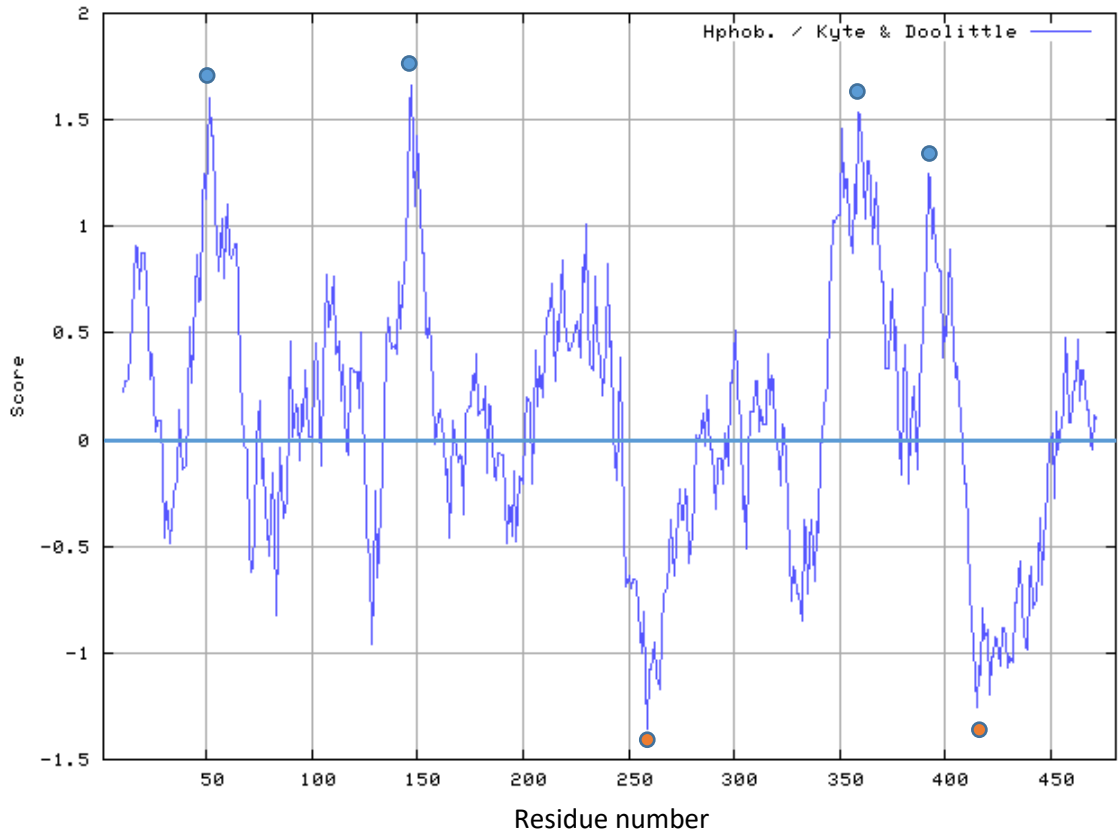


Figure 2.19 Hydrophobicity prediction of PNPLA3 using ProtScale normalised to the grand average of hydrophobicity of the protein database

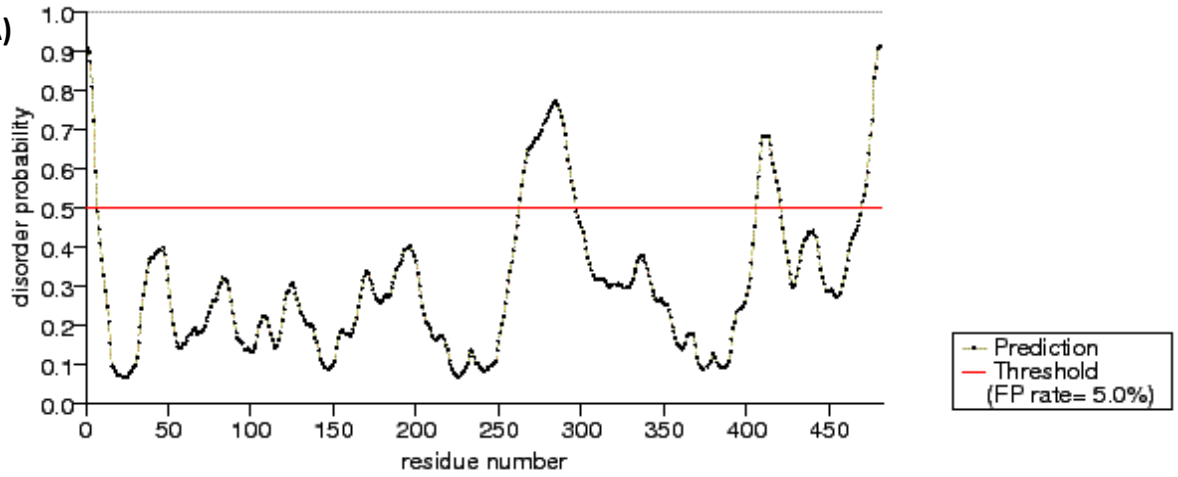
Scores above zero show greater hydrophobic character, scores below zero show more hydrophilic character. **Blue circles:** key hydrophobic peaks. **Red circles:** Key hydrophilic peaks.

2.5.2.2 Disorder prediction:

Both DISopred and PRDOS predict that there is a high probability of a disordered region in the initial 5 N-terminal residues and spanning residues 250-300.

There is an additional region with a high predicted probability of disorder at the C-terminal end of the protein; DISopred predicts this to spans residues 400-481, whereas PRDOS predicts two shorter regions of disorder spanning residues 400-425 and 450-481 (Figures 2.20 and 2.21).

A)



B)

1	**MYDAER**GWSL	SFAGCGFLGF	YHVGATRCLS	EHAPHLLRDA	RMLFGASAGA	50
51	LHCVGVLSGI	PLEQTLQVLS	DLVRKARSRN	IGIFHPSFNL	SKFLRQGLCK	100
101	CLPANVHQLI	SGKIGISLTR	VSDGENVLVS	DFRSKDEVVD	ALVCSCFIPF	150
151	YSGLIPPSFR	GVRYVDGGVS	DNVPFIDAKT	TITVSPFYGE	YDICPKVKST	200
201	NFLHVDITKL	SLRLCTGNLY	LLSRAFVPPD	LKVLGEICLR	GYLDAFRFLE	250
251	EKGICNRPQP	GL**KSSSEGMD**	**PEVAMPSWAN**	**MSLDSSPESA**	**ALAVRL**EGDE	300
301	LLDHLRLSIL	PWDESILDTL	SPRLATALSE	EMKDKGGYMS	KICNLLPIRI	350
351	MSYVMLPCTL	PVESAIAIVQ	RLVTWLPDMP	DDVLWLQWVT	SQVFTRVLMC	400
401	LLPAS**RSQMP**	**VSSQQASPCT**	PEQDWPCWTP	CSPKGCPAET	KAEATPRSIL	450
451	RSSLNFFLGN	KVPAGAEGL**S**	**TFPSFSLEKS**	**L**		500

Figure 2.20 Intrinsic disorder prediction of PNPLA3 using PRDOS

A) The disorder probability is represented as a dotted line on the graph.
B) The amino acid sequence of PNPLA3, with regions of disorder over 0.5 confidence highlighted in red.

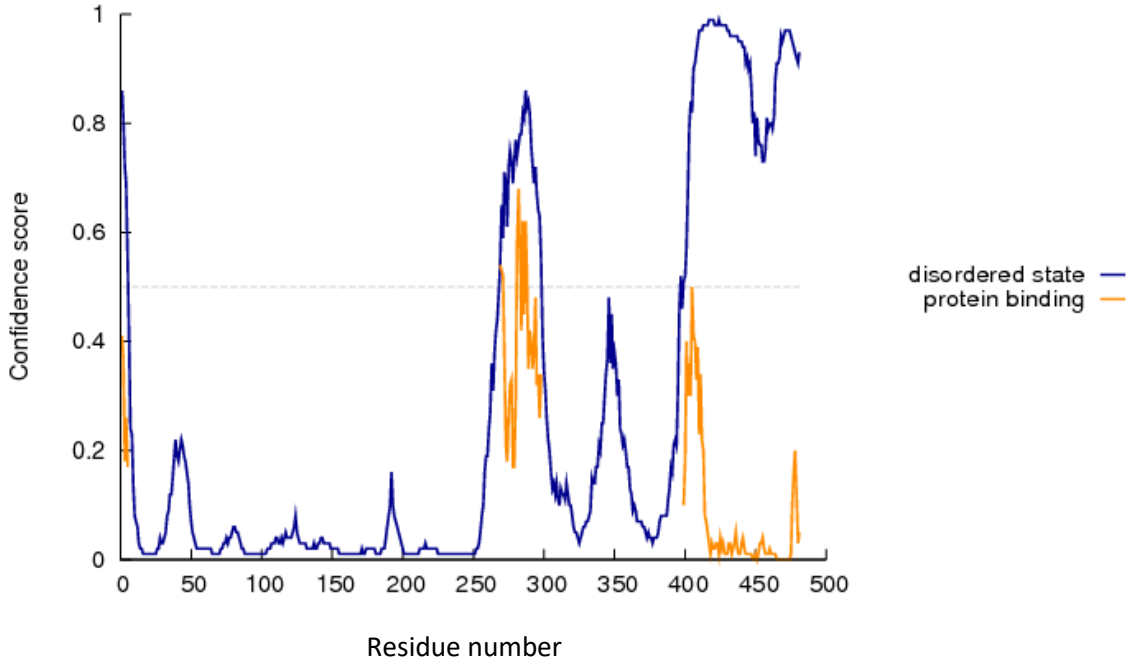


Figure 2.21 Intrinsic disorder prediction of PNPLA3 using DISopred

2.5.2.3 Domain boundaries:

DomPred predicts two domain boundaries. One between residues 160-200 and a further potential boundary between residues 240-260 which was less frequently detected (Figure 2.22).

DomSSEA predicts nine domain boundaries although even the most probable, situated at residue 61, had an alignment score of only 0.67 (Table 2.2).

ThreaDom does not predict the presence of any domain boundaries (Figure 2.23).

Table 2.2 Potential PNPLA3 domain boundaries as determined by multiple protein alignments using DOMSSEA

DomSSEA Results			
Score	**Match**	**No. Doms**	**Boundaries**
0.6722068	1rq5A	2	61
0.6710214	1rfqB	2	201
0.66587955	1ldkB	2	321
0.6604651	3bccC	2	286
0.66046	1qzwG	3	125, 321
0.6581395	2bccC	2	286
0.6536585	1b43B	2	241
0.6492537	1ma7A	2	105
0.64881694	1ouqB	2	134

*Score: normalised alignment score (0-100);[246] Match: PDB code of matched alignment;

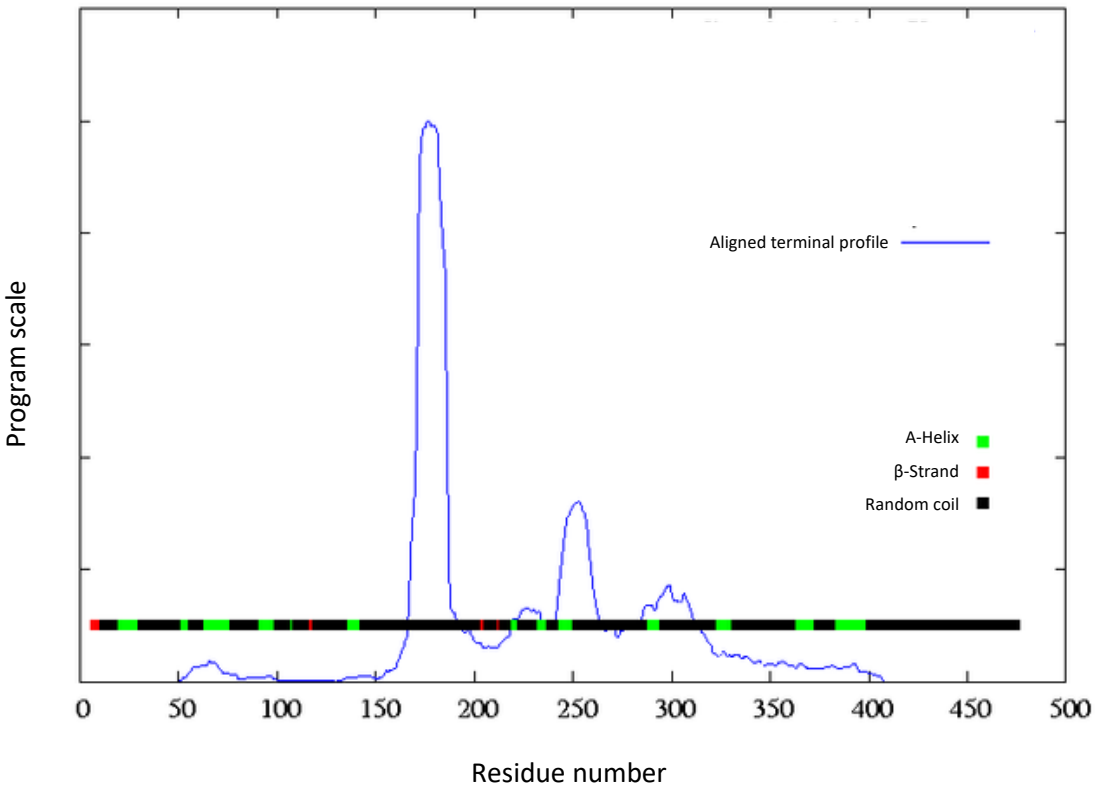


Figure 2.22 Domain boundary prediction of PNPLA3 using DomPred

The aligned terminal profile of PSI-BLAST results is represented by the blue line, reflecting potential domain boundaries. The secondary structure prediction is represented by a continuous bar along the X-axis, showing the predicted sites of helices, β-stands and random coils.

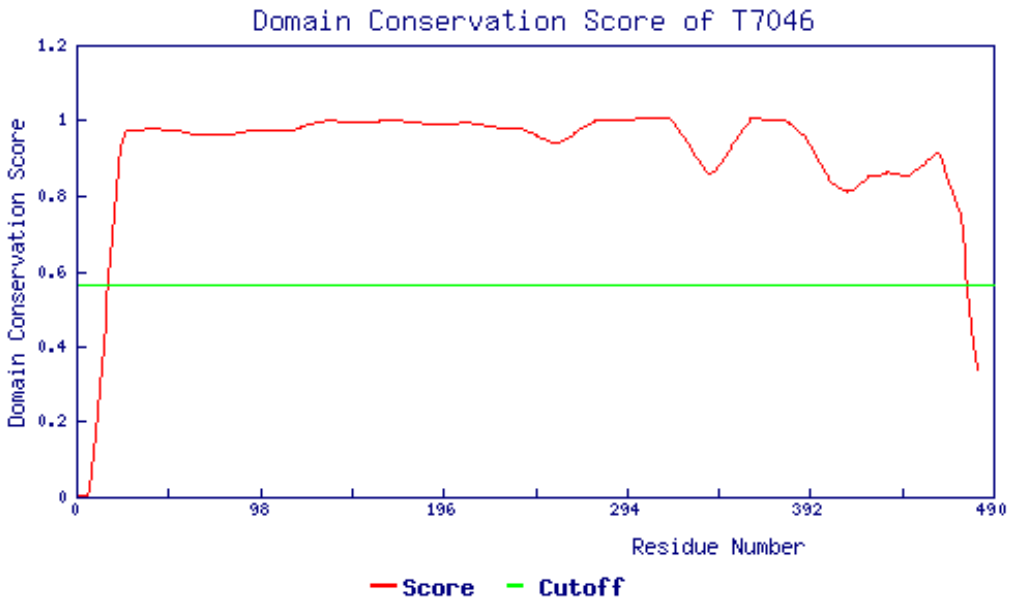


Figure 2.23 Domain boundary prediction of PNPLA3 using Threadom

2.5.3 Secondary structure prediction

The secondary structure of PNPLA3 was predicted with high confidence for most residues, with slight dips in confidence between residues 40-45, 156-165 and 223-235 (Figure 2.24).

A total of eight β-strands are predicted, all within the initial 220 residues of the protein. Five of the β strands comprise of four or more amino acids and span residues 8-12, 115-120, 125-129, 201-208 and 211-215. The remaining three β-strands span less than three residues; two are predicted with low confidence to span residues 158-159, 163-164, and one with high confidence to span residues 183-185 (Figure 2.24).

A total of 15 α-helices are predicted distributed evenly throughout the protein. Of these Six helices are positioned between β strands 1 and 2, spanning residues 18-32, 35-38, 49-57, 62-78, 90-101 and 106-110.

There is a single helix between β strands 2 and 3, spanning residues 134-144. The remaining eight helices lie in the C-terminal region, beyond β sheet eight, spanning residues 218-224, 231-239, 241-252, 286-297, 322-333, 362-375, 382-401 and 448-550 (Figures 2.24).

All of the helices are relatively long coving six or more residues; their prediction is made with a high degree of confidence except for the two helices residing between residues 106-110 and 448-450, which are shorter and predicted with lower confidence (Figures 2.24).

The remainder of the sequence is predicted to comprise disorder regions or random coils. There are several notably large regions of random coil between residues 145-200, 255-285, 335-362 and 405-481.

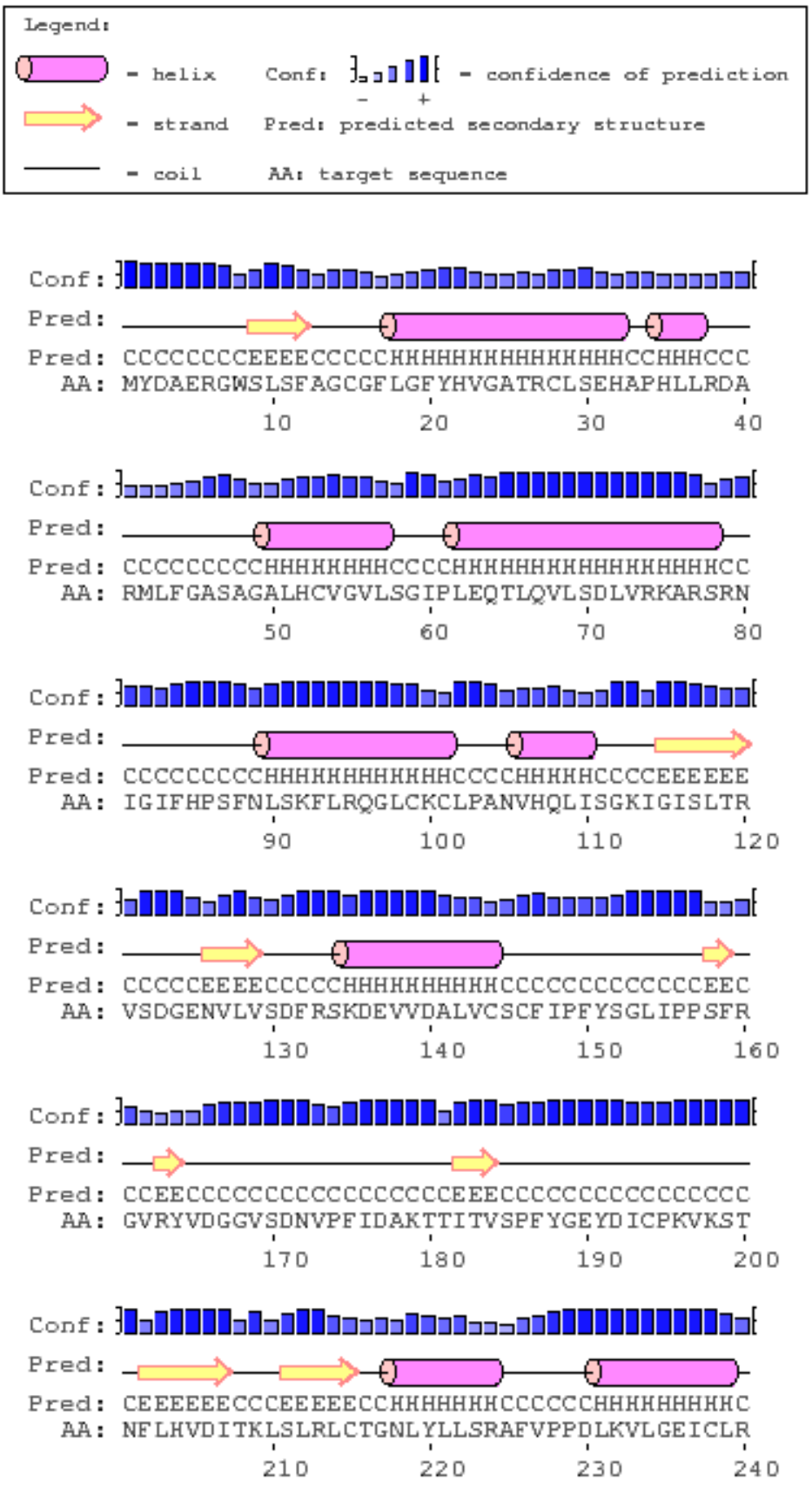


Figure 2.24 Secondary structure prediction using Psipred (cont'd next page)

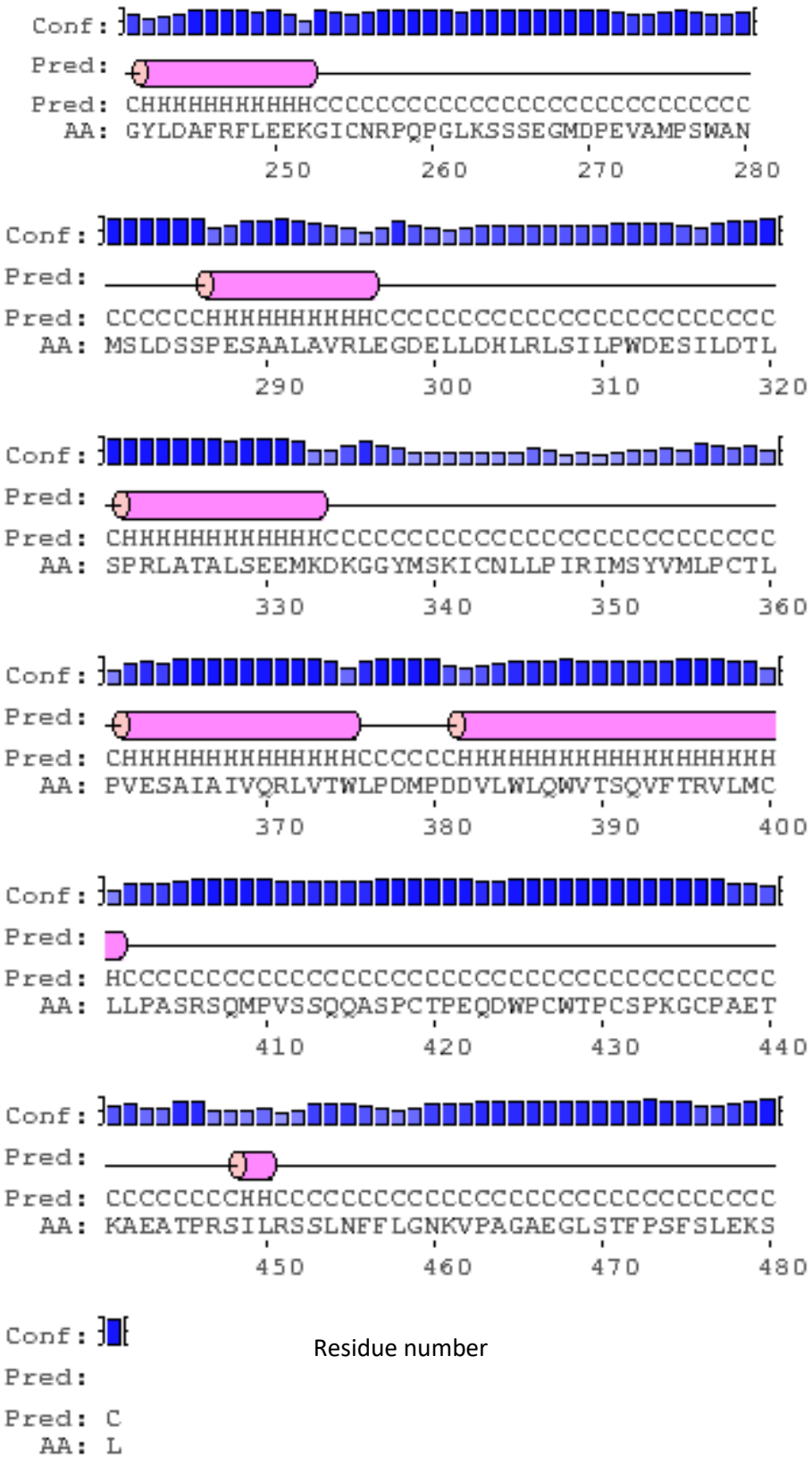


Figure 2.24 (continued)

2.5.4 Transmembrane helices:

A total of 5 potential transmembrane helices were predicted for PNPLA3. However, both the number and positions of the predicted transmembrane helices and in the level of confidence of the predictions varied widely between programs. (Table 2.3).

Table 2.3 Summary of predictions of transmembrane helices inPNPLA3 by various software packages

Platform	Number of Helices	Positions
TMHMM	2	20-35, 45-55
Phobius	5	10-35, 45-55, 145-155, 340-375, 380-410
MEMSAT-SVM	5	11-26, 42-57, 139-154, 212-227, 348-375
MEMSAT3	1	22-52
Spoctopus	2	12-30, 48-65

Two transmembrane helices were predicted using TMHMM, spanning residues 10-35 and 45-55. The prediction of the first is more confident (probability 0.65) than the prediction of the second (probability 0.1) (Figure 2.25).

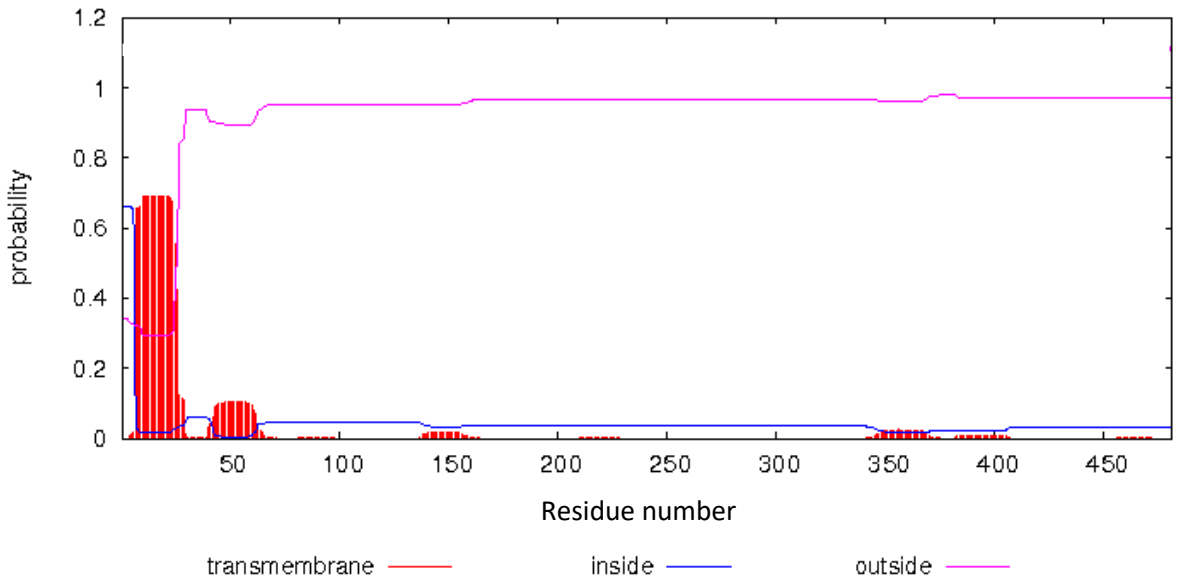


Figure 2.25 PNPLA3 transmembrane helix prediction using TMHMM

The probability of cytoplasmic residues are marked in pink and of non-cytoplasmic residues in blue. Potential transmembrane regions are highlighted with vertical red bars.

Five trans-membrane helices were predicted using Phobius. Two of these, spanning residues 10-35 and 45-55, correspond to the helices predicted by TMHMM, but with notably different probabilities. A short helix was predicted between 145-155, but with low confidence (below 0.2) while two helices in the C-terminal, between residues 340-375 and 380-410 were predicted with high probabilities (above 0.8) (Figure 2.26).

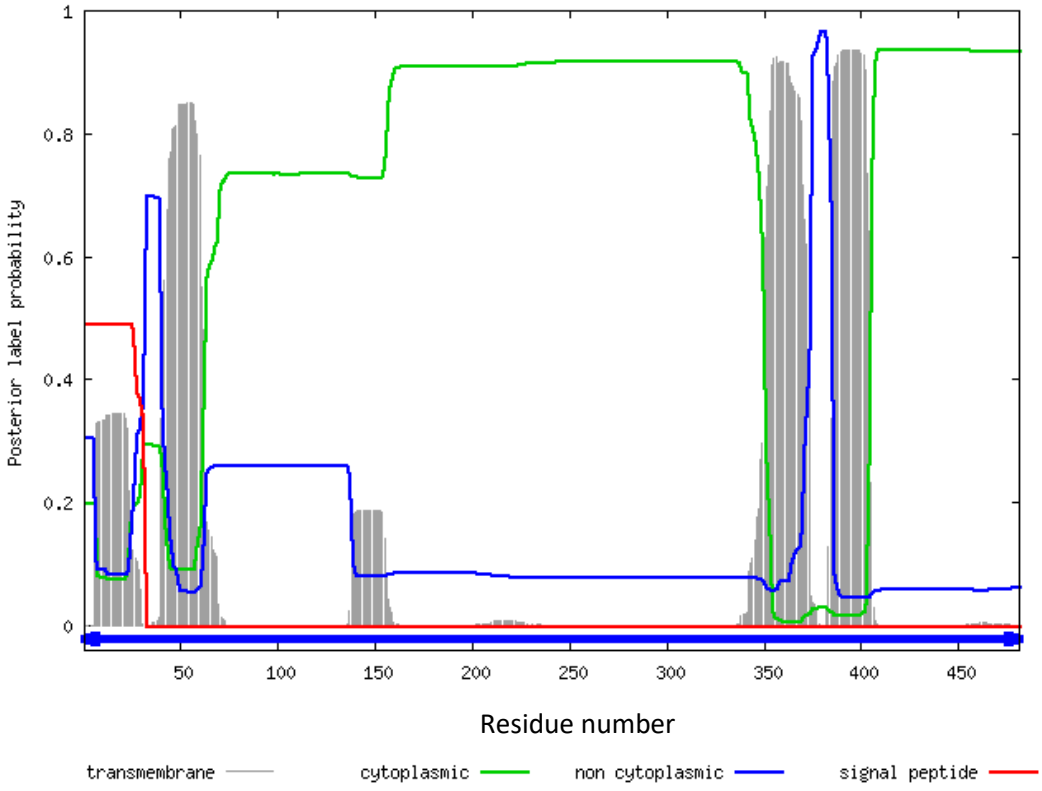


Figure 2.26 PNPLA3 transmembrane helix prediction using Phobius

The probability of cytoplasmic residues are marked in green and of non-cytoplasmic residues in blue and of signal peptide in red. Potential transmembrane regions are highlighted with grey bars.

Five transmembrane helices were predicted using MEMSAT-SVM, spanning residues 11-26, 42-57, 139-154, 212-227 and 348-375 (Figures 2.27) while only one was predicted using MEMSAT3, spanning residues 33-52 (Figure 2.28). Neither of these packages provides an assessment of the confidence of the predictions.

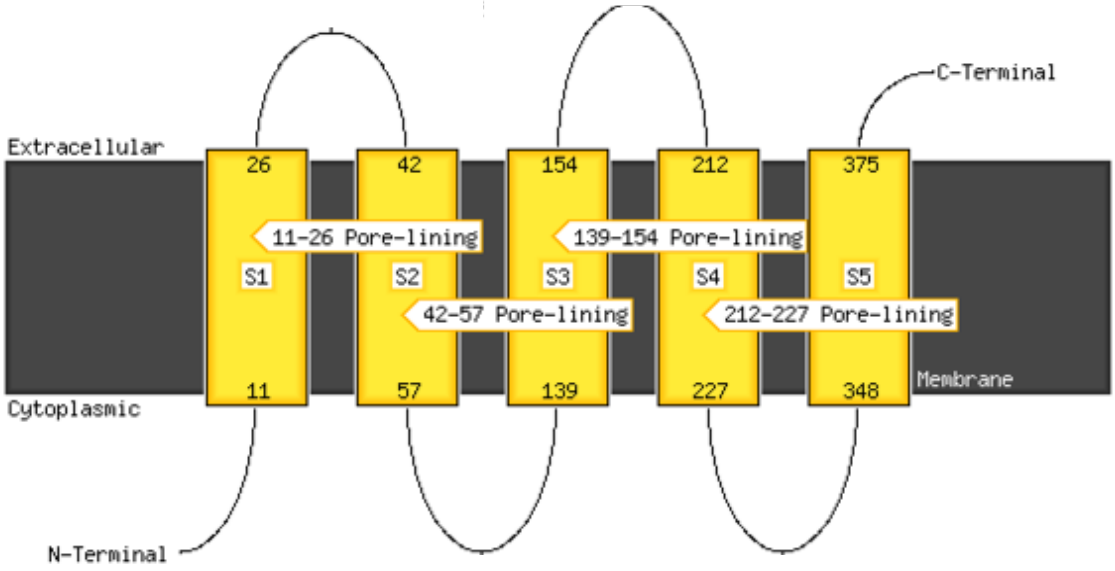


Figure 2.27 PNPLA3 transmembrane prediction using MEMSAT-SVM

The yellow rectangles represent predicted transmembrane helices (labelled S1 to S5). The residues on the membrane cytoplasm interface are numbered.

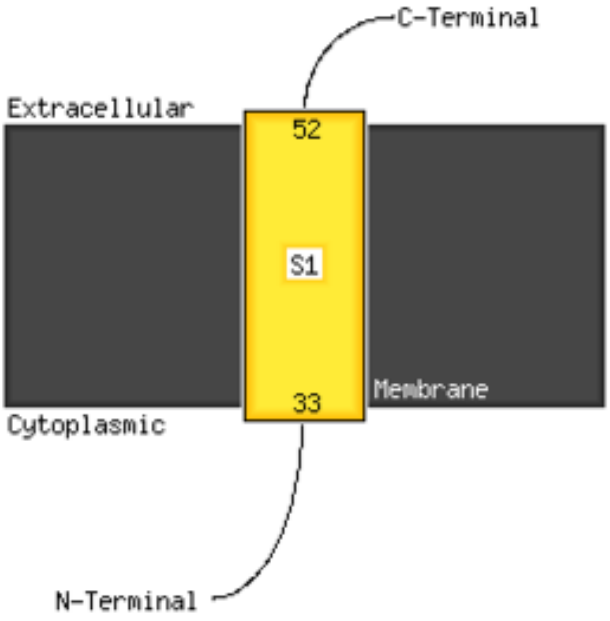


Figure 2.28 PNPLA3 transmembrane prediction using MEMSAT3

The yellow rectangle represents the predicted transmembrane helix (labelled S1). The residues on the membrane cytoplasm interface are numbered

The two helices at the N-terminal of the protein identified using the other packages were also predicted using Octopus with probability close to 1. Spoctopus confirmed these were not signal peptide. however, predicts the majority of the protein is located outside of the cell (Figure 2.29).

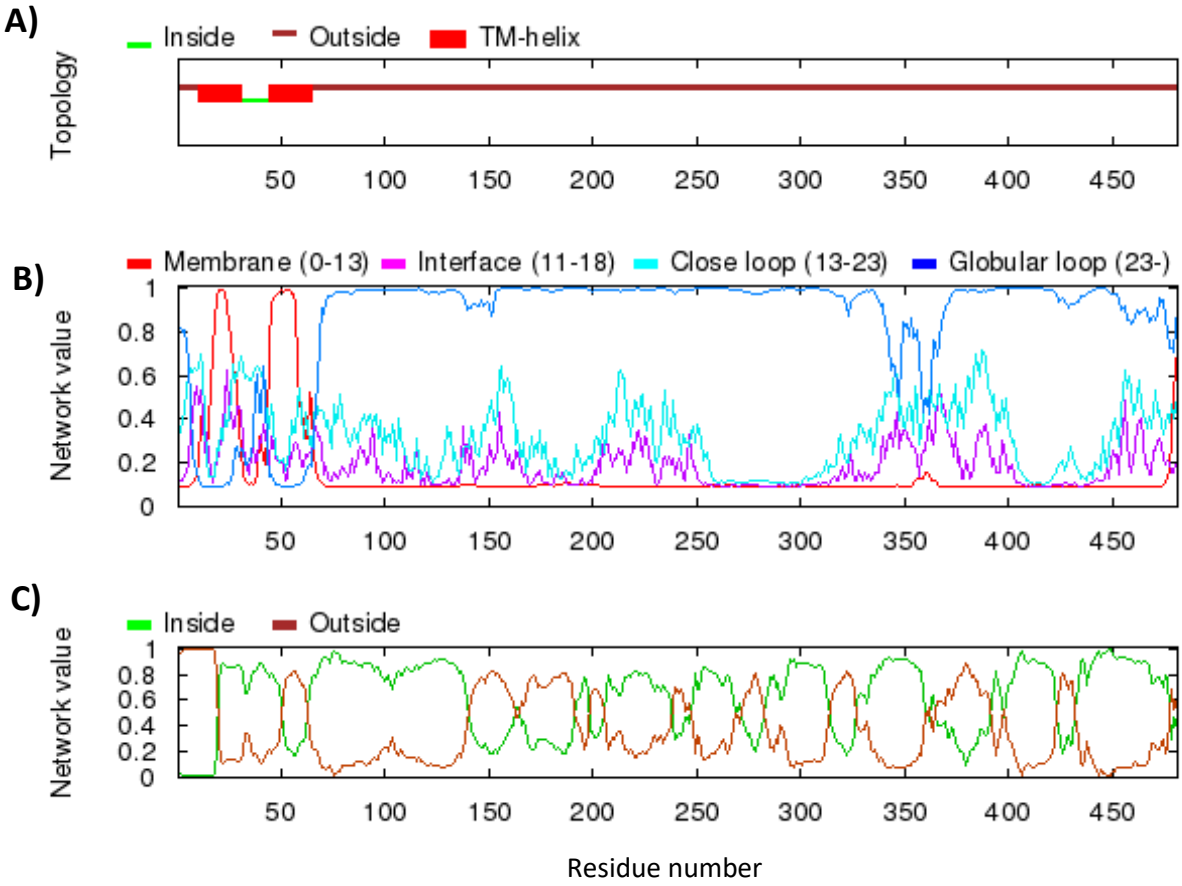


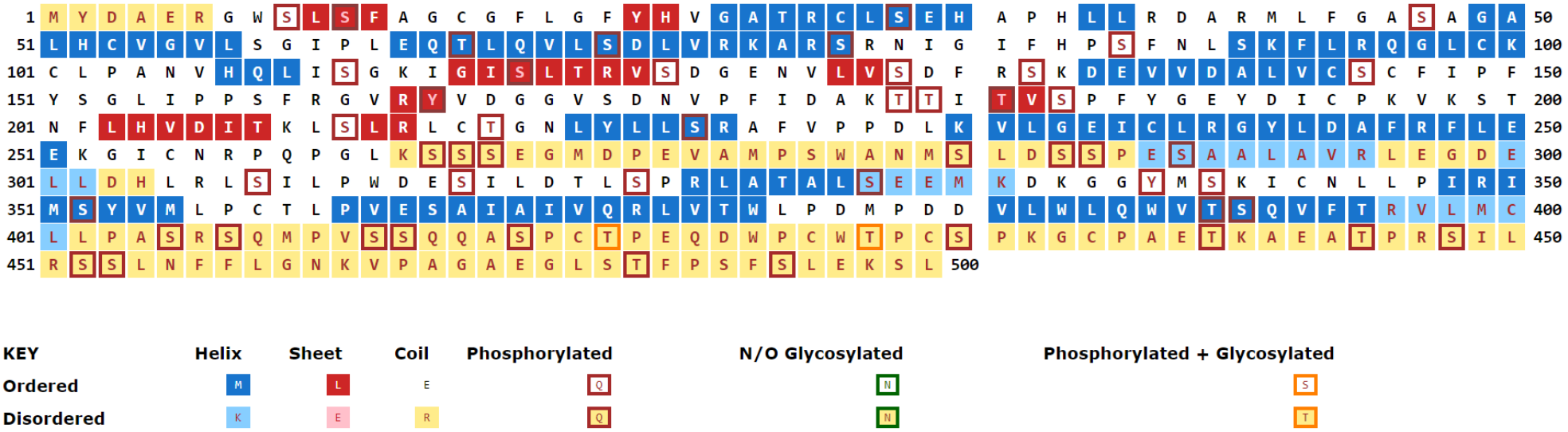
Figure 2.29 PNPLA3 Transmembrane prediction using Spoctopus

A) A summary of the most likely topology of PNPLA3 (no signal peptides were predicted)
B) The estimated preference for each residue to be located at different distances from the membrane
C) The estimated preference of a particular residue to be located either on the inside or outside of the membrane.

2.5.5 Post-translational modification prediction

2.5.5.1 Phosphorylation and glycosylation

Fifty-one potential phosphorylation sites and two glycosylation sites at residues 420 and 429 were predicted using FFPred 2.0. However, none of the sites were predicted with high confidence (Figure 2.30).



KEY	Helix	Sheet	Coil	Phosphorylated	N/O Glycosylated	Phosphorylated + Glycosylated
Ordered	M	L	E	Q	N	S
Disordered	K	E	R	Q	N	T

Figure 2.30 PNPLA3 post-translational modification prediction using FFPred 2.0

Secondary structure highlighted using coloured squares, and post-translational modifications highlighted with coloured outlines.

2.5.5.2 Sites of sumoylation

Lysine 113 was the only residue predicted using Sumoplot to have the potential for sumoylation. However, this prediction had a low confidence score of only 0.5 (Table 2.4).

Table 2.4 PNPLA3 sumoylation sites predicted using Sumoplot

No.	Pos.	Group	Score
1	K113	HQLIS **GKIG** ISLTR	0.5

* Position: amino acid abbreviated letter and residue number; group: the adjacent amino acid residues surrounding the selected residue; score: the confidence score of the prediction.

2.5.4.3 Secretion signals

PNPLA3 was predicted to have the potential for excretion using SecretomeP 2.0 with a low confidence score (2.31).

```
# Name          NN-score
#
# ======================
sp_Q9NST1_P      0.550
# ======================
```

Figure 2.31 Putative secretion signals predicted using SecretomeP 2.0

The NN-score represents the confidence score produced by the neural network within SecretomeP 2.0 and reflects the probability of non-normal secretion.

2.5.5.4 Ubiquitination sites

Six potential ubiquitination sites were predicted using UbPred. The sites at residues 333, 335 and 434 were predicted with low confidence; sites at residues 441 and 479 with medium confidence; and the site at residue 263 with high confidence (Figure 2.32).

In contrast, the only ubiquitination site predicted using iUbiq-Lys was at residue 179.

>sp|Q9NST1|PLPL3_HUMAN Patatin-like phospholipase domain-containing protein 3 OS=Homo sapiens GN=PNPLA3 PE=1 SV=2
MYDAERGWSLSFAGCGFLGFYHVGATRCLSEHAPHLLRDARMLFGASAGALHCVGVLSGIPLEQTLQVLSDLVRKARSRN
IGIFHPSFNLSKFLRQGLCKCLPANVHQLISGKIGISLTRVSDGENVLVSDFRSKDEVVDALVCSCFIPFYSGLIPPSFR
GVRYVDGGVSDNVPFIDAKTTITVSPFYGEYDICPKVKSTNFLHVDITKLSLRLCTGNLYLLSRAFVPPDLKVLGEICLR
GYLDAFRFLEEKGICNRPQPGLKSSSEGMDPEVAMPSWANMSLDSSPESAALAVRLEGDELLDHLRLSILPWDESILDTL
SPRLATALSEEMKDKGGYMSKICNLLPIRIMSYVMLPCTLPVESAIAIVQRLVTWLPDMPDDVLWLQWVTSQVFTRVLMC
LLPASRSQMPVSSQQASPCTPEQDWPCWTPCSPKGCPAETKAEATPRSILRSSLNFFLGNKVPAGAEGLSTFPSFSLEKS
L

Output:

Residue	Score	Ubiquitinated
75	0.36	No
92	0.21	No
100	0.41	No
113	0.42	No
135	0.54	No
179	0.43	No
196	0.50	No
198	0.51	No
209	0.45	No
232	0.30	No
252	0.40	No
263	0.88	Yes High confidence
333	0.63	Yes Low confidence
335	0.67	Yes Low confidence
341	0.29	No
434	0.66	Yes Low confidence
441	0.77	Yes Medium confidence
461	0.41	No
479	0.81	Yes Medium confidence

Legend:

Label	Score range	Sensitivity	Specificity
Low confidence	0.62 ≤ s ≤ 0.69	0.464	0.903
Medium confidence	0.69 ≤ s ≤ 0.84	0.346	0.950
High confidence	0.84 ≤ s ≤ 1.00	0.197	0.989

Figure 2.32 PNPLA3 ubiquitination sites predicted using UbPred.[226]

2.5.5.5 Protease cleavage sites

A total of 1096 cleavage sites were predicted using PeptideCutter. The cleavage sites were evenly spread throughout the protein, with no notable cleavage hotspots. The majority of the cleavage sites were due to proteases with lower specificity, for example, chymotrypsin, pepsin and proteinase K (Table 2.5).

Caspase1, Caspase10, Caspase2, Caspase3, Caspase4, Caspase5, Caspase6, Caspase7, Caspase8, Caspase9, 3C protease, Enterokinase, Factor Xa, GranzymeB, Hydroxylamine, Tobacco etch virus protease had no potential cleavage sites within PNPLA3. Thrombin had only one potential cleavage site at residue 447.

Table 2.5 Protease cleavage sites in PNPLA3 predicted using PeptideCutter

Protease	No. cleavage sites	Residue number of cleavage site
Arg-C proteinase	26	6 27 38 41 74 77 79 95 120 133 160 163 213 224 240 247 257 295 306 323 349 371 396 406 447 451
Asp-N endopeptidase	25	2 38 70 122 130 135 139 165 170 176 191 205 229 243 269 283 298 302 312 317 333 377 380 381 423
Asp-N endopeptidase + N-terminal Glu	48	2 4 30 38 62 70 122 124 130 135 136 139 165 170 176 189 191 205 229 235 243 249 250 266 269 271 283 287 296 298 299 302 312 313 317 329 330 333 362 377 380 381 421 423 438 442 466 477
BNPS-Skatole	8	8 278 312 375 385 388 425 428
CNBr	12	1 42 269 275 281 332 339 351 355 379 399 409
Chymotrypsin-high specificity (C-term to [FYW], not before P)	38	2 8 12 17 20 21 44 84 88 93 132 147 150 151 159 164 175 187 188 191 202 220 226 242 246 248 278 312 338 353 375 385 388 394 428 456 457 475
Chymotrypsin-low specificity (C-term to [FYWML], not before P)	109	2 8 10 12 17 18 20 21 22 29 32 35 36 37 42 43 44 51 52 57 62 66 69 72 84 88 90 93 94 98 107 109 118 128 132 142 147 150 151 154 159 164 175 187 188 191 202 203 204 210 212 214 219 220 221 222 226 231 234 239 242 243 246 248 249 262 269 278 281 283 292 296 301 302 304 305 307 312 317 320 324 328 332 338 339 345 351 353 355 372 375 384 385 386 388 394 398 399 401 428 450 454 456 457 458 469 475 477 481
Clostripain	26	6 27 38 41 74 77 79 95 120 133 160 163 213 224 240 247 257 295 306 323 349 371 396 406 447 451
Formic acid	25	3 39 71 123 131 136 140 166 171 177 192 206 230 244 270 284 299 303 313 318 334 378 381 382 424
Glutamyl endopeptidase	23	5 31 63 125 137 190 236 250 251 267 272 288 297 300 314 330 331 363 422 439 443 467 478
Iodosobenzoic acid	8	8 278 312 375 385 388 425 428
LysC	19	75 92 100 113 135 179 196 198 209 232 252 263 333 335 341 434 441 461 479
LysN	19	74 91 99 112 134 178 195 197 208 231 251 262 332 334 340 433 440 460 478

Table 2.5 (continued)

NTCB (2-nitro-5-thiocyanobenzoic acid)	18	14 27 52 98 100 143 145 193 214 237 254 342 357 399 418 426 430 435
Pepsin (pH1.3)	114	9 10 11 12 16 17 18 19 20 28 36 42 44 50 51 56 57 61 65 66 68 69 71 72 83 88 89 90 92 93 98 108 117 118 127 128 131 132 141 142 146 149 153 159 174 186 201 202 203 209 210 212 214 218 219 220 221 222 225 231 233 238 239 243 245 246 248 262 282 283 291 292 296 300 301 302 304 305 307 310 316 317 319 324 327 328 344 346 356 360 372 376 383 384 385 386 393 394 397 400 402 450 454 455 456 457 458 468 469 472 475 476 477 480
Pepsin (pH>2)	133	1 2 7 9 10 11 12 16 17 18 19 20 21 28 36 42 44 50 51 56 57 61 65 66 68 69 71 72 83 88 89 90 92 93 98 108 117 118 127 128 131 132 141 142 146 149 151 153 159 164 174 186 188 190 191 201 202 203 209 210 212 214 218 219 220 221 222 225 231 233 238 239 241 243 245 246 248 262 278 282 283 291 292 296 300 301 302 304 305 307 310 311 316 317 319 324 327 328 338 344 346 352 353 356 360 372 374 376 383 384 385 386 387 388 393 394 397 400 402 425 450 454 455 456 457 458 468 469 472 475 476 477 480
Proline-endopeptidase [*]	2	86 258
Proteinase K	231	2 4 5 8 10 12 13 17 18 20 21 23 25 26 29 31 33 36 37 40 43 44 46 48 50 51 54 56 57 60 62 63 65 66 68 69 72 73 76 81 83 84 88 90 93 94 98 102 104 106 109 110 114 116 118 119 121 125 127 128 129 132 137 138 139 141 142 143 147 148 150 151 154 155 159 162 164 165 169 173 175 176 178 180 181 182 183 184 187 188 190 191 193 197 200 202 203 205 207 208 210 212 214 216 219 220 221 222 225 226 227 231 233 234 236 237 239 242 243 245 246 248 249 250 251 254 262 267 272 273 274 278 279 283 288 290 291 292 293 294 296 297 300 301 302 305 307 309 310 312 314 316 317 319 320 324 325 326 327 328 330 331 338 342 345 346 348 350 353 354 356 359 360 362 363 365 366 367 368 369 372 373 374 375 376 383 384 385 386 388 389 390 393 394 395 397 398 401 402 404 411 416 420 422 425 428 429 438 439 440 442 443 444 445 449 450 454 456 457 458 462 464 466 467 469 471 472 475 477 478 481

Table 2.5 (continued)

Staphylococcal peptidase I	21	5 31 63 125 137 190 236 250 267 272 288 297 300 314 330 363 422 439 443 467 478
Thermolysin	146	9 11 12 16 17 19 22 24 28 35 36 41 42 43 45 47 49 50 53 55 56 61 65 67 68 72 75 80 82 83 87 89 92 93 97 103 105 108 109 113 115 117 120 126 127 128 138 141 142 146 149 153 158 161 164 168 174 175 181 183 186 196 201 202 204 209 211 213 218 220 221 224 225 232 233 238 242 245 247 248 253 261 268 273 278 280 282 289 290 291 292 293 295 301 304 306 308 315 316 319 323 324 326 327 338 341 344 347 349 350 353 354 361 364 365 366 367 368 371 372 383 385 388 392 393 396 397 398 400 403 410 415 437 441 448 449 453 455 456 457 463 465 468 474 476 480
Thrombin	1	447
Trypsin	44	6 27 38 41 74 75 77 79 92 95 100 113 120 133 135 160 163 179 196 198 209 213 224 232 240 247 252 263 295 306 323 333 335 341 349 371 396 406 434 441 447 451 461 479

2.5.6 Functional and sub-cellular localisation predications

PNPLA3 was predicted to be involved in 25 biological *processes* with high probability using FFPred 2.0 (Table 2.6). The most likely involved process is ion transmembrane transport with a probability score of 0.910. Four out of the five most probable activities are transport related, while regulation of metabolic process is also probable.

Table 2.6 PNPLA3 biological process prediction using FFPred 2.0

GO term	Name	Prob	SVM Reliability*
GO:0034220	ion transmembrane transport	0.910	H
GO:0006810	transport	0.882	H
GO:0019222	regulation of metabolic process	0.876	H
GO:0006812	cation transport	0.833	H
GO:0055085	transmembrane transport	0.804	H
GO:0044281	small molecule metabolic process	0.739	H
GO:0006811	ion transport	0.737	H

Table 2.6 (continued)

GO:0046486	glycerolipid metabolic process	0.725	H
GO:0098655	cation transmembrane transport	0.666	H
GO:0006796	phosphate-containing compound metabolic process	0.656	H
GO:0030001	metal ion transport	0.643	H
GO:0016310	phosphorylation	0.639	H
GO:0019637	organophosphate metabolic process	0.634	H
GO:0006629	lipid metabolic process	0.634	H
GO:0007166	cell surface receptor signalling pathway	0.626	H
GO:0046474	glycerophospholipid biosynthetic process	0.616	H
GO:0051171	regulation of nitrogen compound metabolic process	0.575	H
GO:0006820	anion transport	0.566	H
GO:0015672	monovalent inorganic cation transport	0.559	H
GO:0051649	establishment of localization in cell	0.542	H
GO:0007264	small GTPase mediated signal transduction	0.525	H
GO:0010468	regulation of gene expression	0.524	H
GO:0009059	macromolecule biosynthetic process	0.521	H
GO:0009056	catabolic process	0.510	H
GO:0050877	neurological system process	0.507	H
GO:0044237	cellular metabolic process	0.958	L
GO:0008152	metabolic process	0.915	L
GO:0050896	response to stimulus	0.901	L
GO:0007154	cell communication	0.831	L
GO:0008654	phospholipid biosynthetic process	0.817	L
GO:0051716	cellular response to stimulus	0.811	L
GO:0007165	signal transduction	0.790	L
GO:0032502	developmental process	0.739	L
GO:0023052	signalling	0.736	L
GO:0007275	multicellular organismal development	0.666	L
GO:0006650	glycerophospholipid metabolic process	0.664	L
GO:0009058	biosynthetic process	0.635	L
GO:0048856	anatomical structure development	0.609	L
GO:0006807	nitrogen compound metabolic process	0.597	L

Table 2.6 (continued)

GO:0009966	regulation of signal transduction	0.587	L
GO:0019538	protein metabolic process	0.577	L
GO:0009893	positive regulation of metabolic process	0.572	L
GO:0030154	cell differentiation	0.560	L
GO:0031325	positive regulation of cellular metabolic process	0.526	L
GO:0010033	response to organic substance	0.502	L

* SVM reliability provides information on the degree of reliability of prediction for the specific GO term. H corresponds to GO terms which have high values of the Matthews Correlation Coefficient, sensitivity, specificity and precision in prediction. While L corresponds to GO terms with lower reliability and tend to be less reliably predicted by this algorithm.

PNPLA3 was also predicted, with high probability, to be involved in 12 biological *functions*. Amongst the most probable would be catalytic activity (probability of 0.882), hydrolase activity acting on ester bonds, acyl transferase activity and transmembrane transporter activity (Table 2.7).

Table 2.7 PNPLA3 molecular function prediction using FFPred 2.0

GO term	Name	Prob	SVM Reliability*
GO:0003824	catalytic activity	0.882	H
GO:0016788	hydrolase activity, acting on ester bonds	0.855	H
GO:0016746	transferase activity, transferring acyl groups	0.827	H
GO:0016740	transferase activity	0.770	H
GO:0005215	transporter activity	0.738	H
GO:0022857	transmembrane transporter activity	0.716	H
GO:0015267	channel activity	0.714	H
GO:0022891	substrate-specific transmembrane transporter activity	0.709	H
GO:0015075	ion transmembrane transporter activity	0.688	H
GO:0016747	transferase activity, transferring acyl groups other than amino-acyl groups	0.687	H
GO:0022890	inorganic cation transmembrane transporter activity	0.661	H

GO:0005216	ion channel activity	0.629	H
GO:0016787	hydrolase activity	0.735	L
GO:0036094	small molecule binding	0.638	L
GO:0005102	receptor binding	0.511	L

* SVM reliability informs us of the degree of reliability of prediction for the specific GO term. H corresponds to GO terms which have high values of the Matthews Correlation Coefficient, sensitivity, specificity and precision in prediction. While L corresponds to GO terms with lower reliability and tend to be less reliably predicted by this algorithm.

With regards to the subcellular localisation of PNPLA3, it was predicted to be integral component of membrane (0.991 probability) (Table 2.8). All 16 of the high scoring SVM models were related to the membranes and possible cytoplasmic localisation was predicted with only a low reliability SVM.

Table 2.8 PNPLA3 cellular component prediction using FFPred 2.0

GO term	Name	Prob	SVM Reliability*
GO:0016021	integral component of membrane	0.991	H
GO:0031224	intrinsic component of membrane	0.988	H
GO:0016020	membrane	0.968	H
GO:0005887	integral component of plasma membrane	0.949	H
GO:0031226	intrinsic component of plasma membrane	0.912	H
GO:0005886	plasma membrane	0.889	H
GO:0042175	nuclear outer membrane-endoplasmic reticulum membrane network	0.817	H
GO:0031090	organelle membrane	0.811	H
GO:0098588	bounding membrane of organelle	0.809	H
GO:0005789	endoplasmic reticulum membrane	0.777	H
GO:0012505	endomembrane system	0.755	H
GO:0071944	cell periphery	0.690	H
GO:0005783	endoplasmic reticulum	0.643	H
GO:0034702	ion channel complex	0.639	H
GO:0005739	mitochondrion	0.569	H
GO:1902495	transmembrane transporter complex	0.529	H

Table 2.8 (continued)

GO:0005737	cytoplasm	0.895	L
GO:0043229	intracellular organelle	0.820	L
GO:0043231	intracellular membrane-bounded organelle	0.813	L
GO:0032991	macromolecular complex	0.515	L

* SVM reliability informs us of the degree of reliability of prediction for the specific GO term. H corresponds to GO terms which have high values of the Matthews Correlation Coefficient, sensitivity, specificity and precision in prediction. While L corresponds to GO terms with lower reliability and tend to be less reliably predicted by this algorithm.

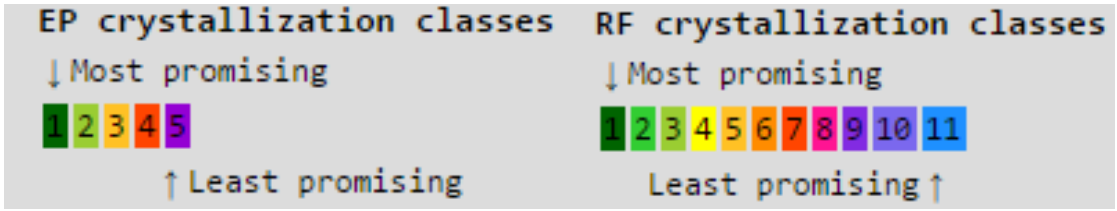
2.5.7 Crystallisation prediction

Full length PNPLA3 was predicted to have a very low propensity to crystallise using bot XtalPred and Crysalis. XtalPred predicts that full length PNPLA3 is in the least promising class for crystallisation, scoring 5 on the expert pool (EP) scale and 11 on the random forest (RF) scale (Figure 2.33).

The main reason for this difficulty is a combination its instability, its high surface hydrophobicity and the fact that it contains a disordered segment, which is longer than that observed in any known structure (Figure 2.34).

Other potential domain fragments selected based on earlier secondary structural properties, also have a low probability of successful crystallisation. XtalPred predicted that the fragment most likely to achieve crystallisation spanned residues 7-179; this achieved scores of 3 and 6 on the EP and RF respectively; this was supported by predictions using Crysalis I but not Crysalis II (Figure 2.34).

The only fragment predicted to perform successfully throughout experimentation to produce a solvable structure by Crysalis spanned residues 179 to 300. All other fragments were predicted to have failed by the purification stage (Figure 2.35).



Target id (link to target details)	EP-class	RF-class	Length	Gravy index	Instability index (II)	Isoelectric point (pI)	Coiled coils	Longest disorder regio
sp|Q9NST1	5	11	481	0.10	55.00	6.27	0	47

Percentage of coil structure	Transmembrane helices (TM)	Signal peptides (SP)	Insertions score	Homologs in NR (clustered to 60% seq. ident.)	Homologs in PDB
45	No	No	0.15	842	9

Target id (link to target details)	EP-class	RF-class	Length	Gravy index	Instability index (II)	Isoelectric point (pI)	Coiled coils	Longest disorder region	Percentage of structure
61-250	2	11	190	0.19	39.03	8.74	0	1	38
61-201	3	11	140	0.08	45.98	8.72	0	4	43
7-250	3	11	244	0.29	35.30	8.69	0	0	37
7-179	3	6	173	0.29	36.48	8.76	0	0	37
7-138	5	6	152	0.35	38.84	8.79	0	1	30
7-300	5	7	294	0.14	42.96	7.15	0	43	46
300-481	5	11	182	0.07	73.50	5.18	0	26	44
1-481	5	11	481	0.10	55.00	6.27	0	47	45
179-300	5	11	122	-0.11	51.79	5.19	0	41	56
300-400	5	10	101	0.40	58.45	4.72	0	3	32
1-250	2	7	250	0.25	36.25	8.52	0	5	39
1-179	2	7	179	0.23	37.77	8.52	0	5	39

Figure 2.33 Crystallisation prediction of full length PNPLA3 using XtalPred

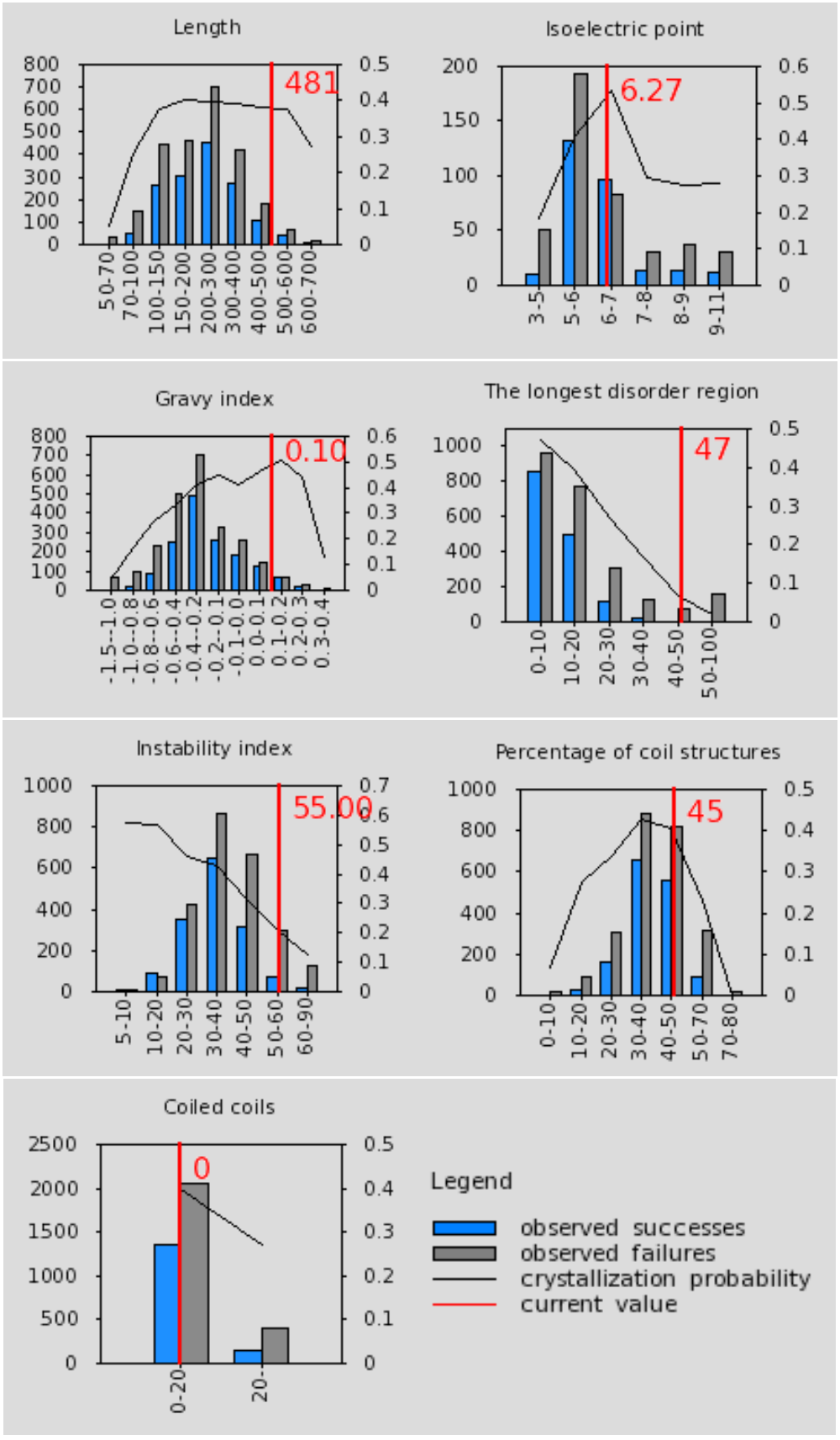


Figure 2.34 Comparison of target features with distributions of crystallization probabilities obtained from TargetDB (Cont'd next page)

Blue bar presents number of successful crystallisations at current value, and grey bar the number of observed failures to crystallise. The Red line denotes the current value for a target protein and the black bar the probability of crystallisation.

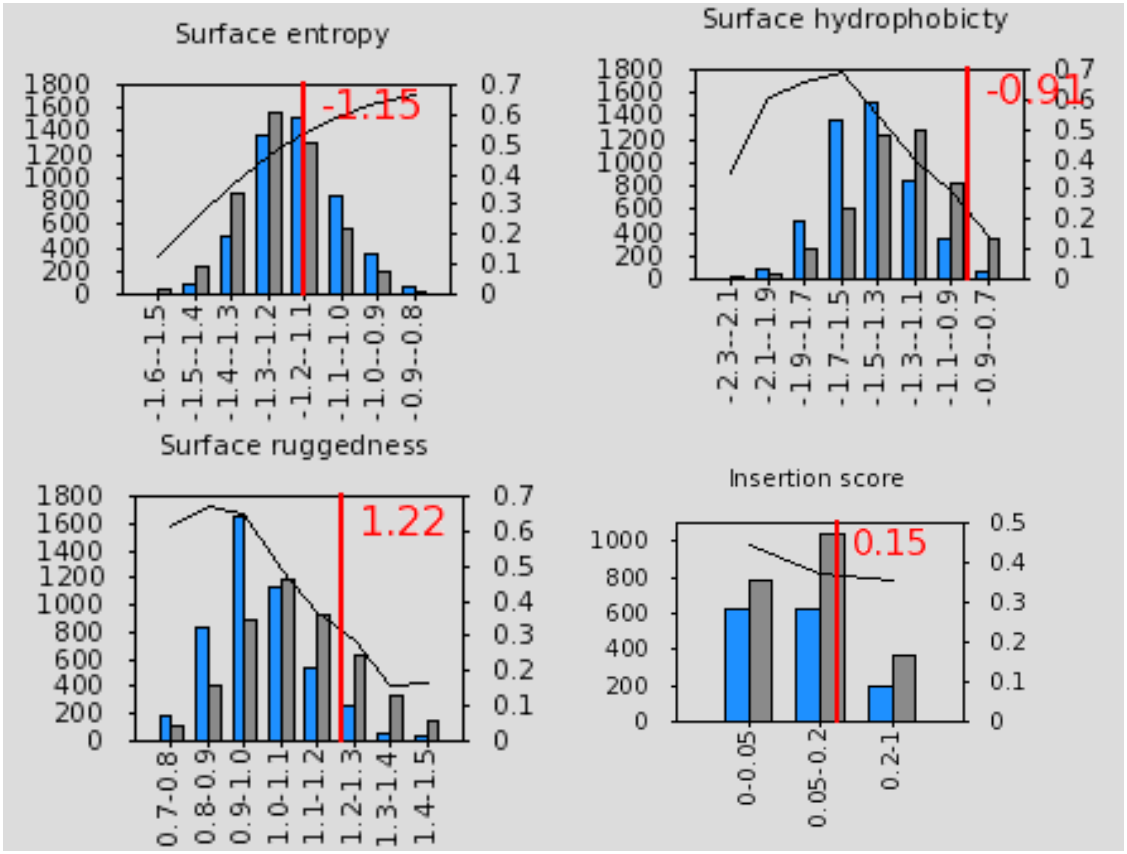


Figure 2.34 (Continued)

Protein	Predictor	CLF	MF	PF	CF	CRYs
1-481_	Crysalis I	failure, 0.449	failure, 0.292	failure, 0.485	success, 0.549	failure, 0.433
	Crysalis II	failure, 0.376	failure, 0.103	failure, 0.483	failure, 0.492	failure, 0.464
179-300_	Crysalis I	success, 0.506	success, 0.522	success, 0.707	success, 0.538	success, 0.666
	Crysalis II	success, 0.51	success, 0.57	success, 0.635	success, 0.5	success, 0.555
300-400_	Crysalis I	success, 0.554	failure, 0.472	failure, 0.35	success, 0.598	success, 0.635
	Crysalis II	success, 0.522	success, 0.513	failure, 0.407	success, 0.544	failure, 0.486
300-481_	Crysalis I	success, 0.502	failure, 0.349	failure, 0.445	success, 0.595	success, 0.533
	Crysalis II	failure, 0.473	failure, 0.249	failure, 0.489	success, 0.524	failure, 0.479
4QMK:A|PDBID|CHAIN|SEQ...	Crysalis I	failure, 0.355	failure, 0.412	success, 0.57	failure, 0.456	failure, 0.407
	Crysalis II	failure, 0.348	failure, 0.176	success, 0.574	failure, 0.455	failure, 0.47
61-201_	Crysalis I	success, 0.523	failure, 0.412	failure, 0.493	failure, 0.335	failure, 0.444
	Crysalis II	failure, 0.478	failure, 0.277	failure, 0.492	failure, 0.41	failure, 0.472
61-250_	Crysalis I	success, 0.51	failure, 0.371	failure, 0.495	failure, 0.388	failure, 0.427
	Crysalis II	failure, 0.448	failure, 0.197	failure, 0.491	failure, 0.433	failure, 0.467
7-138_	Crysalis I	success, 0.548	failure, 0.35	failure, 0.446	success, 0.571	failure, 0.311
	Crysalis II	failure, 0.498	failure, 0.131	failure, 0.305	success, 0.505	failure, 0.459
7-179_	Crysalis I	success, 0.541	failure, 0.352	failure, 0.466	success, 0.51	failure, 0.318
	Crysalis II	failure, 0.477	failure, 0.131	failure, 0.359	failure, 0.482	failure, 0.459
7-250_	Crysalis I	success, 0.522	failure, 0.323	failure, 0.467	success, 0.515	failure, 0.347
	Crysalis II	failure, 0.446	failure, 0.109	failure, 0.395	failure, 0.482	failure, 0.46
7-300_	Crysalis I	success, 0.508	failure, 0.326	success, 0.528	success, 0.52	failure, 0.418
	Crysalis II	failure, 0.433	failure, 0.153	failure, 0.49	failure, 0.487	failure, 0.47

Figure 2.35 Potential domain candidate crystallisation prediction using Crysalis

CLF: sequence cloning failure; MF: material production failure; PF: purification failure; CF: Crystallisation failure; CRYs: structure determination failure.

2.6 Discussion

2.6.1 Homology based domain architecture analysis

Phylogenetic investigations confirmed that PNPLA3 is classified as a member of the Patatin_and_cPLA2 superfamily of proteins, a very large protein family that spans across the tree of life.[150]

Large insertion sequences were observed when the diverse members of the family were aligned between the catalytic residues. It is likely that these regions correspond to solvent-accessible flexible lids which partially cover the active sites of some of these proteins, for example cPLA2. Through the use of "closed lid" and "open lid" forms, cPLA2 displays interfacial activation; as this large sequence is missing from the alignment with PNPLA3, suggests that PNPLA3 will be unable to perform interfacial activation in the same way. [247]

PNPLA3 appears to have diverged from other PNPLA proteins in a common ancestor, resulting in inheritance of all five of these core PNPLA proteins (PNPLA1, PNPLA2, PNPLA3, PNPLA4 and PNPLA5). Thus, humans inherited all of the PNPLA proteins rather than inheriting a single protein which was then subject to duplication events resulting in the creation of multiple homologues.

2.6.1.1 Mouse human comparison

Based on the models produced to date pnpla3 and PNPLA3 not only share high levels of homology based on amino acid conservation, but also retained properties for example hydrophobicity and regions of disorder (Appendix I).

One key difference is that the murine protein lacks the long C-terminal domain present in the human protein. However, the high levels of similarity suggest that it may be a good model to study the functional behaviour of the patatin domain *in vitro*.

2.6.1.2 Phylogenetic relationship between PNPLAs

The present study confirms previous findings that the nine PNPLA proteins roughly fall into two categories. PNPLA1, PNPLA2, PNPLA3, PNPLA4 and PNPLA5 form a specific group of proteins, which share a common ancestral background based on phospholipid specificity.[151] PNPLA6 and

PNPLA7, share a unique common ancestor with NTE like bacterial and fungal proteins, while PNPLA8 and PNPLA9 are more similar to the plant Pat17 proteins.

However, this study provided the novel information that the closest related protein to PNPLA3 is mammalian PNPLA5 rather than PNPLA2 which had previously been believed to be the case.[150]

This is likely due to the fundamental differences in the method used to predict the phylogenetic tree. Previous models were based on a BLAST generated Tree, which computes a pairwise alignment between a query and the database sequences searched. It does not explicitly compute an alignment between the different database sequences in other words it does not perform a multiple alignment. When two database sequences align to different parts of the query, it does not calculate a distance between these two sequences and only the higher scoring sequence is included in the tree.

In contrast, CDART generates superfamily sequence clusters as the basis for observation. This is based on clustering of a refined similarity matrix between sequences. Specific conserved domains within the database are matched and this forms an integral part of the conserved domain curation project. This in turn provides insights into how patterns of residue conservation and divergence in a family relate to functional properties.

Curated alignments contain aligned blocks spanning all rows with no internal gaps within the blocks and unaligned regions between blocks. The blocks are meant to represent conserved structural core motifs of the corresponding domain family, thereby attempting to achieve more complete alignment of distantly related proteins. These analyses are generated from single DNA sequences for the most part and do not account for the prevalence of SNPs in other species.

The identification of the close relationship between PNPLA3 and PNPLA5 is of great importance. Many of the functional inferences made about PNPLA3 have been based on comparison with PNPLA2 and patatin, and the comparison with PNPLA5 is likely to provide more accurate information.

PNPLA5 shows many similarities to PNPLA3, having both been observed to have TG hydrolase, RE hydrolase and transacylase activities; with the predominant activity still under debate. It is also significantly upregulated in the liver in obese mice and inhibited under fasting conditions; features not exhibited by PNPLA2.[151,248]

The further biological importance of PNPLA5 is still under investigation but recent evidence suggests that it is a key enzyme in autophagosome biogenesis.[249] Clearly any future advances in

our understanding of PNPLA5 could offer further insight into the structural and functional behaviour of PNPLA3.

Unfortunately, none of the close homologues of PNPLA3 has a known structure. With such a large database of homologous proteins, many of which have biologically important and disease-causing variations, this is surprising and perhaps implies that there are likely intrinsic challenges in the isolation and purification of this class of proteins as a whole.

2.6.2 Protein properties

The hydrophobicity profile of PNPLA3 shows that the N-terminal is more hydrophobic when compared to the C-terminal domain. This supports evidence that the patatin domain of PNPLA3 forms an interaction with lipid droplets and is in essence the lipid binding domain.

The N-terminal half of PNPLA3 has relatively low disorder, however, there is a large spike in the disorder between residues 250 and 300 predicted with both DISopred and PRDOS, with another large peak in disorder toward the C-terminal 100 residues. This implies a potential domain boundary beginning at residue 250, between two globular domains. Furthermore, the high level of disorder in the C-terminal could suggest potential protein binding activity of this domain.

2.6.3 Secondary structure prediction

The secondary structure of PNPLA3 provided confident prediction for most of the structure. This positions the key I148M variant on a loop at the end of the 7^{th} α helix. meaning that the I148M variation is unlikely to have any impact on the local secondary structure, and likely operates within the tertiary structure.

The exact relationship of PNPLA3 to the lipid membrane is extremely difficult to determine based on its amino acid sequence alone. A range of software was used to predict the membrane topology to help reduce the potential for false errors based on specific algorithms. In consequence there is a degree subjectivity in the appraisal of the results as no one approach can be treated with priority.

The most commonly identified potential transmembrane helices span residues 10-26 and 42-57. These helices are predicted to occur with a range of confidence scores by MEMSAT-SVM, Spoctopus, Phobius and TMHMM. Nevertheless, experiments looking at lipid droplet affinity *in*

vitro suggests that this early N-terminal helix facilitates lipid interface association, but is not an integral membrane protein.[158]

These early N-terminal residues do occur around the conserved pattern encoding an oxyanion hole. This could form an extremely hydrophobic tunnel like pocket which ensures hydrophobic substrate specificity and in turn appears to have transmembrane properties. Other α/β fold proteins have shown similar hydrophobic protperties.[250] Indeed the same features were predicted when testing PNPLA2, which is known not to form transmembrane contacts (Figure 2.36)

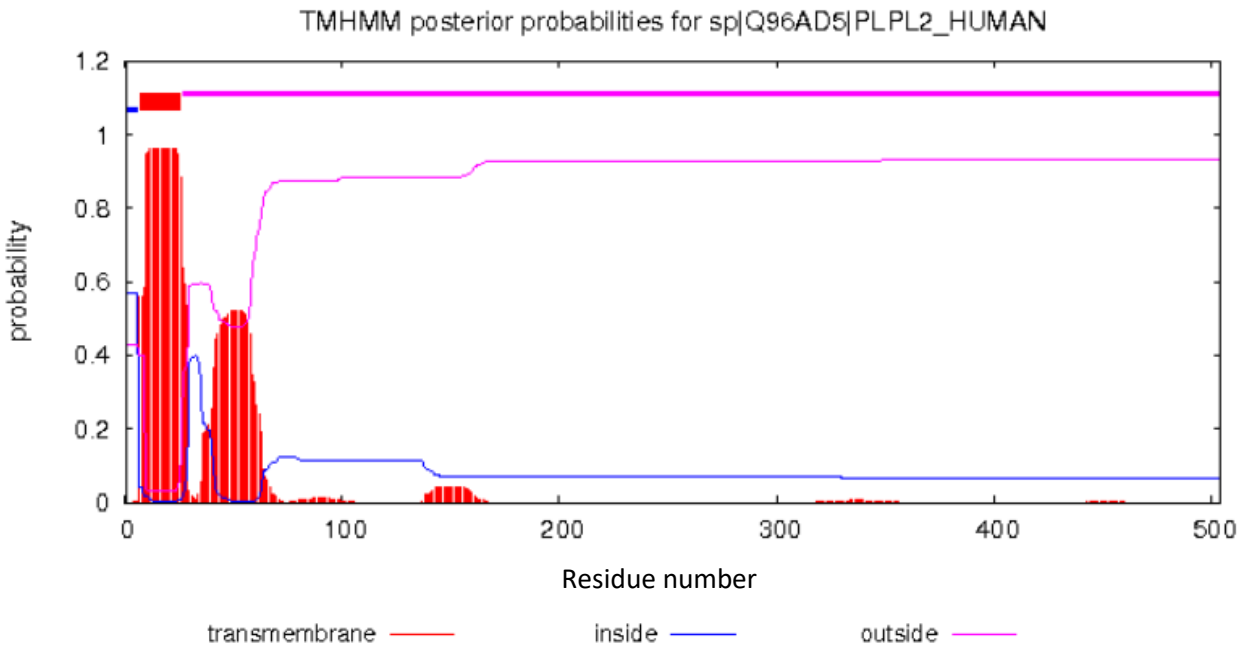


Figure 2.36 PNPLA2 transmembrane helix prediction with TMHMM

Probability of cytoplasmic residues in pink and non-cytoplasmic in blue. Potential transmembrane regions highlighted with red bars.

Additional transmembrane helices are predicted around residues 139-154. However, it is more likely that they are internal structural helices given the proximity of the catalytic aspartate.

While it is unlikely that PNPLA3 is an integral membrane protein, it does interact with lipid droplets *in vivo* and has some membrane related character.[158] While membrane proteins are highly stable in their membrane environment, they show high levels of instability once solubilised in solution. It is thought that the high level of flexibility needed to function within the membrane causes stability issues outside of the membrane or when solubilised in soution.[251] This suggests that PNPLA3 may present stability and solubility issues *in vitro*.

Affinity tags can be used to aid solubility. However, the flexibility of the linker region causes low conformational stability which can often inhibit crystallisation. In such situations, this can be overcome by modifying the linker region to form a more rigid structure, which has been used to crystallise some tagged proteins.[252]

2.6.4 Domain boundaries

Determining the domain architecture of PNPLA3 is extremely challenging without the ability to visually inspect the protein structure, or reproducibly express and observe specific protein activity *in vitro*. Obtaining a correct domain boundary is not only important for the understanding of the protein, but also to help identify potential fragments of the protein which may be amenable to structural investigation.

Predicting domain boundaries based on secondary structure remains somewhat an artform. However, there are guideline which must be observed *viz*.: (i) remove predicted membrane-spanning regions; (ii) avoid disrupting predicted secondary structural elements; (iii) respect the boundaries of globular domains, if known; and (iv) avoid inclusion of low-complexity regions or hydrophobic residues at the termini.[253,254]

The prediction of the domain boundaries in PNPLA3 was inconclusive and varied widely between the packages used. DomPred predicted a domain boundary with high confidence around residue 179, which matches the previously predicted patatin domain boundary. DOMSSEA predicted 10 domain boundaries, while Threadom predicted no boundaries.

Because of the unreliability of the domain architecture as well as the potential transmembrane helix, a range of potential domains was predicted. Use of these may facilitate the chance of developing soluble protein fragments which are able to crystallise.

Viz. residues: 1-481, 7-250, 7-179, 300-400, 300-481, 61-201, 61-250, 179-300, 7-138

2.6.5 Post-translational modification

While there are several sites for potential glycosylation and phosphorylation, there are no clearly important key post translational modifications which relate to the enzyme function. The only modification probable is ubiquitination which is a normal part of the degradation pathway for eukaryotic lipid droplet binding enzymes such as PNPLA2.[173] This means that there is no need for specific post translational machinery within an expression system.

2.6.6 functional prediction

Functional predictions, based on sequence, provides a range of potential biological activities. FFPred2.0 predicted with greatest confidence that PNPLA3 would have both hydrolase and acyl transferase activity. As both of these activities have been detected separately *in vitro*, this does not give any further insight into which the predominant function of the protein may be.

2.6.7 Propensity to crystallise:

PNPLA3 was predicted to be very unlikely to crystallise, be successfully expressed or purified. This is reflected in the lack of publications which have achieved purification of PNPLA3.

These predictions may in part be due to the low homology shared between PNPLA3 and any known structures, particularly in the C-terminal domain; as the prediction software used to make these predictions is informed by proteins which have been successfully crystallised. This means that targets which have low homology with previously crystallised proteins will necessarily be predicted to have a low propensity to crystallise, due to a confirmation bias introduced by the training data set.

However, it is important to note that PNPLA3 belongs to a large superfamily of proteins, many of which have vital biological function and are highly relevant in several models of disease. Yet none of these proteins has a known structure. This suggests even with the caveat above, PNPLA3 is highly likely to be a challenging target for crystallisation.

These potential added difficulties should be carefully taken into account in any attempts to delineate the structure and function of this protein. In particular, truncated variants are recommended to try to remove the long region predicted to be disordered.

2.7 Conclusion

Key findings:

Support for previous findings:

- Experimental evidence that PNPLA3 likely exists as a soluble, but lipid associated protein was supported.
- Predictions of PNPLA3 consisting of multiple domains were supported; *vis:* the patatin domain spanning the initial 179 residues and a C-terminal domain with unknown homology.
- Work suggesting that PNPLA3 has a hydrolase activity and independent work suggesting it has an acyl-transferase activity were both equally supported.

Novel findings:

- PNPLA3 shares stronger homology with PNPLA5, rather than with PNPLA2 as was previously described.
- The secondary structure of PNPLA3 was clarified, and the position of the I148M variant defined to be within a coil region.
- PNPLA3 was predicted not likely to be amenable to *in vitro* expression, purification or crystallisation.
- A range of recombinant constructs of PNPLA3 were designed which may facilitate further structural investigation

A bioinformatic investigation using the sequence of PNPLA3 has yielded several novel insights into the properties of the protein which will provide a foundation for further experimental and computational work.

The fact the predicted amenability of PNPLA3 to expression, purification and crystallisation is low, means that further experimentation is likely to be challenging. However, these predictions are heavily influenced by the lack of homology with other known protein structures, and therefore could be misleading. Despite this, the low levels of homology with other known proteins, and extensive genetic studies make further functional and structural characterisation the next key step.

Since PNPLA3 is believed to be a multi-domain protein, useful information may be derived from working with smaller recombinant fragments of PNPLA3. A range of potentially useful fragments were identified during this investigation which will be taken forward.

Chapter 3

PNPLA3 expression and purification *in vitro*

"The night is in its darkest just before dawn. But keep your eyes open. If you avert your eyes from the dark, you'll be blinded by the rays of a new day. So keep your eyes open, no matter how dark the night ahead may be."

Gintoki Sakata

3.1 Overview

The high-resolution structure of a protein facilitates insight into its biochemical function and provides a platform for drug discovery. Numerous previous attempts have been made to express and purify PNPLA3, but these did not produce either the quantity or purity needed for structural investigation; hence the structure of the protein remains unknown.

In this chapter expression and subsequent purification of PNPLA3 is attempted in an *E. coli* expression system, using a broad range of *PNPLA3* clones.

Although low levels of PNPLA3 were expressed within the cell, the protein was rapidly lost from solution. Further purification and isolation of the protein to a level suitable for structural investigation could not achieved under any conditions tested.

These difficulties were ascribed to:

i) contamination from an *E. coli* protein, ArnA, which both exhibits similar characteristics and co-purifies with PNPLA3; and

ii) the intrinsic insolubility of PNPLA3 under *in vitro* conditions.

This implies that *E. coli* is likely not the optimal expression host for PNPLA3. Any further studies should utilise a different host system or focus on *E. coli* strains which do not produce ArnA, such as LOBSTR.

3.2 Introduction

Proteins are responsible for the key processes required in living cells; and thus, form the crucial building blocks to understand cellular metabolism. To understand the role of a target protein within the greater framework of the cell and the organism, you must first understand the specific behaviours of the protein.

The ultimate detail in understanding a protein arises from the combination of biochemical functional data with high resolution structural data.[255] While methods to investigate the structure and function of a protein *in silico* have been developed (see Chapter 2: *Primary-sequence based investigation of PNPLA3*), they suffer from limitations, which mean direct evaluation of the protein is often still necessary.

An *in vitro* approach is required for the direct study of a protein without interference from other cellular components. This means *in vitro* functional assays can be used to investigate specific reactions catalysed by a protein in detail and measure the kinetics of the reaction. High resolution structural studies can then be used to provide further insight into mechanisms of reaction, ligand binding and protein-protein interactions.

Structural investigations into a protein can be undertaken using a range of experimental techniques including X-ray crystallography, Nuclear Magnetic Resonance (NMR) and cryo-electron microscopy (Cryo-EM). Of these, X-ray crystallography, which was first used in the early 1930s, is the most frequently used method.[256]

X-ray crystallography has historically been able to provide higher resolution structural models than any other technique and hence is particularly valuable as an investigatory tool in structure-based drug design and ligand docking experiments.[257,258] Advancements in the technology, in particular the development of third generation synchrotron radiation sources and hybrid methods for processing the collected data now allow proteins to be studied in higher resolutions than ever before.[259]

3.2.1 Structure determination using X-ray crystallography

The investigation into protein structure using X-ray crystallography can be described using five distinct stages (Figure 3.1); each stage presenting a wide range of experimental choices which can dramatically impact the ability to determine a high-resolution structure (*vide infra*).

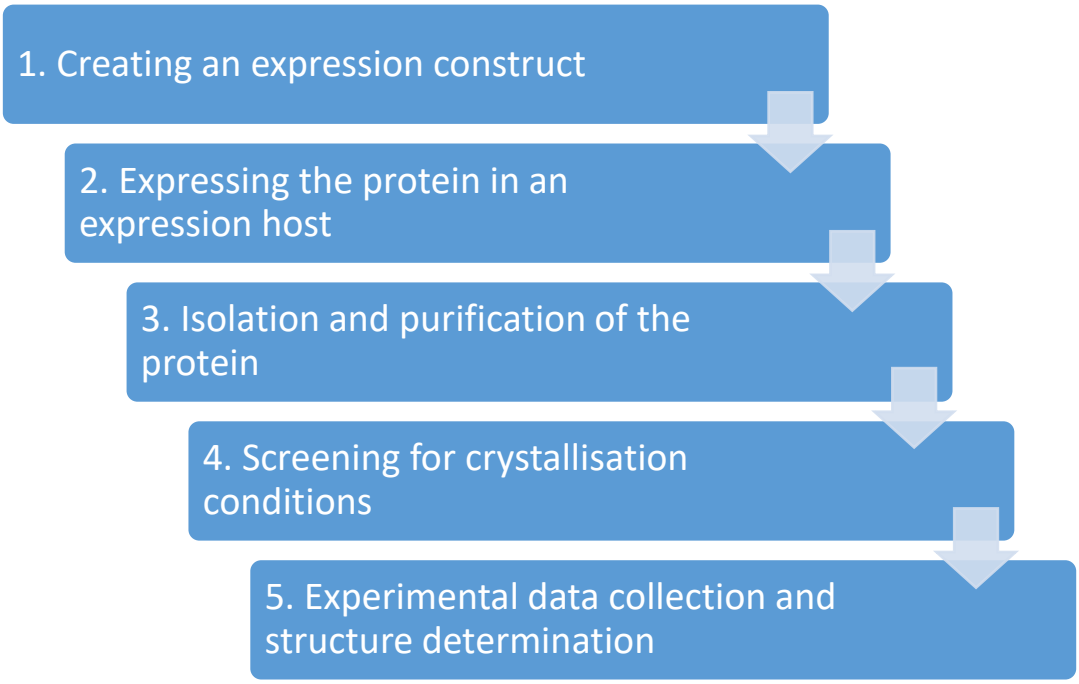


Figure 3.1 Experimental steps in determining protein structure using X-ray crystallography

Specific protocols for the production of high-quality crystals from all proteins cannot be defined, because each protein is unique. In consequence, choices are made at each stage of the procedure, which becomes an iterative process based on the information gathered in preceding stages and the experience gained from previous crystallisation attempts.[260]

3.2.2 Developing expression constructs

An expression vector, is the genetically engineered vehicle used to insert a gene of interest into a cell, and appropriate the protein synthesis machinery for efficient expression of a target protein. Both plasmids and viruses can be engineered to act as expression vectors.

In addition to the gene of interest, these expression vectors also contain regulatory elements such as enhancers and promoters to control protein expression; an origin of replication to allow replication of the plasmid; and several customisable elements *vide infra* (Figure 3.2).

A large number of basic expression vectors, which have been optimised for efficient expression in a range of biological systems, are now available commercially. These facilitate experimentation based on highly polished established systems, without the need to establish the efficacy of the underlying plasmid within a cell.[261]

However, there are several key customisable elements of the expression vector, which must be carefully selected in each case: *viz:* which complementary DNA (cDNA) to use, whether a fusion tag is needed, which cleavage sites to engineer and which antibiotic resistance gene to use.

As expression vector design relies on the knowledge of which expression hosts and purification systems will be needed downstream in the investigation, and in practise is an iterative process. The information gained after an initial investigation with a specific construct, can be used to improve the construct and the cycle repeats until adequate expression has been achieved.[254]

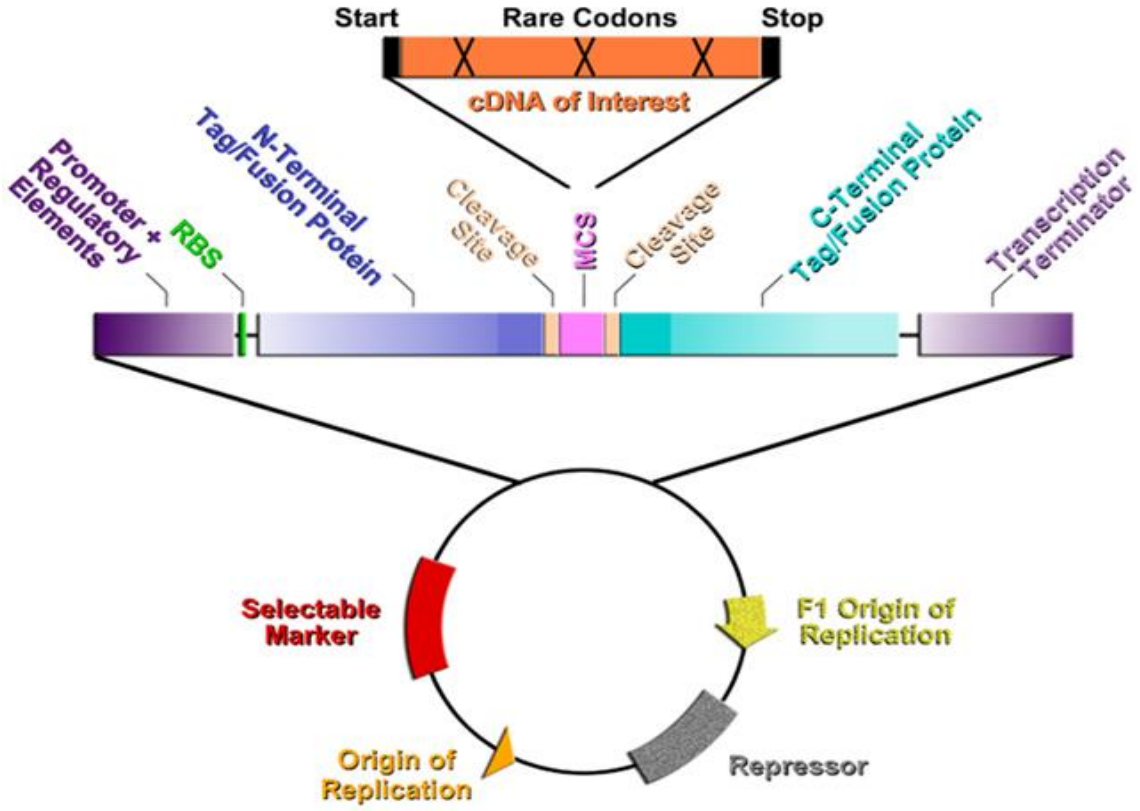


Figure 3.2 An overview of an expression vector

RBS: Ribosome binding site; MCS: multiple cloning site. (Adapted from Tayar & Kleinberger-Doron 2014).[262]

3.2.2.1 cDNA optimisation for the expression host

The gene of interest is translated into the target protein based on a series of nucleotide triplets known as codons; where each codon corresponds to a specific amino acid. During transcription

of the gene, these codons are sequentially bound by tRNA isoacceptors, which facilitates extension of the protein chain with the corresponding amino acid.

The normal codon usage pattern is different between host organisms and is based on the relative abundance of tRNA isoacceptors within the cell; this creates an innate codon bias within each species. Matching the codon distribution within the genetic sequence encoding your protein and the host cell is a crucial determinant of the expression level of the target recombinant protein; as the occurrence of rare codons within a cell will drastically slow the transcription of the gene and can even result in premature termination of transcription.[263]

While this has historically limited the ability to express genes in non-optimal host organisms, the rapid evolution of DNA synthesis technologies has enabled the direct modification of DNAs to adjust for this bias, while maintaining the same final polypeptide product.[264]

Several tools have been developed to facilitate systematic optimisation of codons, such as the software package Codon Optimization OnLine (COOL), which can even allow some degree of optimisation for multiple hosts.[265] Because of the complexity of the protein synthesis machinery and the need to balance codon optimisation with GC content, mRNA secondary structure and stability, some trial and error is often still required for complete optimisation of the cDNA.[266,267]

3.2.2.2 cDNA truncation

In proteins with multiple domains, often not all domains are needed for a specific function. This means that it is sometimes possible to design smaller truncated recombinant proteins consisting of smaller discrete domains, in order to undertake relevant research into certain behaviours of a protein.

As the N and C-terminal boundaries of the expressed protein have significant impact on both its solubility and tendency to aggregate in solution, this can facilitate improvements in downstream protein expression, purification and crystallisation.

The behaviour of the protein is extremely sensitive to changes in domain boundaries, and is subject to change when only a few terminal residues are modified; making boundary selection a crucial step in design.[268]

There are a range of approaches which attempt to predict truncated candidates which are most likely to express successfully, using a combination of sequence predicted domain boundaries and regions of high disorder (see Chapter 2: *Primary-sequence based investigation of PNPLA3*).

However, due to the unnatural state of the protein, any unexpected biological results should be interpreted with caution.[254]

3.2.2.3 Tag / Fusion proteins

Protein tags or fusion proteins, are peptide sequences genetically grafted onto the target protein to allow the expression of a recombinant protein with altered characteristics. There are two main types of tags, solubilisation tags and affinity tags, which can be attached at either the N or C terminal of the protein.

Solubilisation tags are used to assist in the proper folding of proteins and to keep them from aggregating and precipitating. These tags are they are generally large and readily fold into soluble globular forms during overexpression, for example the widely used tag, maltose binding protein (MBP). Due to them often being large tags, solubilisation tags are more likely to impact the native function and folds of the target protein and are therefore more likely to require removal for downstream investigation.[269]

Affinity tags are used to facilitate protein purification using affinity techniques. These tags have an affinity for specific ions or media, for example the widely used poly-histidine tag, which has a high affinity for Ni(II) ions. Due to its small size, it rarely interferes with native function of the protein, and in some cases still allows protein crystallisation to occur.[270,271]

3.2.2.4 Cleavage sites

In cases where use of a tag has been deemed appropriate, it is often desirable to later remove the tag, in order to recover the target protein in a more natural form. While it is not always necessary to remove small tags such as the poly-histidine tag, larger tags can impact on the proteins function and propensity to crystallise and so are best removed.[196]

To facilitate removal of a tag, a protease cleavage site must be engineered into the linking region between the target protein and the tag. This polypeptide sequence should not be present within the target protein to avoid unwanted degradation.

A range of proteases have been engineered to enzymatically remove the tag at the pre-selected cleavage site; these have high efficiency at low temperatures so cleavage can be performed at 4^{o}C to maintain protein stability.[272]

3.2.2.5 Selective resistance

Once the construct has been transformed into the expression host, selective pressure must be used to select for positively transformed cells. This is most frequently achieved through antibiotic resistance, which can allow fast acting selection in most host cells.

The key factor when choosing an antibiotic resistant gene to use, is compatibility with the expression host, to allow selection for only those cells positively transformed. Traditionally ampicillin was the most commonly used antibiotic, although the more stable derivative carbenicillin is recommended for longer experiments.[254]

3.2.3 Protein expression

Recombinant proteins can be expressed in a range of expression host organisms including *E. coli*, yeast, insect cells and mammalian cell culture. However 90% of the protein structures in the protein databank were purified using an *E. coli* expression system.[273]

This is due to the fact protein expression using *E. coli* is inexpensive and able to produce large quantities of protein over very short time periods. In addition, because of the large body of accumulated experience with the use of *E. coli,* extremely robust expression systems have been developed which are much more amenable to investigating novel proteins.[274]

However, there are some potential issues which might limit the use of *E.coli* systems to express mammalian proteins, namely codon bias, a lack of chaperones and post translational machinery.

This has been largely overcome by development of a range of novel recombinant *E. coli* strains. The most commonly used being BL21(DE3), which have been genetically modified to remove over 30 native non-essential *E. coli* proteins, including *lon* and *ompT* proteases, and to be compatible with the T7 *lac*O promotor system.[275,276]

Other strains, have been modified to further adapt for Eukaryotic protein expression, including Rosetta, which include rare codons to correct for the codon bias,[277] and C41 which are more robust against proteins with native toxicity.[278]

3.2.4 Purification techniques

Protein samples obtained from live expression hosts, exist in a complex mixture with a range of other host proteins and chemicals; therefore, requiring purification. The degree of purity required will vary depending on the intended application; purity of 95% or above is generally required for structural studies.

A large range of purification techniques have been developed which rely on separating proteins based on their unique properties. In practice several different purification techniques must be applied sequentially to the same protein sample, in order to achieve the required purity; this often consists of several stages of liquid chromatography.

In liquid chromatography methods, proteins applied in a liquid mobile phase are separated based on their different physiochemical interactions with a stationary phase; these interactions can be based on molecular size, charge, hydrophobicity, or specific binding interactions.

Two specific liquid chromatography techniques, affinity-based chromatography and size exclusion chromatography, are of particular relevance to this study.

3.2.4.1 Affinity chromatography

Affinity chromatography is one of the most versatile and selective forms of liquid chromatography. It relies on the presence of a biologically-related agent in the stationary phase which will bind both selectively and reversibly to the target protein.

When the sample including the target protein is applied to the affinity column (stationary phase) within the mobile phase, the target protein will selectively bind to the stationary phase and be retained; while other elements of the sample are lost. The retained target protein can then be eluted from the column using a different mobile phase (Figure 3.3).[279]

Immobilised metal-ion affinity chromatography (IMAC) is one class of affinity chromatography, which facilitates the binding of targets with electron donor groups to an immobilised metal ion, bound to a chelating agent on the stationary phase support.[280] Although a number of residues can theoretically bind in an IMAC column, binding is almost solely dependent on the availability of histidine residues.[281] Therefore, the addition of poly-histidine tags to a target protein can facilitate IMAC purification.[282]

The most frequently used IMAC method is nickel affinity chromatography which relies on the formation of specific coordinate bonds between the nickel within the column matrix, and the

poly-histidine tag of the target protein.[260] However, other proteins with a high number of accessible histidine residues or metal binding complexes (prevalent in *E. Coli*), may also bind to the column reducing the effectiveness of the technique. This can be circumvented, to a degree, by using Zn(II) ions instead of Ni(ii) in the stationary phase, which have lower binding efficiency to known *E. Coli* proteins; or by using additional purification steps.[283]

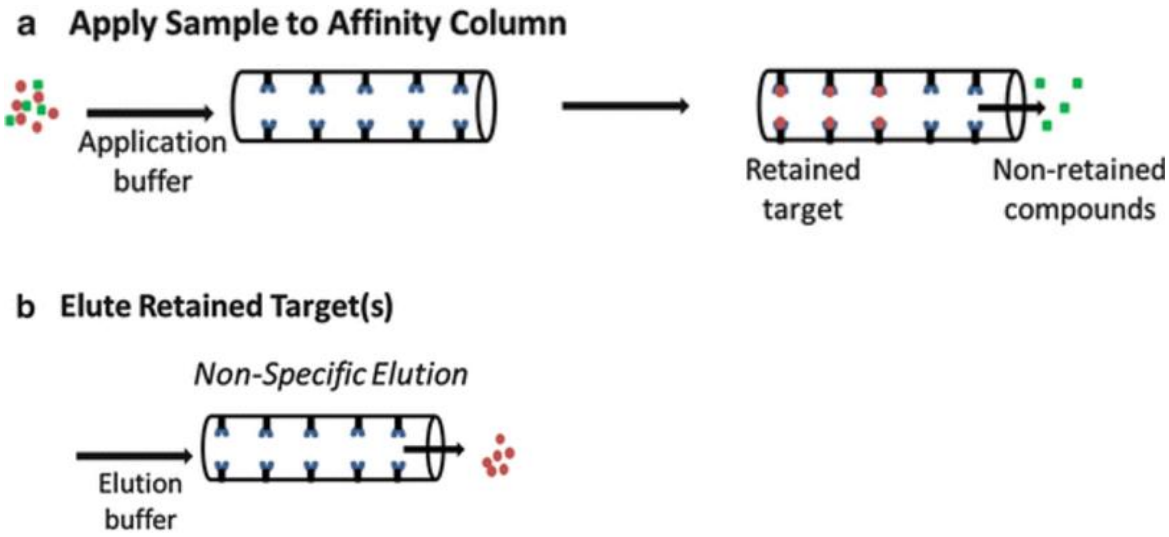


Figure 3.3 Typical application of affinity chromatography

Blue dots within the column matrix represent the affinity matrix; green dots unwanted protein mixture components and red dots the protein of interest (Adapted from Hage & Matsuda 2015).[279]

3.2.4.2 Size exclusion chromatography

Size exclusion chromatography (SEC) uses an inert porous matrix, to separate proteins by size.[284,285] The media used to facilitate separation can vary broadly; however, are commonly formed of irregular porous spheres with varying degrees of crosslinking (Figure 3.4).[286,287]

SEC not only allows separation of proteins but also allows their native molecular size in solution to be estimated. This is facilitated by use of a calibration curve which estimates the relative molecular mass of each protein using the time taken for known reference proteins to elute from the column. Although the separations are often used to describe the molecular mass of studied proteins, this is an approximation based on the Stokes radii, which is the true separating factor.[288–290]

SEC relies on applied proteins within the sample not chemically interacting with the column matrix. In practice, however, non-ideal interactions can impact the efficacy and results of SEC,

by deleteriously affecting the retention time, peak shape and recovery of the protein; this must be carefully addressed and minimised through matrix selection.[291]

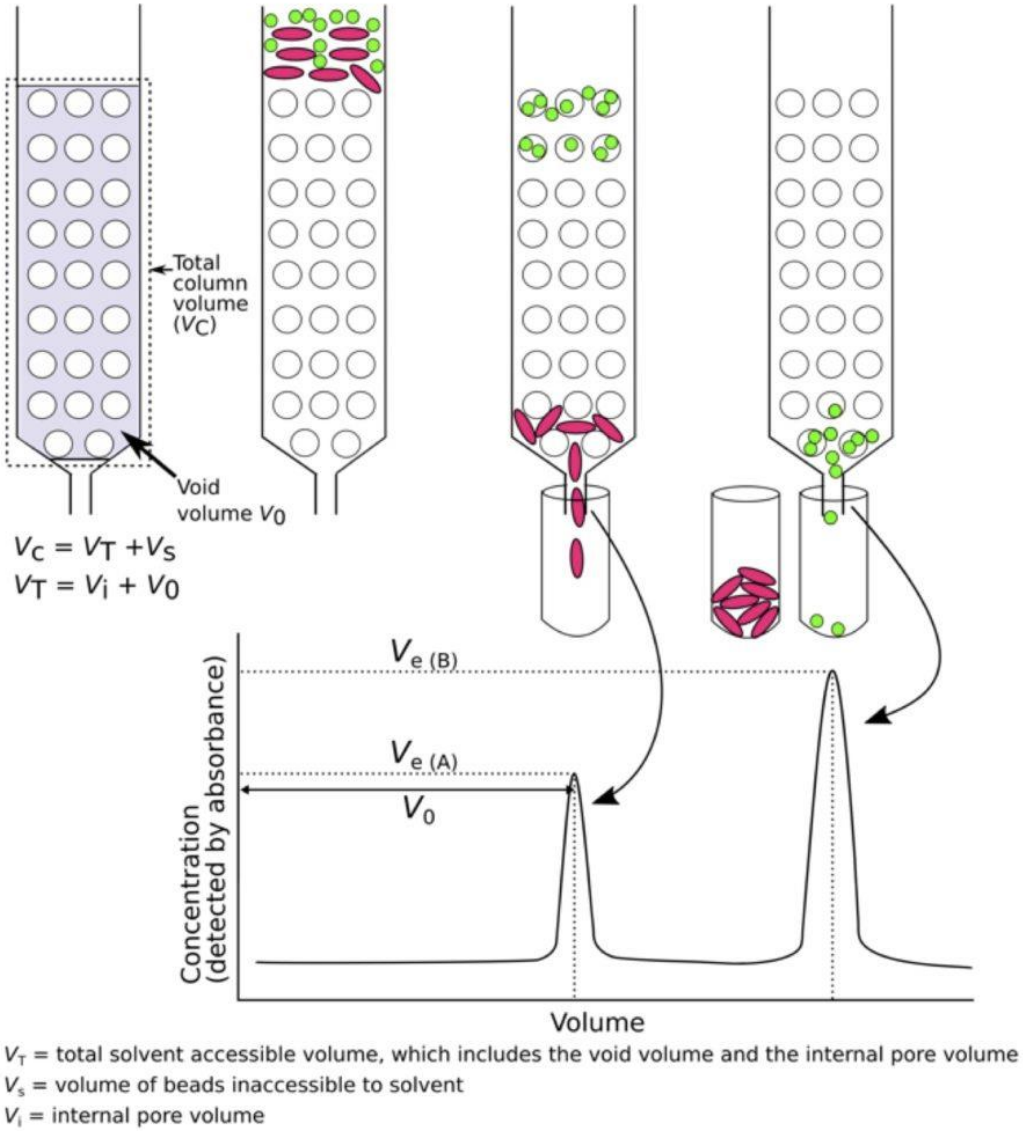


Figure 3.4 Time sequence representing methodology of size-exclusion chromatography

(Adapted from BitesizeBio accessed: 2018).[292]

Because both affinity chromatography and SEC select for different physiochemical properties of the target protein, the efficacy of purification is greatly improved by using both methods in combination.

The ability of affinity chromatography to rapidly purify proteins from a highly contaminated solution means this is usually used as an early step in the purification process; this helps

eliminate the majority of contaminating proteins facilitating the use of more precise polishing steps. SEC is often used as a refining step as it is not able to efficiently separate proteins of similar size. However, it has the advantage of not only separating different proteins but different multimeric forms within the same sample, producing samples of a single oligomer.

3.2.5 Analytical methods

Protein identity and purity is tracked throughout the purification process, often using a combination of sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and western blots. This allows identification of when a sample has reached a purity deemed adequate for further functional and structural investigation.

3.2.5.1 Sodium dodecyl sulfate poly acrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE is a destructive technique for the efficient separation of proteins based on their molecular weight (Figure 3.5).

The sample, containing the target protein, is heated in excess SDS and reducing agent to fully denature the proteins. Under these conditions, proteins will bind molecules of SDS in a constant weight ratio; resulting in a mixture of proteins which have an almost constant charge density, similar shapes, and thus, similar electrophoretic mobility.[293]

A polyacrylamide gel is formed by the reaction of acrylamide and bis-acrylamide to form a highly crosslinked gel matrix. The SDS denatured sample is added to the gel matrix and a current applied, causing the proteins to migrate toward the anode.

A sieving mechanism will occur based on the matrix pore size, which causes smaller peptides to migrate faster through the gel than the larger peptides.

Once the proteins have migrated adequately to achieve visible separation, the proteins can be stained in the gel using a range of protein dyes. The most common stain is Coomassie blue, which can detect as little as 30ng of protein.[294] A ladder containing proteins of known masses is run adjacent to the sample and is used to predict the molecular weights of the unknown proteins.

The purity of the sample is often predicted based on the SDS-PAGE results, although this does not account for the large number of contaminants present below the detection limit. While SDS-

PAGE presents an accurate prediction of the molecular mass of a target protein, the prediction will almost always differ slightly from the true molecular mass due to additional interactions within the sample.[295]

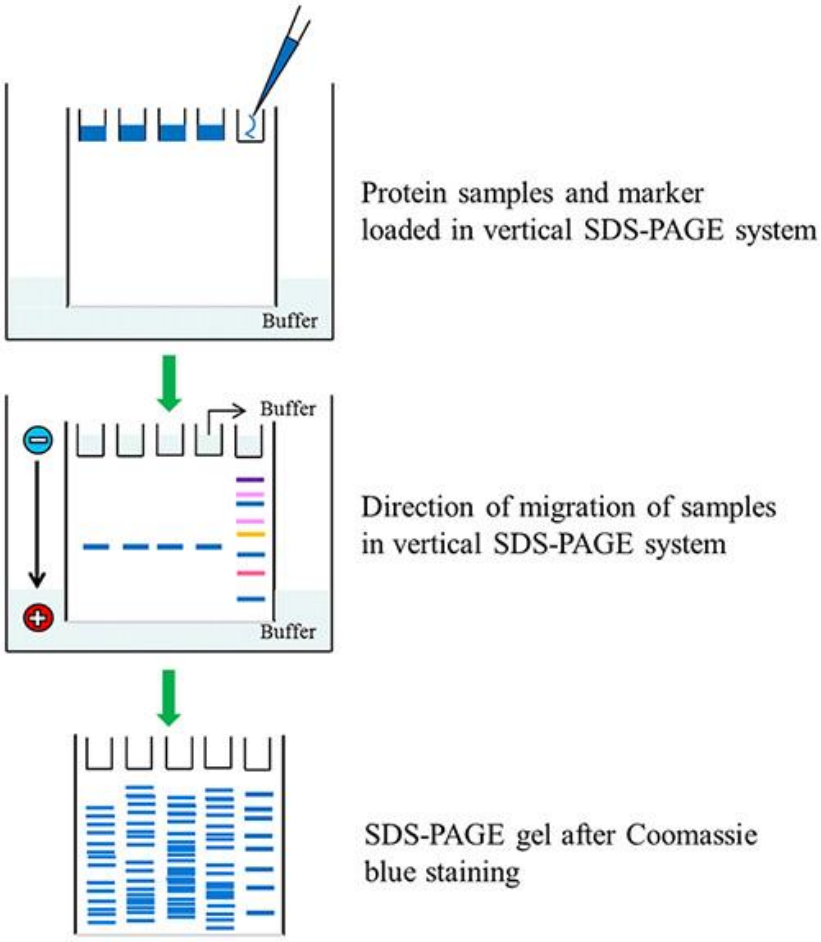


Figure 3.5 Theory of SDS-PAGE of sample protein

(Adapted from Sigma Aldrich accessed:2017).[296]

3.2.5.2 Western blotting

Western blotting is a highly sensitive analytical technique which can be used to detect the presence or absence of a specific target protein.[297]

A western blot involves the transfer of proteins separated by SDS-PAGE, from the polyacrylamide gel to a nitrocellulose membrane, replicating the original gel pattern. The membrane pores are then blocked, to inhibit non-specific interactions between other proteins and the membrane.

In order to identify the presence of a target protein, the membrane is probed using specific labelled antibodies, which can undergo detection. This can either be done in a single step; by which a labelled primary protein specific antibody is raised directly to the protein, or in two steps; where a labelled secondary antibody is used to bind a non-labelled primary antibody (Figure 3.6).

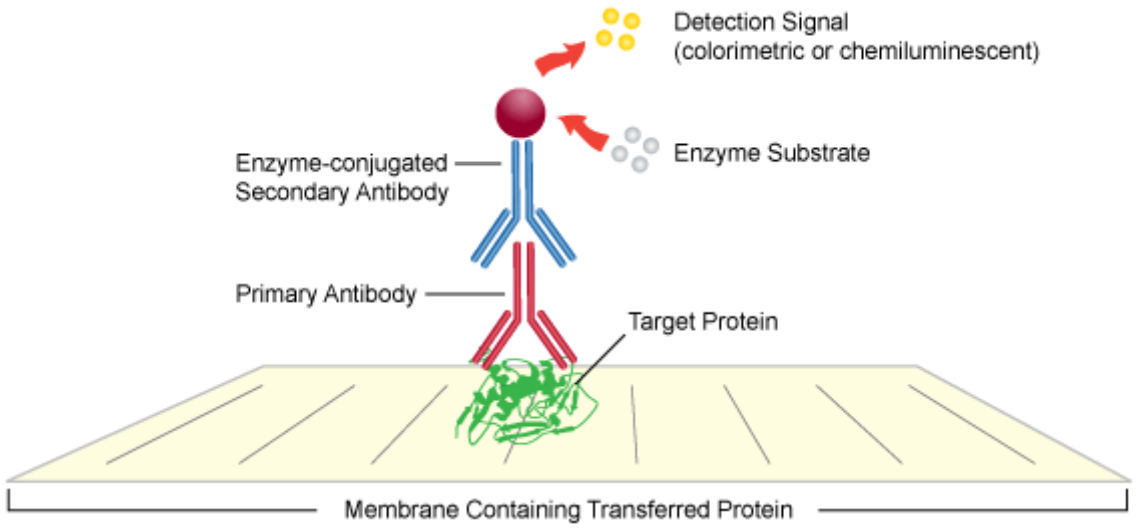


Figure 3.6 The detection of proteins using western blots

(Adapted from Leinco accessed: 2017)[298]

Antibodies can be labelled with a range of elements, including enzymes, fluorophores, biotin, gold, or radioisotopes. The enzyme Horse-radish peroxidise is commonly used as a label as it is able to oxidase a range of substrates to produce a colorimetric or chemiluminescent signal, which can be measured through UV detection, fluorescence or visible inspection.

Western blotting does rely on the availability of antibodies to the target protein. However, antibodies have been raised to poly-histidine tags, which can be used to detect tagged proteins when specific antibodies are not available.

3.2.6 Protein crystallisation

A protein crystal is a unique structural protein state, which consists of a highly ordered crystalline solid. Protein crystallisation is a complex multiparametric technique, which facilitates the formation of protein crystals from a purified protein solution. For crystallographic investigation into a structure, obtaining a high fidelity crystal is a vital yet challenging step.

Despite the experimental importance, protein crystallization has only recently begun to be understood in greater depth because of its complex multiparametric nature.[299]

3.2.6.1 Theory of crystallisation

The crystallisation of a protein from highly purified protein solution requires the formation of a solid phase which adopts a high degree of internal three-dimensional order.[300] This process is driven by the minimisation of free energy within the system, by forming multiple new stable non covalent intermolecular chemical bonds, when there is insufficient solvent to maintain full hydration of the protein.[301]

The process of crystallisation has three distinct stages: nucleation, growth, and cessation of growth.

Nucleation is the process by which the nuclei of the crystal is formed and is the most poorly understood step in crystal formation. While it is clear an ordered nucleus is a requirement for further crystal growth, it has not yet been determined whether nuclei are formed by strict monomer or oligomer addition or the coalescence of sub-nuclear structures. Notably, unstructured protein precipitants are believed to form in a similar fashion, however, instead of ordered regular interactions, they exhibit random intermolecular interactions. This allows precipitants to generally form much faster and therefore dominate the protein solid phase, over the rarer crystal forms.[302]

The growth phase is a phase of the process in which additional protein subunits are added to the surface of the crystal, thereby increasing its size. The rate of growth is highly dependent on the nature of the crystal as well as the conditions of the solution. In general, growth of crystals with a rough surface requires less energy than from a smooth surface, facilitating faster crystal growth.[303]

Cessation of growth, the final phase of the process, can occur for a number of reasons. First, the concentration of protein in the solution can decrease until the solid and liquid phases reach an exchange equilibrium. This is accelerated with fast growing crystals, where a local decrease in protein concentration is observed around the protein crystal during formation.

Second, growth may cease as a result of crystal poisoning the process by which damaged or non-identical subunits are incorporated into the crystal until successive defects interrupt the crystal lattice. This is a common problem and highlights the importance of using highly purified protein solutions for the crystallisation step. Crystal poisoning not only limits the size of the crystals

obtained but, depending on its extent, may have downstream effects on the quality and resolutions of the crystal diffraction.[304]

The process of growing protein crystals can be summarised by use of a phase diagram (Figure 3.7) which divides the protein containing solution into four main phases; the stable phase, metastable phase, labile phase and precipitation phase.

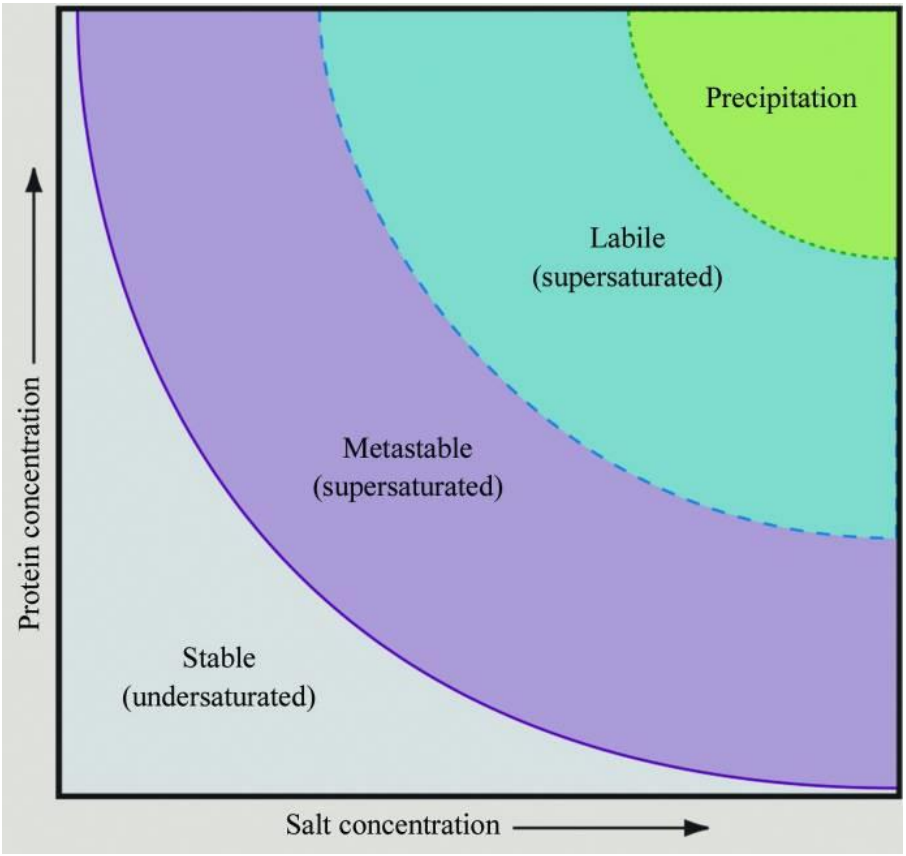


Figure 3.7 The phase diagram for the crystallisation of proteins

(Adapted from McPherson & Gavira 2014).[305]

The stable phase, also known as the under saturated state, is the typical state of the protein in solution, where the protein concentration is below the solubility limit.

The supersaturated state, is a non-equilibrium condition where the macromolecule is in excess of the solubility limit, but none-the less still in solution. This state is facilitated in the space between the solubility limit and the kinetic energy barrier between forming the second phase. Nucleation and growth of crystals are both critically dependent upon reaching this supersaturation state.

The supersaturation state includes both the metastable and labile zones. In the labile region, both nucleation and crystal growth can occur, while only crystal growth occurs in the metastable region.

Finally, in the precipitation region, the protein will aggregate, and form disordered solids. Achieving quality crystals is dependent upon being able to produce and maintain the labile and metastable supersaturated states.[305]

3.2.6.2 Experimental crystallisation

There are a number of techniques which can be applied to help grow protein crystals including vapour diffusion, solvent diffusion, and vacuum sublimation.

The most widely used is vapour diffusion in which a droplet of solution containing both the protein of interest and the determined precipitant is suspended over a reservoir containing a more concentrated solution of the same precipitant, in a closed system. As the reservoir draws water from the drop via air-gap separation, the protein concentration in the droplet gradually increases until the supersaturated region is reached (Figure 3.8) .[306]

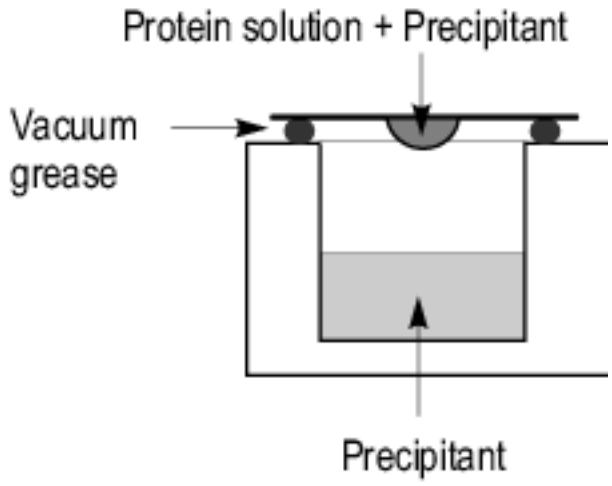


Figure 3.8 Process of hanging drop vapour diffusion crystallisation technique

(Adapted from Chirgadze 2001)[307]

The stages of crystal growth during the process of vapour diffusion can be visualized on the phase diagram (Figure 3.9).

The screen is set up, with protein concentrations in the under saturated state. As the experiment proceeds and the protein concentration within the droplet increases, the solution will enter the supersaturation state. At this stage, if the conditions are correct, nuclei will form, and crystals will begin to grow. As the crystals grow in solution, the protein concentration will decrease; however, the precipitant concentration continues to increase, which aids in a prolonged maintenance of the supersaturated state, to maximize the crystal growth.[308,309]

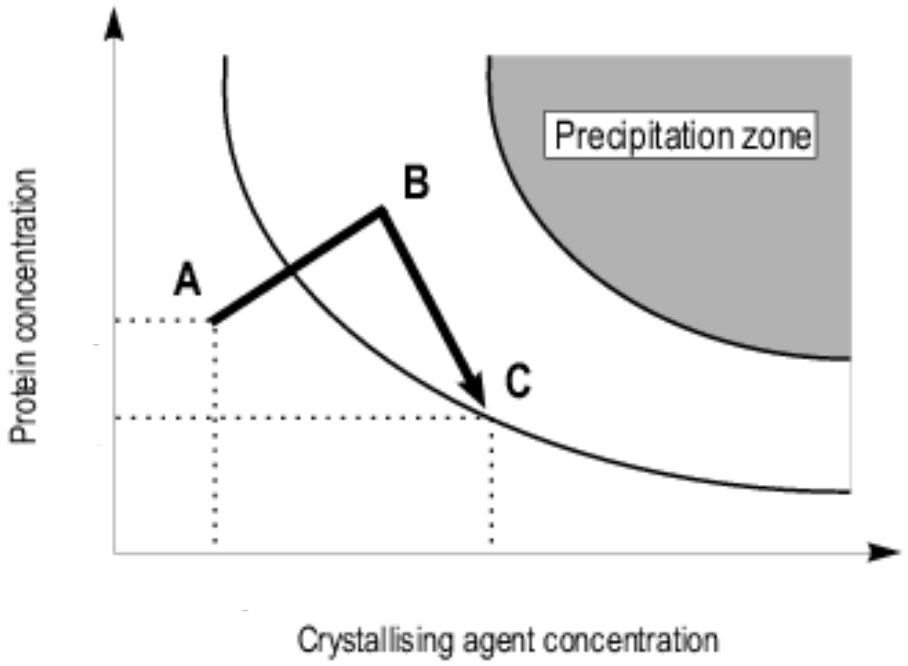


Figure 3.9 Theory of crystal formation in vapour diffusion X-ray crystallography

A: Conditions at the start of the crystallisation experiment.
B: Conditions at which nucleation occurs.
C: Conditions at which cessation of growth occurs
(Adapted from Chirgadze 2001)[307]

A significant number of factors can impact protein crystallisation (Table 3.1). In general, protein concentration, pH, salt concentration and the precipitant selected are often deemed the most important factors, however, other rarely used parameters can be critical in certain cases.

Each of these variables can affect the overall probability of crystal formation, nucleation rate, growth rate and the ultimate size and quality of the crystal. Finding crystallisation conditions has remained the rate limiting step of structural investigations, due to the large number of conditions which must often be screened before crystallisation is acheived.[305]

Table 3.1 Factors affecting crystallisation

Physical	Chemical	Biochemical
Temperature/temperature variation	pH	Purity of the macromolecule/nature of impurities
Surfaces/heterogeneous nucleants	Precipitant type	Ligands, inhibitors, effectors
Methodology/approach to equilibrium	Final precipitant concentration	Aggregation state of the macromolecule
Mother-liquor volume	Ionic strength	Post-translational modifications
Geometry of chamber or capillary	Cation type and concentration	Source of macromolecule
Gravity	Anion type and concentration	Proteolysis/hydrolysis
Pressure	Degree of supersaturation	Chemical modifications
Time	Reductive/oxidative environment	Genetic modifications
Vibrations/sound/mechanical perturbations	Concentration of the macromolecule	Inherent symmetry of the macromolecule
Electrostatic/magnetic fields	Metal ions	Degree of denaturation
Dielectric properties of the medium	Initial precipitant concentration	Isoelectric point
Viscosity of the medium	Cross-linkers/polyions	Unstructured regions
Rate of equilibration	Detergents/surfactants/amphophiles	purification tags
	Non-macromolecular impurities	α-Helix content
	Chaotropes	Conformational states
		Thermal stability
		Allowable pH range
		History of the sample

Adapted from McPherson & Gavira 2014.[305]

The introduction of both screening kits and high throughput screening robots have made the identification of crystallisation conditions much faster, particularly when they deviate from typical condition variations. The potential to screen conditions in this way has facilitated the crystallisation of large numbers of proteins for which crystallisation conditions are otherwise unlikely to have been found.[306]

The Mosquito from TTP LabTech is one such robot, which allows the automated set up of 96 well vapour diffusion plates. Because of the disposable plunger tip design, it can achieve highly accurate pipetting down to nanolitre volumes, and avoids the challenges needed to wash tips.

The development of such high throughput techniques allows the screening of thousands of conditions in a single day, and requires significantly less protein sample and screening buffer than manual screening.[310]

There are now over 15,000 commercially available screening conditions which are available for use in the high throughput screening for crystallisation conditions. This does contain many overlapping conditions; however, attempts are continuously made to build upon past success and reduce conditions only to those most likely to produce crystals.

Using commercially available high throughput screens has proven very successful. This can be seen by observing data in the PDB, where it was shown 37% of crystals were obtained in a condition which varied by less than 0.1 in a single condition to a commercial counterpart.[311]

While this can create higher efficiency in screening by only using conditions which have proved effective in the past, if used alone, risks continued crystallisation of proteins similar to those already crystallised. This means novel proteins with no known structural homologues may require follow up screening with a broader array of conditions.

Even with the development of high-throughput screening methods, crystallisation often remains the slowest investigatory step. Attempts are being made to address this by making computational predictions of crystallisation conditions, rather than the purely trial and error screening approach; however, this is not yet a viable with any current tools.[312]

3.2.7 Expression and purification of PNPLA3

To date PNPLA3 has been successfully expressed in a wide range of expression hosts, including *E. coli*,[166] *P. pastoris*,[164] SF9 insect cells[165] and human cells including HEK293[151] and HuH-7.[158] In addition a number functional *in vitro* and *in vivo* studies have been undertaken. Nevertheless, PNPLA3 remains only partially characterised biochemically. The inability to produce the quantity and purity of PNPLA3 needed for structural investigation has hindered progress into understanding the link between PNPLA3 and disease.

The purification of PNPLA3 in previous studies has so far always involved a single crude purification step to prepare samples for functional assays. Initial experiments have used both immunoprecipitation with an anti-V5 mouse monoclonal antibody,[151] and anti-FLAG affinity chromatography with partial success.[165] However, more recent publications have shown great promise in achieving up to 90% purity with a single Ni-affinity chromatography step.[6,164,166]

3.3 Aims

The overarching aim of this chapter was to gain insight into the function and structure of PNPLA3 using an *in vitro* approach.

This will involve the completion of three distinct steps: 1) Developing a reproducible protocol to enable the production of milligram quantities of PNPLA3 with purity above 95%. 2) Developing an *in vitro* assay to measure the activity of PNPLA3. 3) Solving an experimental structure of PNPLA3.

Insights learned from this investigation can later be used to improve understanding of PNPLA3 and the I148M variant, as well as provide a framework for future drug discovery investigations.

3.4 Methods

3.4.1 Expression constructs

The protein sequences of human PNPLA3 (Uniprot Accession: Q9NST1) and murine pnpla3 (Uniprot Accession: Q91WW7) were retrieved from Uniprot database (Figure 3.10).[121] All truncated variants of PNPLA3 were based on these sequences.

Human PNPLA3 sequence

MYDAERGWSLSFAGCGFLGFYHVGATRCLSEHAPHLLRDARMLFGASAGALHCVGVLSGIPLEQTLQVLS
DLVRKARSRNIGIFHPSFNLSKFLRQGLCKCLPANVHQLISGKIGISLTRVSDGENVLVSDFRSKDEVVD
ALVCSCFIPFYSGLIPPSFRGVRYVDGGVSDNVPFIDAKTTITVSPFYGEYDICPKVKSTNFLHVDITKL
SLRLCTGNLYLLSRAFVPPDLKVLGEICLRGYLDAFRFLEEKGICNRPQPGLKSSSEGMDPEVAMPSWAN
MSLDSSPESAALAVRLEGDELLDHLRLSILPWDESILDTLSPRLATALSEEMKDKGGYMSKICNLLPIRI
MSYVMLPCTLPVESAIAIVQRLVTWLPDMPDDVLWLQWVTSQVFTRVLMCLLPASRSQMPVSSQQASPCT
PEQDWPCWTPCSPKGCPAETKAEATPRSILRSSLNFFLGNKVPAGAEGLSTFPSFSLEKSL

Murine pnpla3 sequence

MYDPERRWSLSFAGCGFLGFYHVGATLCLSERAPHLLRDARTFFGCSAGALHAVTFVCSLPLGRIMEILM
DLVRKARSRNIGTLHPFFNINKCIRDGLQESLPDNVHQVISGKVHISLTRVSDGENVLVSEFHSKDEVVD
ALVCSCFIPLFSGLIPPSFRGERYVDGGVSDNVPVLDAKTTITVSPFYGEHDICPKVKSTNFFHVNITNL
SLRLCTGNLQLLTRALFPSDVKVMGELCYQGYLDAFRFLEENGICNGPQRSLSLSLVAPEACLENGKLVG
DKVPVSLCFTDENIWETLSPELSTALSEAIKDREGYLSKVCNLLPVRILSYIMLPCSLPVESAIAAVHRL
VTWLPDIQDDIQWLQWATSQVCARMTMCLLPSTRSRASKDDHRMLKHGHHPSPHKPQGNSAGL

Figure 3.10 PNPLA3 and pnpla3 protein sequences retrieved from Uniprot database.

Each letter corresponds to a single amino acid within the protein sequence.

Ninety-six *PNPLA3* DNA constructs were available from a previous high-throughput expression screen at the Oxford Protein Production Facility, UK (OPPF-UK) (Appendix II). The DNA constructs consist of varying length fragments of *PNPLA3* cloned into a range of vectors from the OPPF-UK pOPIN vector suit based on a commercial pTriEx2 plasmid (Novagen, Nottingham, UK).

The vectors share core features for recombinant protein expression *viz:* pUC origins of replication for high-copy replication in *E. coli*; the ampicillin resistance gene (AmpR) for positive selection; T7 promoter, T7 terminator, a lac operator and a lacZ promotor for enhanced

translational control; and a poly-histidine tag for purification. The vectors differ only in the recombinant protein tag which is co-expressed with the gene of interest.

Eleven of these DNA constructs had prior evidence of low level protein expression, suggesting the potential for optimisation and were selected for further investigation (Table 3.2).[313] These constructs were cloned into one of four pOPIN vectors, viz: pOPINM, pOPINS3C , pOPINE-3c-eGFP and pOPINE-3c-HALO7.[314] Each of these has a tag which is used for enhanced stability and folding of proteins.[269]

Table 3.2 PNPLA3 expression constructs from the Oxford Protein Production Facility, UK

Expression construct	Plasmid	Protein insert	Expected molecular weight
A4	pOPINM	full length PNPLA3	96.3
A6	pOPINS3C	300-400	24.5
B4	pOPINM	1-175	62.5
B6	pOPINS3C	271-400	27.5
D1	pOPINE-3C-eGFP	NA GFP control	27
D2	pOPINE-3C-eGFP	271-400	42.5
D4	pOPINM	300-481	63.5
E4	pOPINM	300-400	55
F2	pOPINE-3C-HALO7	300-481	53
F4	pOPINM	271-400	15.5
G1	pOPINE-3C-eGFP	2-481	81
pnpla3	pCold TF	1-413	94.8

pOPINM has an N-terminal maltose binding protein (MBP) tag and poly-histidine tag with an engineered 3C protease cleavage site. pOPINS3C has an N-terminal SUMO and poly-histidine tag with an engineered 3C protease cleavage site. pOPINE-3c-eGFP has a C-terminal green fluorescent protein and poly-histidine tag with an engineered 3C protease site. pOPINE-3c-eGFP has a C-terminal green fluorescent protein and poly-histidine tag with an engineered 3C protease site. pOPINE-3c-HALO7 has a C-terminal halo and poly-histidine tag with an engineered 3 protease site (Figures 3.11-3.14).

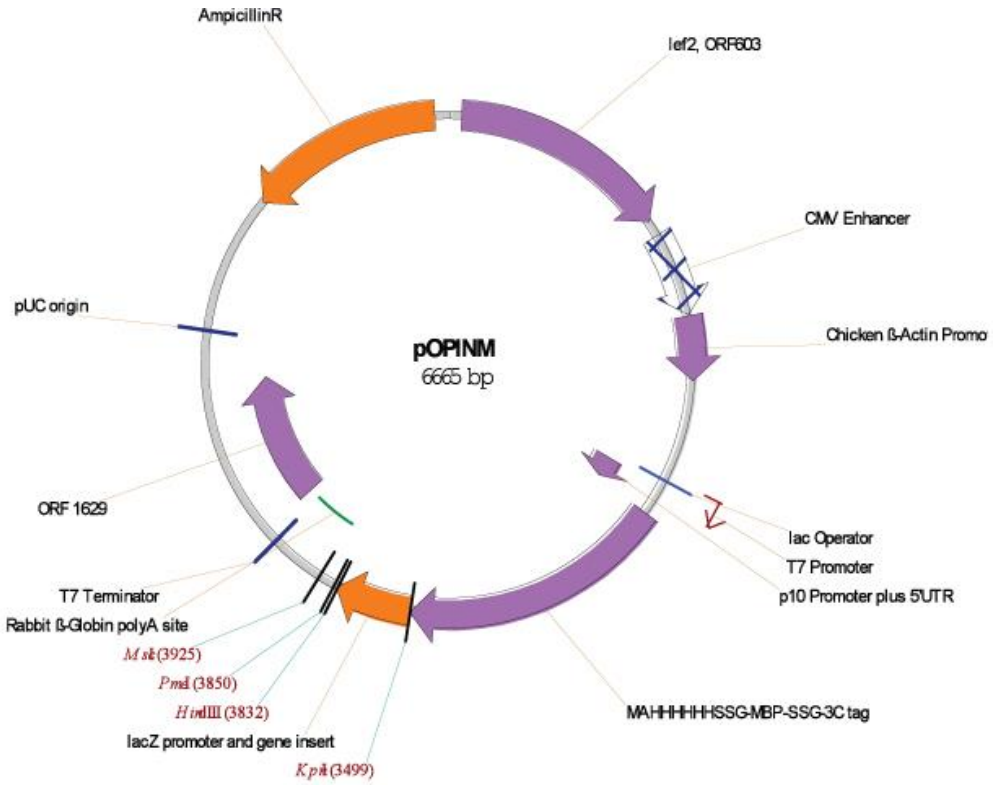


Figure 3.11 Plasmid map of pOPINM.

Adapted from Addgene plasmid #26044.[315]

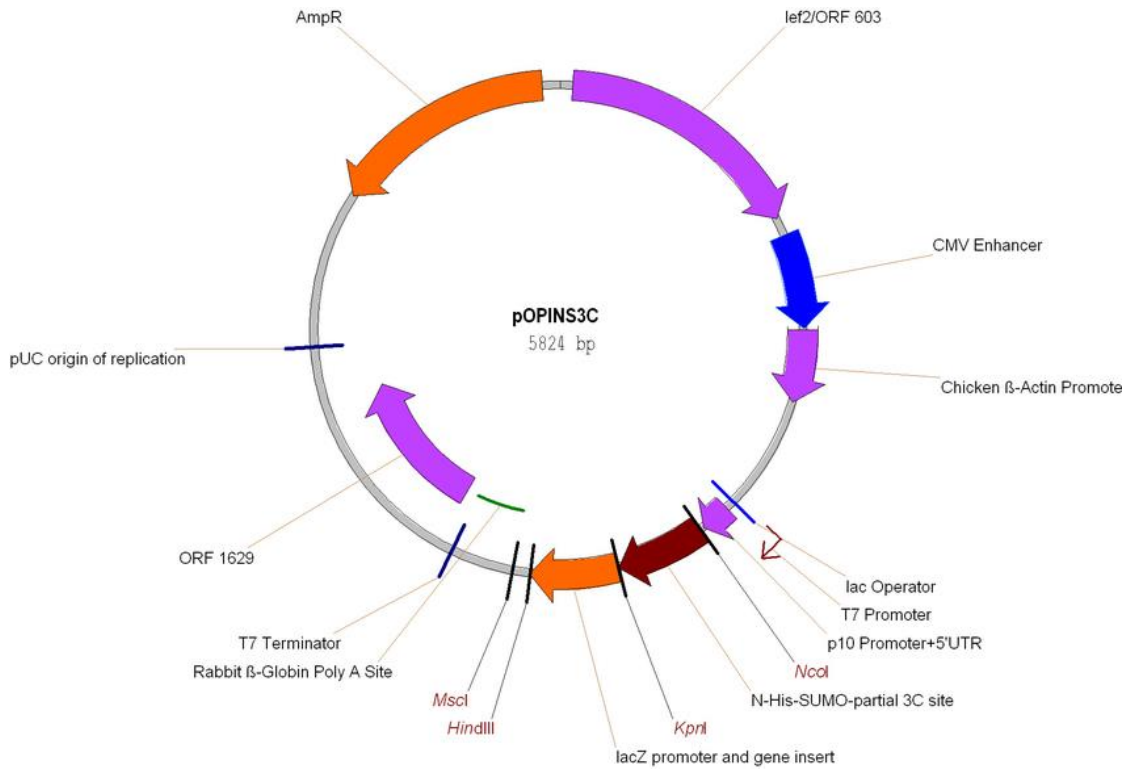


Figure 3.12 Plasmid map of pOPINS3C.

Adapted from Addgene plasmid #41115.[315]

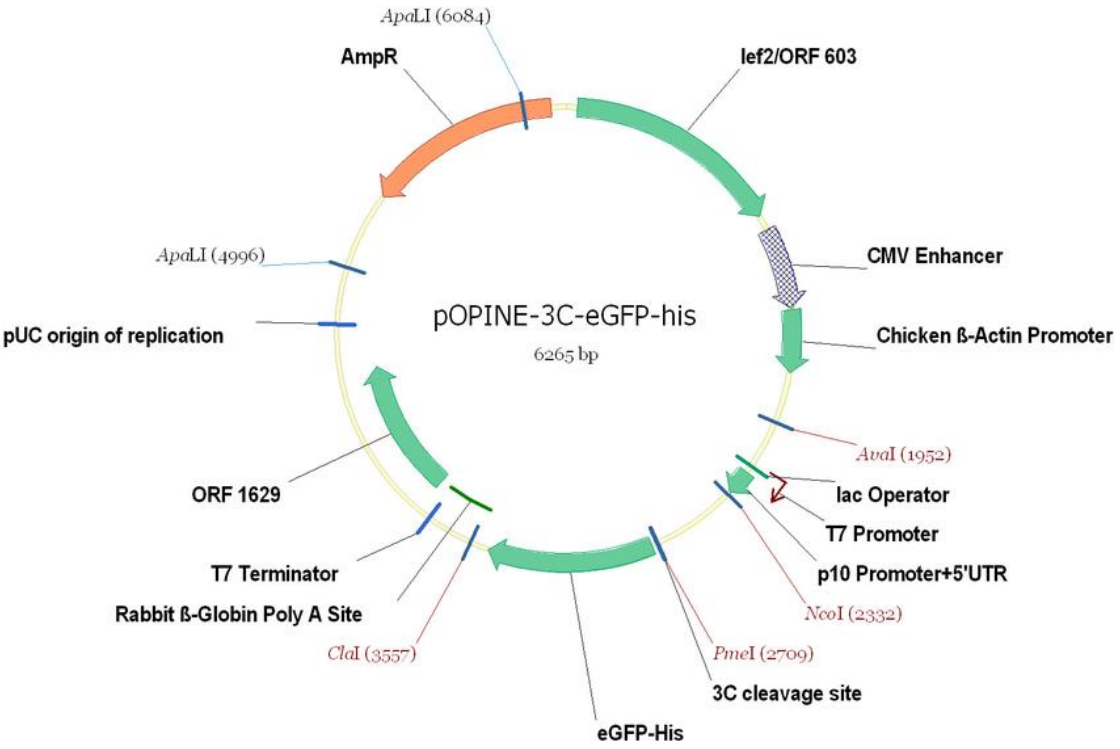


Figure 3.13 Plasmid map of pOPINE-3C-eGFP-his.

Adapted from Addgene plasmid #41125.[315]

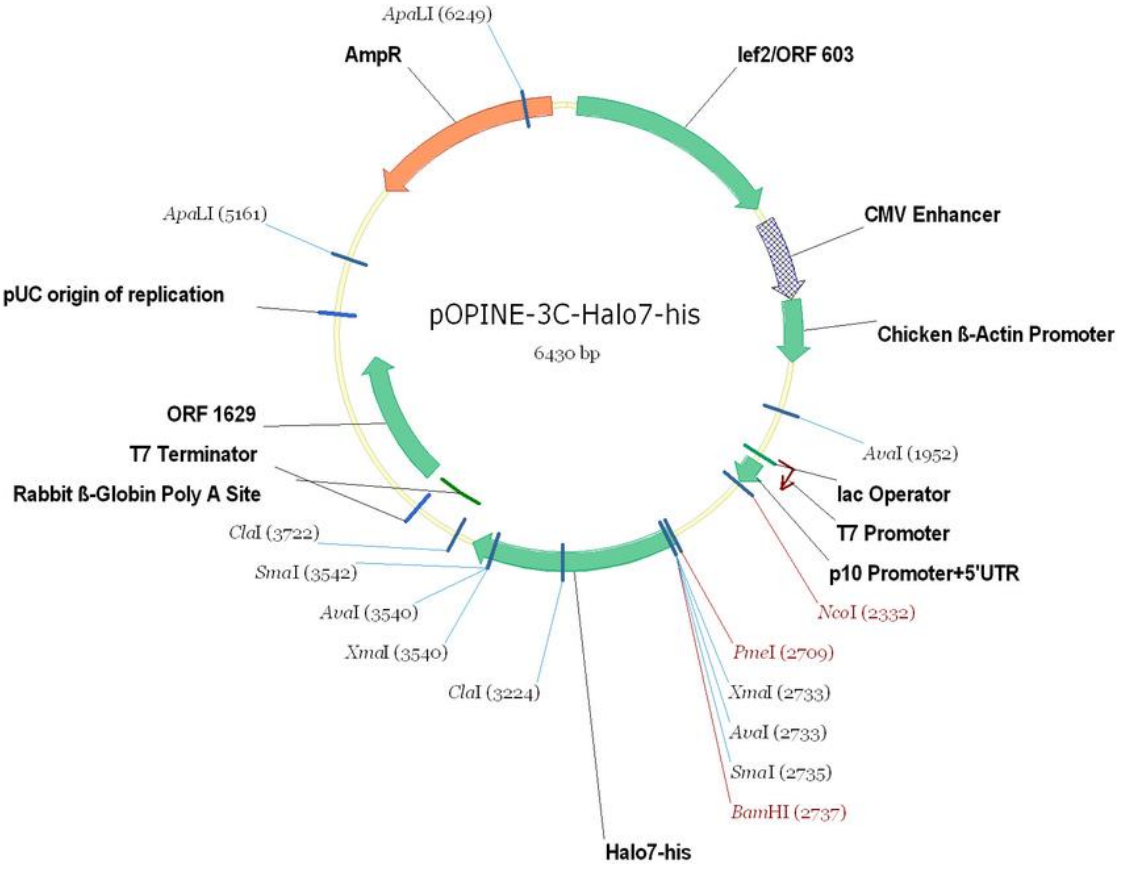


Figure 3.14 Plasmid map of pOPINE-3c-Halo7.

Adapted from Addgene plasmid #41126.[315]

An additional construct using the pCold expression vector containing an mPNPLA3 insert was generously donated by the Institute of Molecular Biosciences, University of Graz (Figure 3.15).

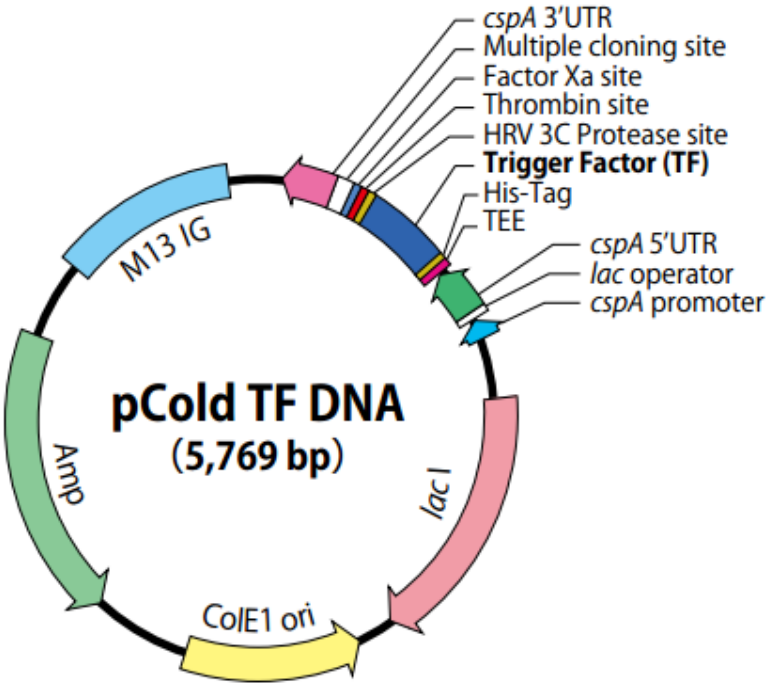


Figure 3.15 Plasmid map of pCold TF plasmid.

Adapted from Takura Bioscience.[316]

This was included as the murine protein is a shorter but highly homologous protein which has previously shown expression within an *E. coli* host and hence has a greater chance of crystallisation. Of note, this construct contains trigger factor (TF) which is thought to facilitate correct protein folding by shielding hydrophobic regions of the protein from the cytoplasm and slowing protein folding.[317]

3.4.2 Plasmid DNA isolation

3.4.2.1 Transformation.

All clones were transformed into DH5α high efficiency *E. coli* using the standard Novagen heat shock protocol. Transformed *E. coli* were plated with a standard spread plating technique onto LB Agar (Fluka Analytical) supplemented with 100 µg ml^{-1} ampicillin and incubated overnight at 37°C (Appendix II).

3.4.2.2 Miniprep

Liquid cultures containing 10ml LB broth (Miller) supplemented with 100 µg ml^{-1} ampicillin were inoculated with single colonies from the previously transformed plates using a sterile filter tip. Cultures were incubated overnight (18 hours) at 37°C with shaking at 120rpm.

The plasmid DNA was extracted from the cells using the Axygen Axyprep plasmid mini-prep kit, using the manufacturer's guidelines (Appendix II).

3.4.3 Expression trials

3.4.3.1 Transformations

All clones were transformed into Rosetta 2(DE3)pLysS (Novagen), BL21(DE3) and OverExpress C41(DE3)pLysS *E. coli* strains using the standard Novagen heat shock protocol. Transformed *E. coli* were plated with a standard spread plating technique onto LB Agar (Fluka Analytical) supplemented with 100 µg ml^{-1} ampicillin and incubated overnight at 37°C (Appendix II).

LB agar used for Rosetta 2(DE3)pLysS (Novagen) and BL21(DE3) were additionally supplemented with 34 µg ml^{-1} chloramphenicol.

Successive trials were performed in order to obtain optimal conditions for protein expression and purification (Figure 3.16).

3.4.3.2 Initial screening for overexpression

Liquid starter culture containing 10ml LB broth (Miller) supplemented with 100 µg ml^{-1} ampicillin were inoculated with single colonies from previously transformed plates, using a sterile filter tip. Cultures were incubated overnight (18 hours) at 37°C with shaking at 120rpm.

Fifty µl of the starter culture were inoculated into 5ml LB broth supplemented with 100 µg ml-1 ampicillin. The samples were grown at 37°C to an optical density (OD) of 0.6 and induced with 0.5mM isopropyl β-D-1-thiogalactopyranoside (IPTG). and incubated for an additional 18 hours at 18°C. Twelve clones were tested for small scale expression: A4, A6, B4, B6, D1, D2, D4, E4, F2, F4, G1 and Pnpla3.

The growth media used for Rosetta 2(DE3)pLysS (Novagen), BL21(DE3) were additionally supplemented with 34 µg ml^{-1} chloramphenicol at each step.

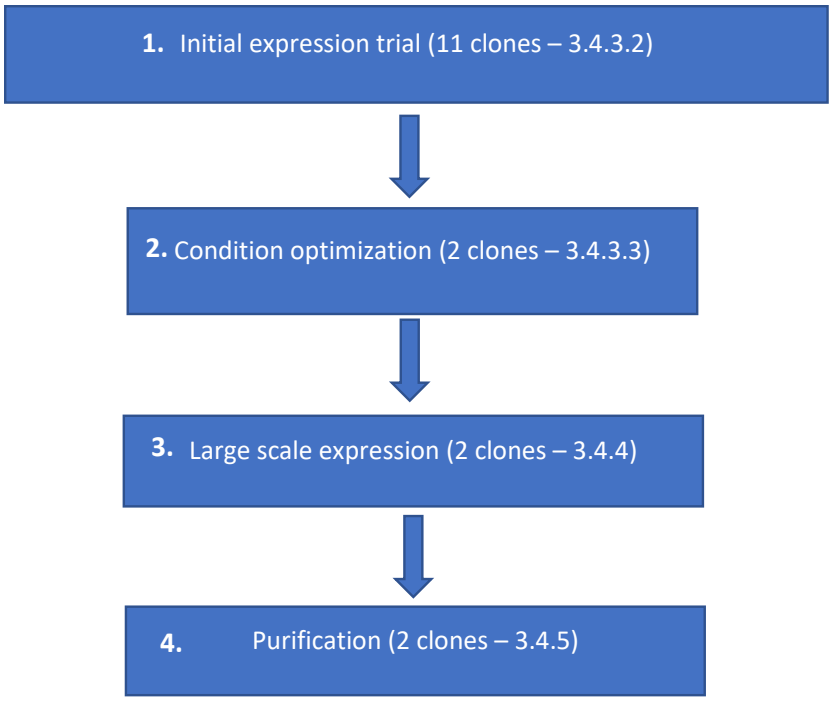


Figure 3.16 Optimisation overview for protein expression and purification
Each successively numbered step was used to identify the best conditions for expression which were taken forward into the next step. The number of clones taken forward are highlighted and the section corresponding to each method is numbered on the diagram.

3.4.3.3 Condition screening for PNPLA3 A4 and pnpla3 clones

Liquid starter culture containing 10ml LB broth (Miller) supplemented with 34 µg ml^{-1} chloramphenicol and 100 µg ml^{-1} ampicillin were inoculated with a single colony from a previously transformed plate of clone A4, using a sterile filter tip. Cultures were incubated overnight (18 hours) at 37°C with shaking at 120rpm.

Fifty µl of the starter culture were inoculated into 5ml LB broth supplemented with 30 µg ml-1 chloramphenicol and 100 µg ml-1 ampicillin.

Four growth conditions were used to test for expression of protein product:

Grown overnight at 37°C, induced at OD 1.4 for 16 hours at 15°C.

Grown overnight at 37°C, induced at OD 1.4 for 24 hours at 15°C.

Grown overnight at 27°C, induced at OD of 0.6 for 16 hours at 15°C.

Grown overnight at 27°C, induced at OD of 0.6 for 24 hours at 15°C.

3.4.4 Large scale protein expression

Protein expression was replicated on a larger scale using the most promising human (A4) and murine clones. These were run in parallel with the murine line of experimentation running slightly ahead.

Transformations were performed as above (see 3.4.3.2). Liquid starter cultures using the conditions above grown overnight at 27^{o}C, induced at OD of 0.6 for 24 hours at 15^{o}C (see 3.4.3.3).

3.4.4.2 Inoculation and growth

One-hundred and fifty µl of the starter culture were inoculated into larger expression cultures LB broth supplemented with 30 µg ml-1 chloramphenicol and 100 µg ml-1 ampicillin and incubated at 37^{o}C with shaking until an OD of 0.6 was reached determined using a spectrophotometer. The starter culture allowed for a consistent number of *E. coli,* and in turn growth rate across flasks which could not be achieved by inoculation directly from the plate.

3.4.4.3 Induction

Cultures were then cooled in an ice bath, induced with 0.5mM IPTG and incubated for 24 hours at 15^{o}C for pnpla3 and 18^{o}C for PNPLA3. A low induction temperature was used to facilitate correct folding of the protein. A small sample was taken and incubated without the addition of IPTG as a negative control.

Cells were pelleted from the expression culture by centrifugation at 4^{o}C for 20 minutes at 12000g in order to halt the expression. The cell pellet was immediately frozen at -20^{o}C for storage.

3.4.4.4 Protein extraction

The cell pellet was allowed to thaw on ice and then re-suspended in cell lysis buffer (described below). The cells were lysed by sonication, with 10 cycles each comprising of 30 seconds sonication and 45 seconds cooling. The insoluble fraction containing cell membranes was separated by centrifugation at 4^{o}C for 20 minutes at 12000g and the cell lysate collected for further purification.

3.4.5 Purification

3.4.5.1 Buffer screens

The protein sample was dialysed into a range of buffers and analysed by further size exclusion chromatography to determine the effect on protein stability and select optimum conditions for purification (results not shown). Each buffer was modified by one single factor as follows:

NaCl 50mM 150mM 200mM 500mM
2mM Dithiothreitol (DTT), 200M Tris(2-carboxyethyl)phosphine hydrochloride (TCEP)
Tris(hydroxymethyl)aminomethane (Tris) 50mM, 100mM, 150mM
pH 6, 7, 8
3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS) 1mM and Polyethylene glycol sorbitan monolaurate (tween-20) 0.06mM

3.4.5.2 Final buffers:

Cell lysis buffer: 50mM Tris, 200mM NaCl, 10% glycerol, 200µM TCEP, 30mM imidazole, 100µg/ml lysozyme, ethylenediaminetetraacetic acid (EDTA) free protease inhibitor tablet (Roche), 500µM benzamidine pH 7.5

Ni-binding buffer: 50mM Tris, 200mM NaCl, 10% glycerol, 200µM TCEP, 30mM imidazole pH 7.5

Ni-elution buffer: 50mM Tris, 200mM NaCl, 10% glycerol, 200µM TCEP, 500mM imidazole pH 7.5.

SEC buffer: 50mM Tris, 200mM NaCl, 10% glycerol, 200µM TCEP, 30mM imidazole pH 7.5.

3.4.5.3 Nickel affinity

Nickel affinity chromatography was used for purification of the crude cell lysate, and to remove Ni binding contaminants after cleavage of the poly-histidine tag.

A range of buffers were used as a basis for the binding and elution buffers as described previously; all binding buffers contained 30mM imidazole and all elution buffers 500mM imidazole. Buffers were de-gassed before use, with a Büchner flask under a vacuum and filtered through a 0.22µm filter.

The column was equilibrated with two column volumes of binding buffer before loading the protein sample. The column was washed with binding buffer until no further protein eluted.

The bound protein was eluted with elution buffer. The flow through of the loaded sample, the wash elute, and the eluted protein peak were collected in fractions at pre-specified time intervals.

3.4.5.4 Size exclusion

Size exclusion chromatography was used as the second step of purification following nickel affinity chromatography.

Initial tests were undertaken, with small volumes of protein (100µl), on Superdex 200 10/300 GL and Superose 6 increase 10/300 GL columns using a range of buffers.

Molecular weight estimates were obtained by interpolating data points on an equilibration curve using Microsoft Excel, assuming a linear relationship.

3.4.5.5 Protein concentration and buffer exchange

The protein samples were concentrated using membrane ultrafiltration. Solutions smaller than 3ml were concentrated using a Vivaspin 500 while solutions over 3ml were concentrated using a Vivaspin 20 (GE Healthcare Life Sciences, Buckinghamshire, UK).

This protocol was also used for buffer exchange to reduce the concentration of imidazole for additional nickel-affinity columns. Adequate buffer exchange was performed to reach discrepancies of less than 0.3M reagent concentrations from the target buffer.

3.4.5.6 Cleavage of tag

Thrombin and Turbo 3C protease were tested for cleavage activity on the pnpla3 clone after the first nickel affinity chromatography column. The protein samples were incubated at 4^{o}C with either protease at a 1:200 dilution for 48 hours.

PNPLA3 clone A4 was cleaved using turbo 3C protease at 4^{o}C.at a 1:200 dilution for 48 hours.

3.4.6 Analysis

3.4.6.1 Plasmid concentration

The concentration of the plasmid DNA was estimated using spectrophotometric methods on a Nanodrop 1000 (Thermo Scientific Products). Two µl samples were loaded and the absorbance compared with pure buffer as a blank.

3.4.6.2 DNA sequencing

The resulting purified plasmid DNA was sequenced using the Source Bioscience Sequencing service (SourceBioscience, Nottingham, UK). The plasmid was sequenced in the forward and reverse direction using the standard plasmid recommended primers.

pOPIN primers:

T7-F 5'- TAATACGACTCACTATAGGG

T7-Terminal-R 5'- GCTAGTTATTGCTCAGCGG

pCold Primers:

pCold-TF-F2 5'-GCGAAAGTGACTGAAAAAG

pCold-R 5'-GGCAGGGATCTTAGATTCTG

3.4.6.3 Protein concentration

Protein concentration was estimated on a Nanodrop 1000 by measuring absorbance at 280mm. Two µl samples were loaded and compared with pure buffer as a baseline.

The OD values were used to more accurately account for protein concentration by adjusting for their respective predicted extinction co-efficient. This adjustment was not applied to the crude samples as the presence of multiple contaminants with different extinction coefficients would make the calculation unreliable.

Protein concentration was measured regularly to ensure no loss had occurred and to determine the yields from each protocol.

3.4.6.4 SDS-PAGE:

SDS-PAGE electrophoresis was used to determine the composition of the protein samples.

Soluble protein samples were diluted to a 1:1 ratio with Laemmli buffer. For cell membrane samples, 50µl of lysed cell debris was centrifuged at 12000g for 2 minutes; the supernatant was then discarded and cell debris diluted to 100 µl with Laemmli buffer.

All of the buffered samples were heated for 20 minutes at 97^{o}C prior to loading. Ten µl of each sample were loaded into each well and the gel was run at 25mA until the dye front reached the bottom of the gel. Either unstained molecular weight markers or full range rainbow markers were used for molecular weight comparisons.

Several gels, including: 12% home cast, 10% NuPage precast gels and 4-20% NuPage gradient gels were used in order to achieve optimal resolution of protein samples,. The run gels were initially stained with Coomassie blue for 1 hour, and de-stained overnight with de-stain solution.

If protein detection was not clear, the gels were destained for a further two days and re-stained using standard silver staining protocol (Appendix II).

3.4.6.5 Western blot:

Western blotting was performed to confirm the presence of the target protein using the poly-histidine tag as a marker.

Western blots were performed directly after SDS-PAGE and were transferred for 1 hour and 45 minutes at 100V. Blocking of the nitrocellulose membrane was performed in 5% milk in tris-buffered saline and tween-20 (TBST) solution for one hour with shaking at room temperature.

The membrane was then incubated overnight with horse radish peroxidase conjugated anti-his primary antibody (Abcam) at 4^{o}C in a 1:4000 dilution to 5% milk in TBST.

The membrane was washed three times for two minutes with TBST. The antibodies were then visualised using an AEC Chromogen Staining Kit (Sigma Aldrich, product number AEC101) and staining was stopped by washing with H_2O. Full-length rainbow markers were used in all western blots to allow for comparison against standard molecular weights.

3.4.6.6 Circular Dichroism:

Circular dichroism (CD) was used to assess protein folding. Both a pnpla3-trigger factor (TF) complex and a cleaved sample were exchanged into buffer containing less than 5mM NaCl. The samples were analysed with circular dichroism at the far UV-end of the light spectrum.

CD experiments were performed using a nitrogen- flushed Module B (b) end-station spectrophotometer at B23 Synchrotron Radiation CD Beamline at the Diamond Light Source, Oxfordshire, UK. The absorbance spectrum data was analysed using CDApps oftware[318].

3.4.6.7 Mass spectroscopy

Protein mass spectrometry was performed on the purified pnpla3 protein sample after size exclusion chromatography. The SDS-polyacrylamide gel was fully destained for 72 hours; the

band of interest was excised from the gel and prepared for mass spectrometry using standard in-gel digestion.[319]

Liquid chromatography-mass spectroscopy (LC-MS) was performed on the sample using a Velos Orbitrap mass spectrometer. The proteins present in each sample were determined using Proteome Discoverer 1.3 via searches of the UniProt human and *E. coli* database with the Mascot search engine.

3.4.7 Activity assay

The substrate 1,2-O-dilauryl-rac-glycero-3-glutaric acid-(6'- methylresorufin)-ester (DGGR) can be hydrolysed by lipases to 1,2-O-dilauryl-rac-glycerol, and an unstable dicarbonic acid ester intermediate, glutaric acid-(6'-methylresorufin)-ester. The glutaric acid-(6'-methylresorufin)-ester spontaneously hydrolyses to form glutaric acid and methylresorufin, a bluish-purple chromophore with peak absorption at ~580nm (Figure 3.17).[320]

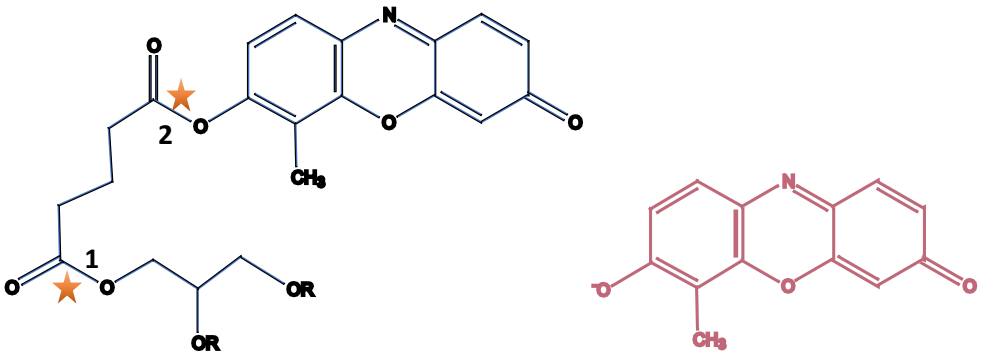


Figure 3.17 The hydrolysis reaction of DGGR to methylresorufin.

DGGR is shown in blue and methylresorufin in pink. R groups represent $-(CH_2)_{11}CH_3$. Bond 1 is hydrolysed in the reaction by lipase class enzymes while Bond 2 spontaneously hydrolyses as a second step from the unstable intermediate product.

Lipase activity was measured for partially purified pnpla3 and PNPLA3 clone A4. The increase in absorbance at 580nm was used to represent the lipase activity of the protein sample.

The assay buffer consisted of 50mM Tris-HCl, 500mM NaCl, 10% glycerol, 1mM tris-(2-carboxyethyl)-phosphine (TCEP) at pH 7.5. DGGR was dissolved in the minimum quantity of methanol to achieve a solution and then rapidly added to the assay buffer to form a micro-suspension. A fresh DGGR sample was prepared for each assay.

Protein samples, DGGR and the 96-well reaction plate were separately incubated at 37^{o}C for five minutes prior to reaction. Reactions were initiated by the addition of protein to DGGR solutions in a 1:1 ratio, to a final volume of 250µl in the 96 well plate. The final concentrations of protein and DGGR were 0.5mg/ml and 36µg/ml respectively.

The reaction was monitored in a SpectraMaxM2 spectrophotometer with readings at 20 second intervals for two hours with automixing between readings. The DGGR buffer mixture without protein was used as a blank.

The first 60 seconds of collected data were discarded to allow for the initial delay in reaction initiation.; The reaction rates were calculated over the subsequent 10 minutes using linear regression in Microsoft Excel.

3.4.8 Crystal screening

Crystal screening was performed using a mosquito Crystal liquid handler for partially purified clones PNPLA3-A4 and pnpla3 after size exclusion chromatography.

Screens were performed using a sitting drop method with 400nl drops, consisting of 200nl of protein and 200nl of well solution. Structure screen 1&2, JCSG-plus, MIDAS, Morpheus II and PACT-Premier screening kits (Molecular Dimensions, Suffolk, UK) were all tested at a protein concentration of 14mg/ml and 10mg/ml at room temperature.

The screens were manually checked for crystals with an optical stereo microscope after 24 hours, 36 hours, 1 week and one month.

3.5 Results

3.5.1 Transformation into *E. coli* vectors.

All clones successfully transformed into the DH5α *E Coli* cloning strain and produced adequate concentrations of DNA for sequencing. Sequencing confirmed all sequences were as expected (Data not shown).

Transformation into each *E. coli* expression strain was also successful with a large number of distinct colonies growing under antibiotic selection and with each plate presenting normal cell morphology (Figure 3.18).

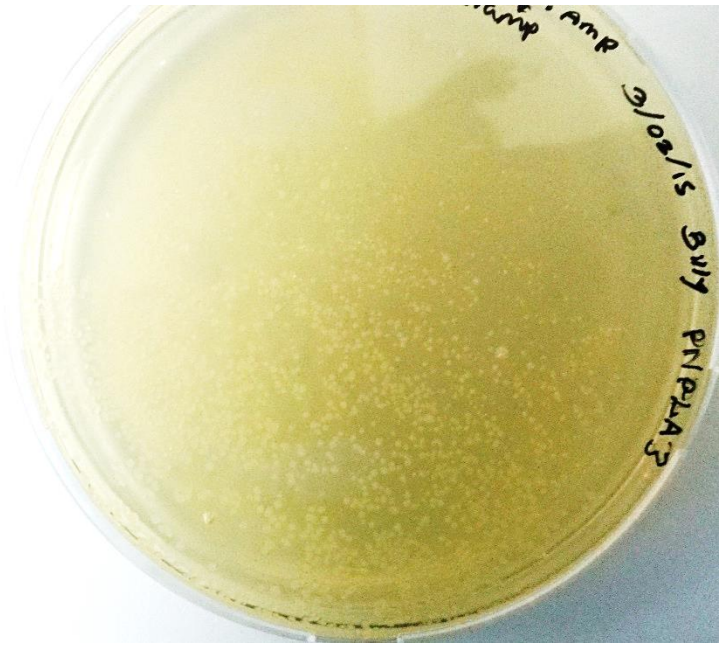


Figure 3.18 Example of positively transformed colonies grown overnight at 37°C

3.5.2 Pnpla3 protein expression

3.5.2.1 Pnpla3 expression trials

An additional band with an estimated molecular weight of around 98kDa was observed in colonies grown overnight at 37°C and induced to an OD of 1.4 at 15°C (Figure 3.18: Lanes 1 and 3). This band was only weakly observed or absent in the uninduced samples grown under the same conditions (Figure 3.18: Lanes 2 and 4).

The most prominent and clearest expression band was observed in colonies grown overnight at 27°C, and induced to an OD of 0.6 for 24 hours at 15°C (Figure 3.19; Lane 7). These are the conditions recommended for expression by Takura Bioscience.[316] However, the estimated molecular weight of this band is 45kDa and hence much lower than the expected 95kDa expected for pnpla3-tf recombinant protein, more likely corresponding to trigger factor alone.

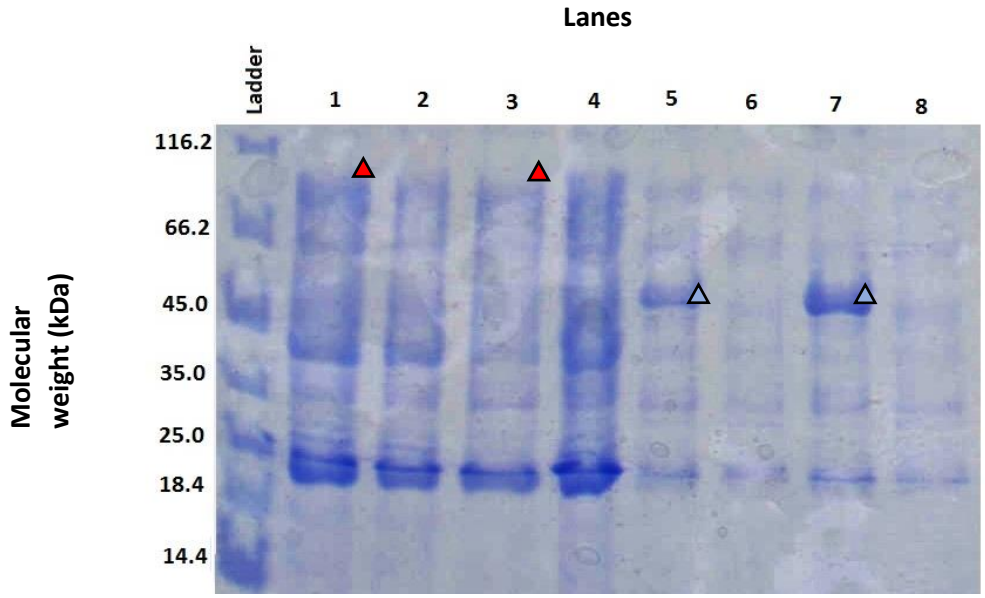


Figure 3.19 SDS-PAGE analysis of the pnpla3 small scale expression trials.

Red triangles indicate pnpla3-TF expression bands at approximately 98kDa
Blue triangle indicates degraded TF only expression band at 45kDa
Lane:
1: Grown overnight at 37°C, induced at OD 1.4 for 16 hours at 15°C
2: Un-induced grown overnight at 37°C, incubated for 16 hours at 15°C
3: Grown overnight at 37°C, induced at OD 1.4 for 24 hours at 15°C
4: Un-induced grown overnight at 37°C, incubated for 24 hours at 15°C
5: Grown overnight at 27°C, induced at OD of 0.6 for 16 hours at 15°C
6: Un-induced grown overnight at 27°C, incubated for 16 hours at 15°C
7: Grown overnight at 27°C, induced at OD of 0.6 for 24 hours at 15°C
8: Un-induced grown overnight at 27°C, incubated for 24 hours at 15°C

3.5.2.2 PNPLA3 expression trials

The human PNPLA3 clones showed only low levels of expression. Additional bands with an estimated expected molecular weight were only detected in clones A4 and D2 (Figure 3.20; Lanes 2 and 7). Clone A4 was selected for further investigation since this corresponded to a full-length clone.

Changing expression conditions had minimal impact on the expression of PNPLA3, but the best expression conditions were growing overnight at 27°C and inducing at an OD of 0.6 for 24 hours at 15°C (Figure 3.21; Lane 4).

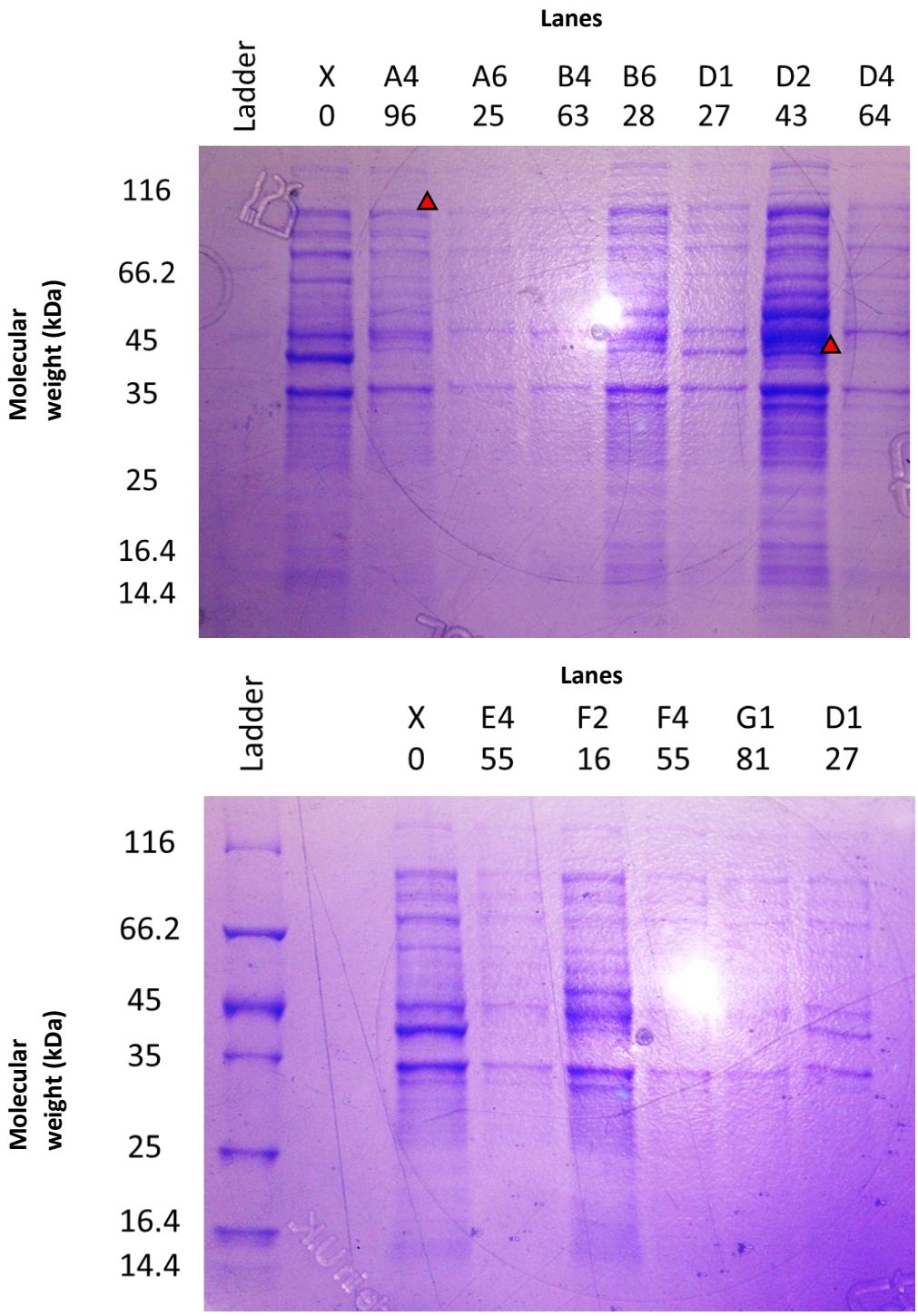


Figure 3.20: SDS-PAGE analysis of PNPLA3 small scale expression trials

Each lane corresponds to a different clone labelled above with ID and its expected molecular weight.
Red triangles indicate PNPLA3 expression bands on the gel.
Lane X is an un-induced control sample.

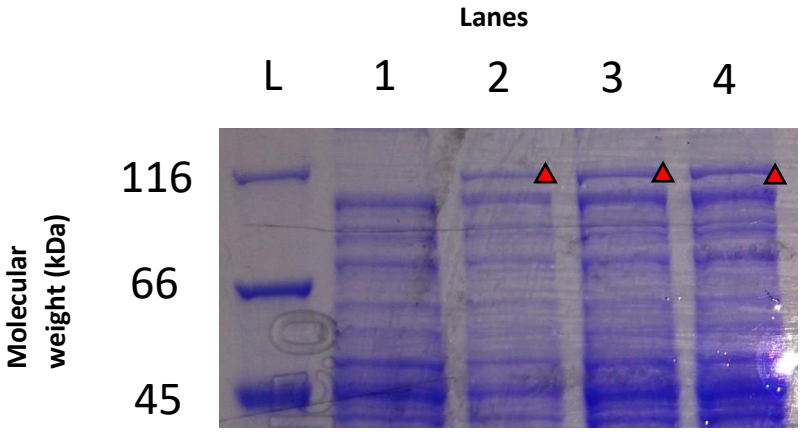


Figure 3.21 SDS-PAGE analysis of PNPLA3 clone A4 small scale expression trials.

Red triangles used to indicate PNPLA3 expression bands on the gel; **L:** Molecular marker
Growth conditions
1: Un-induced grown overnight at 37°C, incubated for 16 hours at 15°C
2: Grown overnight at 37°C, induced at OD 1.4 for 16 hours at 15°C
3: Grown overnight at 37°C, induced at OD 1.4 for 24 hours at 15°C
4: Grown overnight at 27°C, induced at OD of 0.6 for 24 hours at 15°C

3.5.3 Purification

3.5.3.1 Pnpla3 nickel affinity chromatography

Purification with a nickel affinity chromatography column was successfully able to remove the majority of contaminants from the cell lysate (Figure 3.22). Elution with high concentration imidazole produced a significant amount of protein (Figure 3.23), with each litre of culture yielding approximately 10ml of eluted protein at a concentration of 2mg/ml.

Further analysis of these fractions using SDS-PAGE and western blot confirmed the presence of pnpla3. The main body of protein was present at a molecular weight of around 98kda. However, the band was a diffuse combination of protein suggesting high levels of degradation. Additional high weight bands were also detected suggesting the presence of molecular multimers in solution.

The obvious high level of contamination within the sample indicated the need for further purification steps.

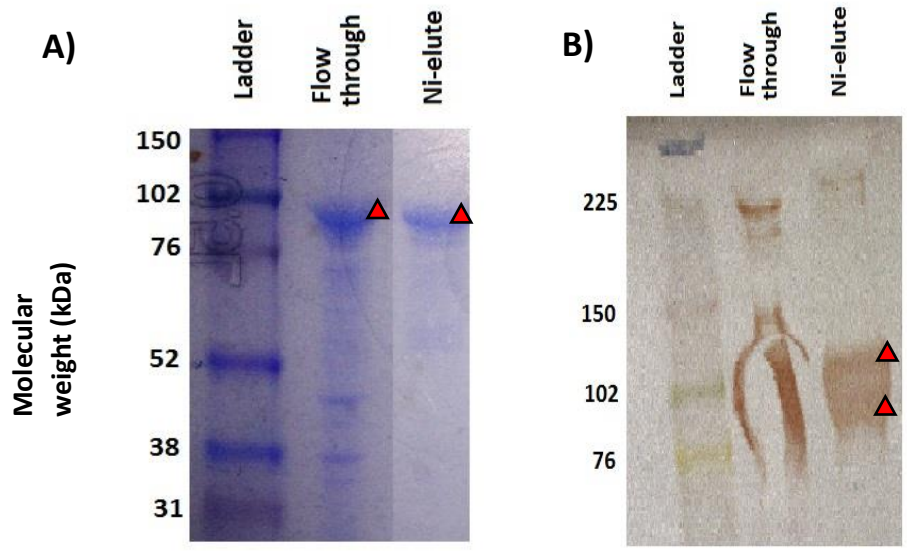


Figure 3.22 SDS-PAGE and western blot analysis of nickel purified pnpla3.

Red triangles used to indicate pnpla3 expression bands.
A) SDS-PAGE of the unbound and bound protein samples.
B) Western blot using a his-tagged antibody.

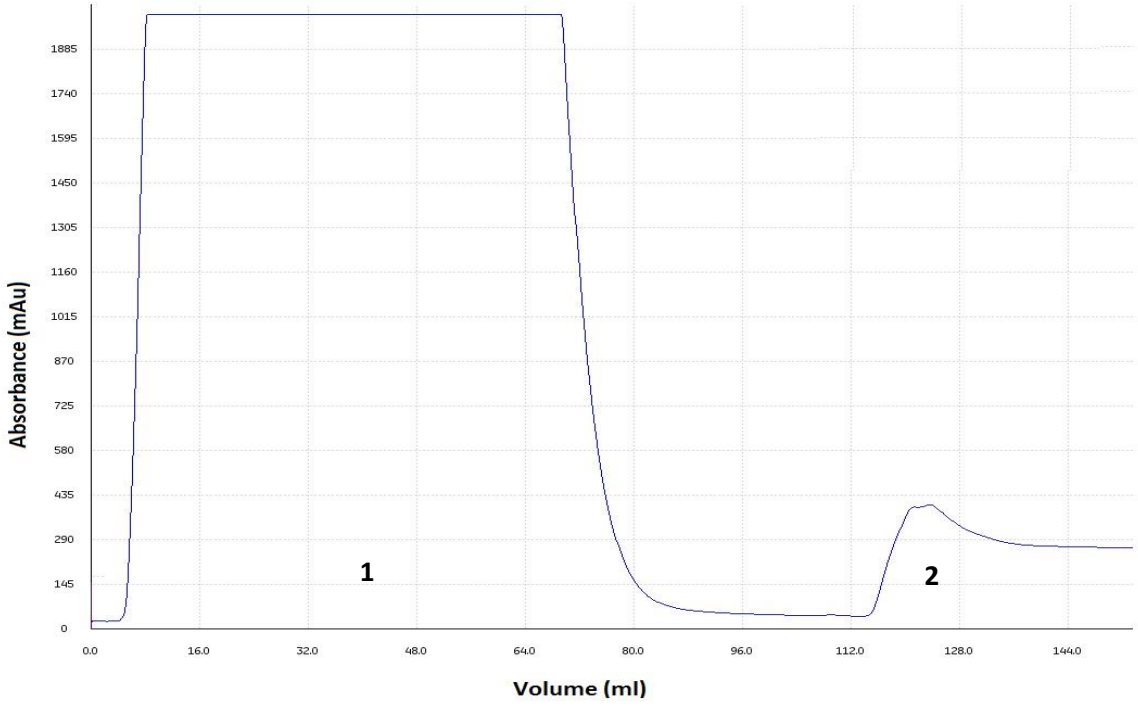


Figure 3.23 Absorbance trace of Nickel affinity purification on cell lysates

Peak 1 shows loading of cell lysate;
peak 2 shows elution of bound proteins using imidazole.

Lanes **Lanes**

3.5.3.2 PNPLA3 nickel affinity chromatography

Purification with a nickel affinity chromatography column removed the majority of contaminants from the cell lysate (Figure 3.24). Each litre of culture yielding approximately 10ml of eluted protein at a concentration of 1.3mg/ml.

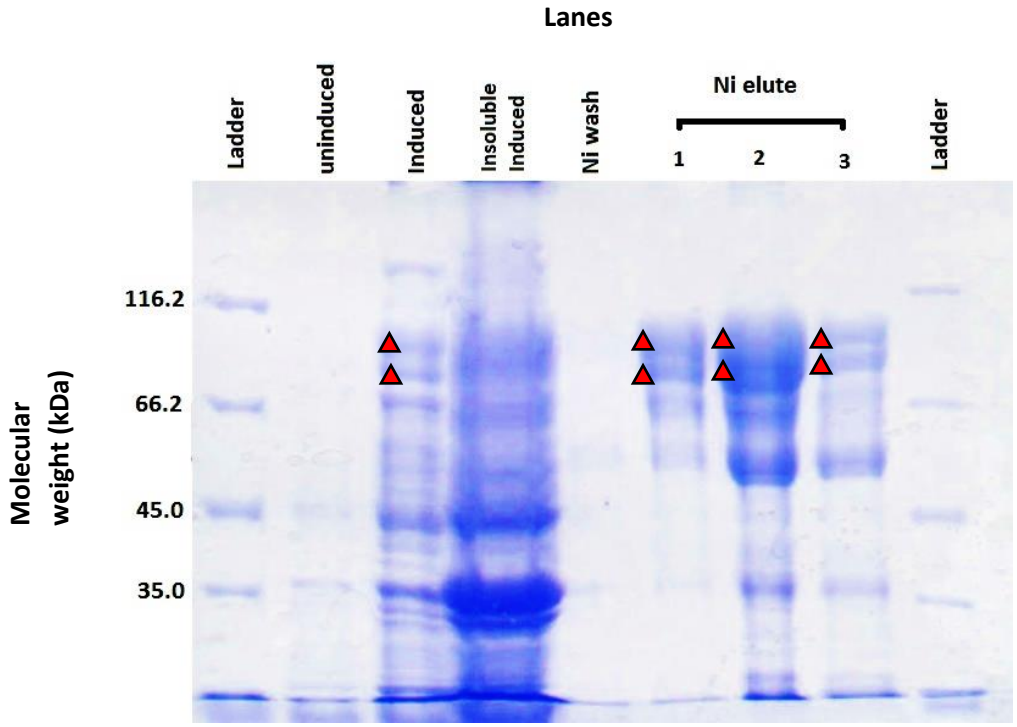


Figure 3.24 SDS-PAGE analysis of nickel purified PNPLA3.

Red triangles used to indicate PNPLA3 expression bands on the gel.

Further analysis of these fractions using SDS-PAGE and western blot confirmed the presence of PNPLA3. The main body of protein was present as two predominant bands, at molecular weights of around 96kDa and 76kDa. These bands were broad and diffuse suggesting high levels of degradation (Figure 3.25). The level of contamination within the sample showed the need for further purification steps.

3.5.3.3 Pnpla3 size exclusion chromatography

Calibration curves were generated for the Superdex 200 and Superose 6 size exclusion chromatography columns to assess the sizes of the eluted fractions (Figure 3.26).

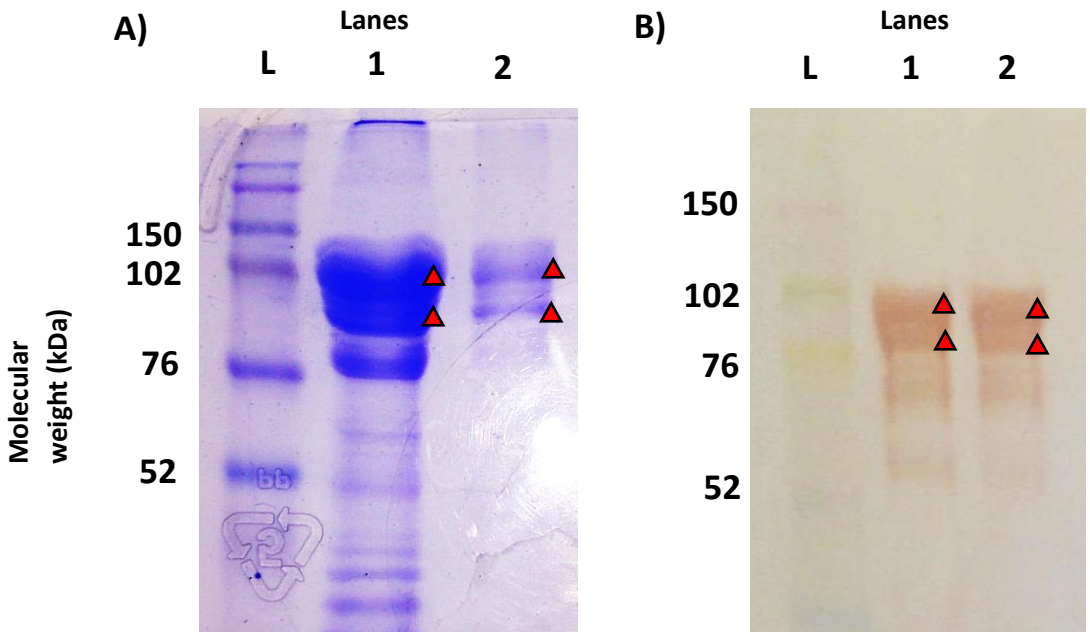


Figure 3.25 SDS-PAGE and western blot analysis of nickel purified PNPLA3.

Both lanes contain the same sample at different concentrations
Red triangles indicate PNPLA3 expression bands.
A) SDS-PAGE **B)** Western blot using an anti-his monoclonal antibody.

The molecular weights for fragments of the pnpla3-TF recombinant enzyme should vary between 45 and 100 kDa; however, multimers with higher molecular weights might exist (Tables 3.3 and 3.4).

Purification of the nickel elute from the previous step using size exclusion chromatography showed that the largest component of the solution had an apparent molecular weight of around 700kda. Further analysis of this peak using a Superose 6 column revealed a broad symmetric peak of around 670kda as the main component of the solution (Figure 3.27). There were also a range of other potential contaminant peaks within the sample (Figure 3.8).

Further analysis of each eluted peak using SDS-PAGE and western blotting revealed that only this large 670kDa multimer contained pnpla3-TF. The purity of the protein sample was improved overall, but still contained large amount of degradation product (Figure 3.28).

Changing NaCl concentration, Tris concentration and pH had no effect on the elution profile. (chromatography traces not shown). Addition of CHAPS and Tween-20 resulted in sharpening of the 700kda peak and an apparent reduction in other peaks but no improvement in recovered protein.

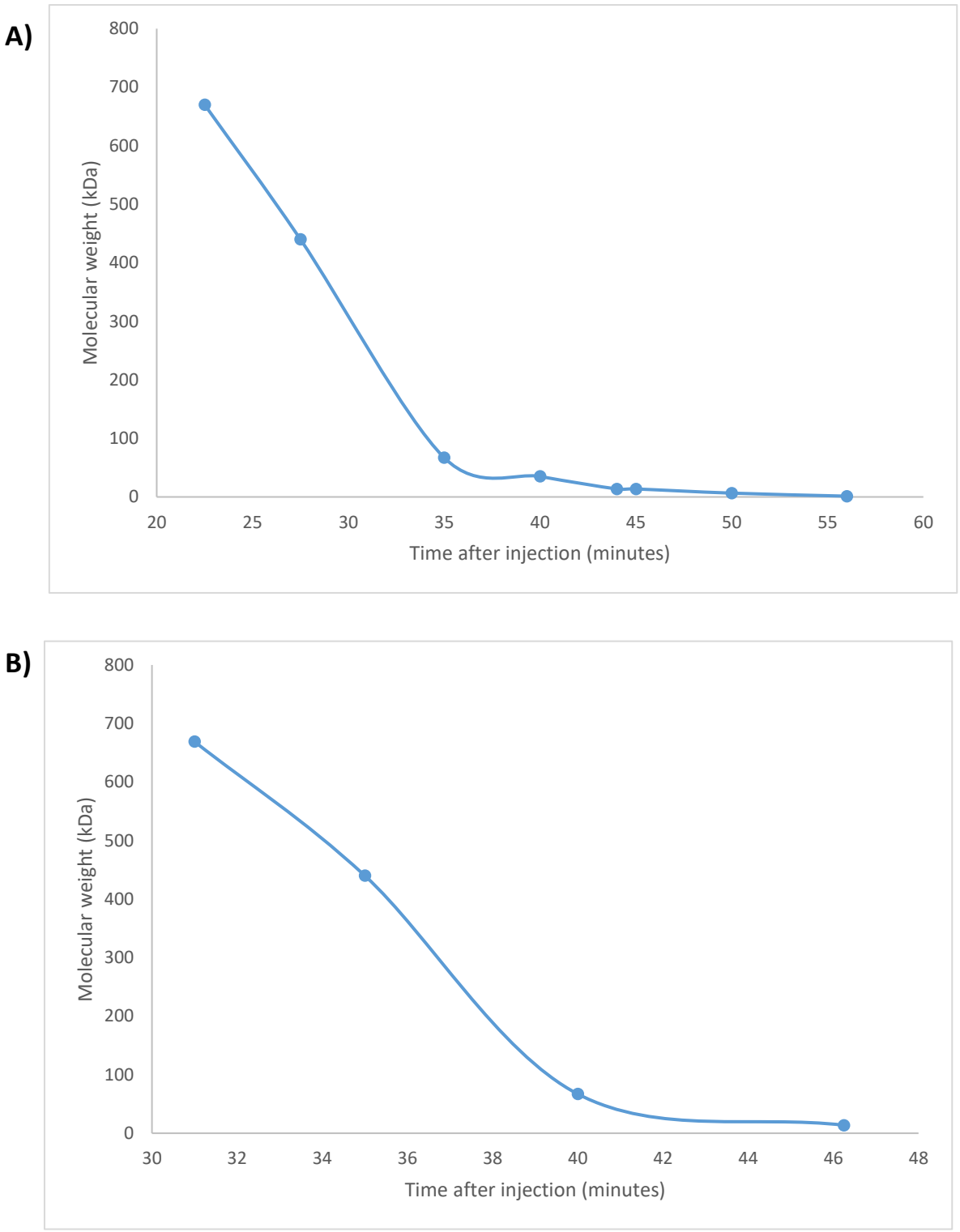


Figure 3.26 The calibration curve for size exclusion chromatography columns.

A) Superdex 200 column.
B) Superose 6 column.

Table 3.3 Predicted elution time from Superdex 200 column for key protein elements of recombinant pnpla3 based on the equilibration curve

Element	Molecular weight (kDa)	Expected time of elution (mins)
octamer	758.1	20.6
heptamer	663.3	22.6
hexamer	568.6	24.7
pentamer	473.8	26.8
tetramer	379.1	28.7
trimer	284.3	30.6
dimer	189.5	32.5
Pnpla3 -TF	94.8	34.4
TF	49.0	37.8
Pnpla3	45.8	38.3

TF: Trigger factor

Table 3.4 Predicted elution time from Superose 6 column for key protein elements of recombinant pnpla3 based on the equilibration curve

Element	Molecular weight (kDa)	Expected time of elution (mins)
Octamer	758.1	29.4
Heptamer	663.3	31.1
Hexamer	568.6	32.8
Pentamer	473.8	34.4
Tetramer	379.1	35.8
Trimer	284.3	37.1
Dimer	189.5	38.4
PNPLA3-TF	94.8	39.6
TF	49.0	42.1
PNPLA3	45.8	42.5

TF: Trigger factor

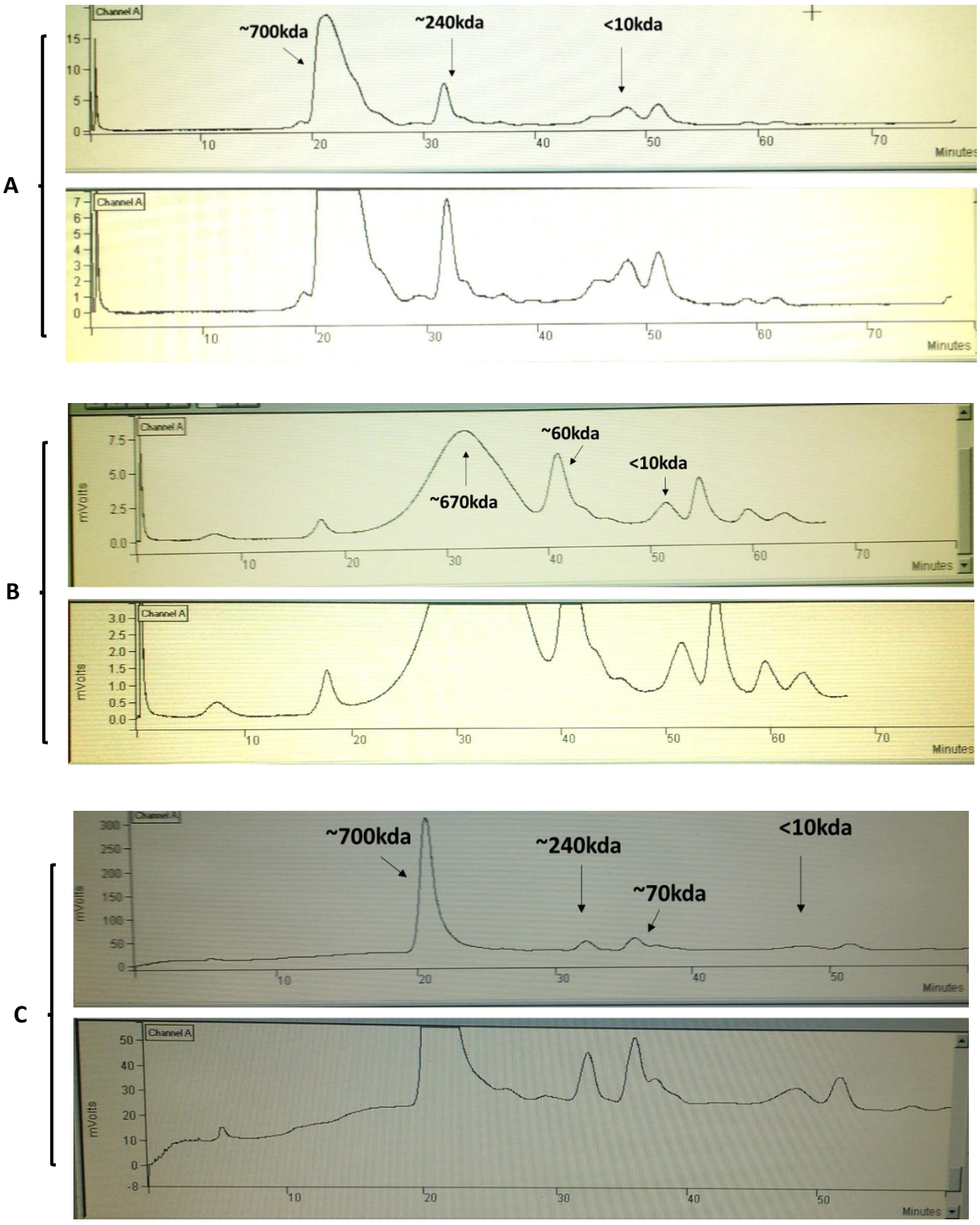


Figure 3.27 Size exclusion chromatography of pnpla3-TF

The top panel shows a trace of absorbance against time after injection on to the size exclusion column Predicted molecular weights for each large peak indicated in bold. The bottom panel is enlarged to more clearly show impurities within the sample

A) pnpla3-TF complex using a Superdex 200 column

B) pnpla3-TF complex using a Superose 6 column

C) detergent stabilised pnpla3-TF complex on Superdex 200 column

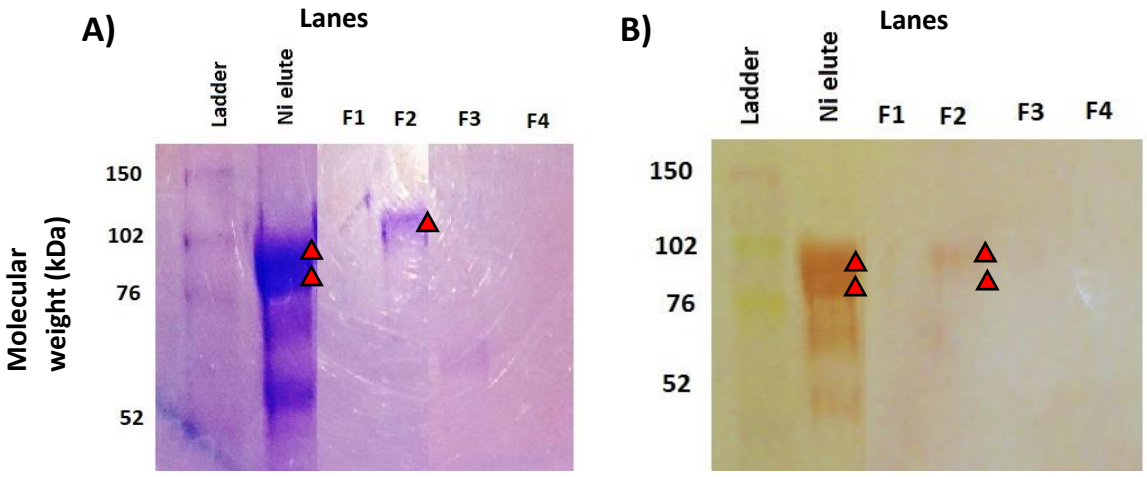


Figure 3.28 SDS-PAGE and corresponding western blot of size exclusion chromatography separated peaks

A) SDS-PAGE of the unbound and bound protein samples

B) corresponding western blot using a his-tagged antibody.

Red triangles used to indicate PNPLA3 expression bands on the gel. The F1 fraction contained the protein eluted in void volume, F2 large multimeric peak, F3 eluted at 32 minutes and F4 at 36 minutes.

3.5.3.4 PNPLA3 size exclusion chromatography

Calibration curves were generated for the Superdex 200 and Superose 6 size exclusion chromatography columns to assess the sizes of the eluted fractions.

The molecular weights for fragments of the PNPLA3-MBP recombinant enzyme should vary between 45 and 100 kDa; however, multimers with higher molecular weights might exist (Tables 3.5 and 3.6).

Purification of the nickel elute with size exclusion chromatography showed that the largest component of the solution has an apparent molecular weight of around 100kda. There are also a range of potential other contaminant peaks within the sample including a peak corresponding to the large 670kDa multimer observed during purification of pnpla3 (Figure 3.29).

Further analysis of each eluted peak revealed only the large 670kDa peak was observed on western blots and therefore contained PNPLA3-MBP. The purity of the protein sample in this peak was improved overall but it still contained large amounts of degradation product.

Additional size exclusion chromatography of this 670kDa peak, showed that smaller contaminants had been removed and that there was no dynamic equilibria with smaller peaks. Changing NaCl concentration, Tris concentration or pH had no effect on the elution profiles (chromatography traces not shown).

Table 3.5 Predicted elution time from Superdex 200 column for key protein elements of recombinant PNPLA3 based on the equilibration curve

Element	Molecular weight (kDa)	Expected time of elution (mins)
octamer	770.4	20.3
heptamer	674.1	22.4
hexamer	577.8	24.5
pentamer	481.5	26.6
tetramer	385.2	28.6
trimer	288.9	30.5
dimer	192.6	32.5
PNPLA3-MBP	96.3	34.4
MBP	43.5	38.7
PNPLA3	52.8	37.2

MBP: Maltose binding protein

Table 3.6 Predicted elution time from Superose 6 column for key protein elements of recombinant PNPLA3 based on the equilibration curve

Element	Molecular weight (kDa)	Expected time of elution (mins)
octamer	770.4	29.2
heptamer	674.1	30.9
hexamer	577.8	32.6
pentamer	481.5	34.3
tetramer	385.2	35.7
trimer	288.9	37.0
dimer	192.6	38.3
PNPLA3-MBP	96.3	39.6
MBP	43.5	42.8
PNPLA3	52.8	41.7

MBP: Maltose binding protein

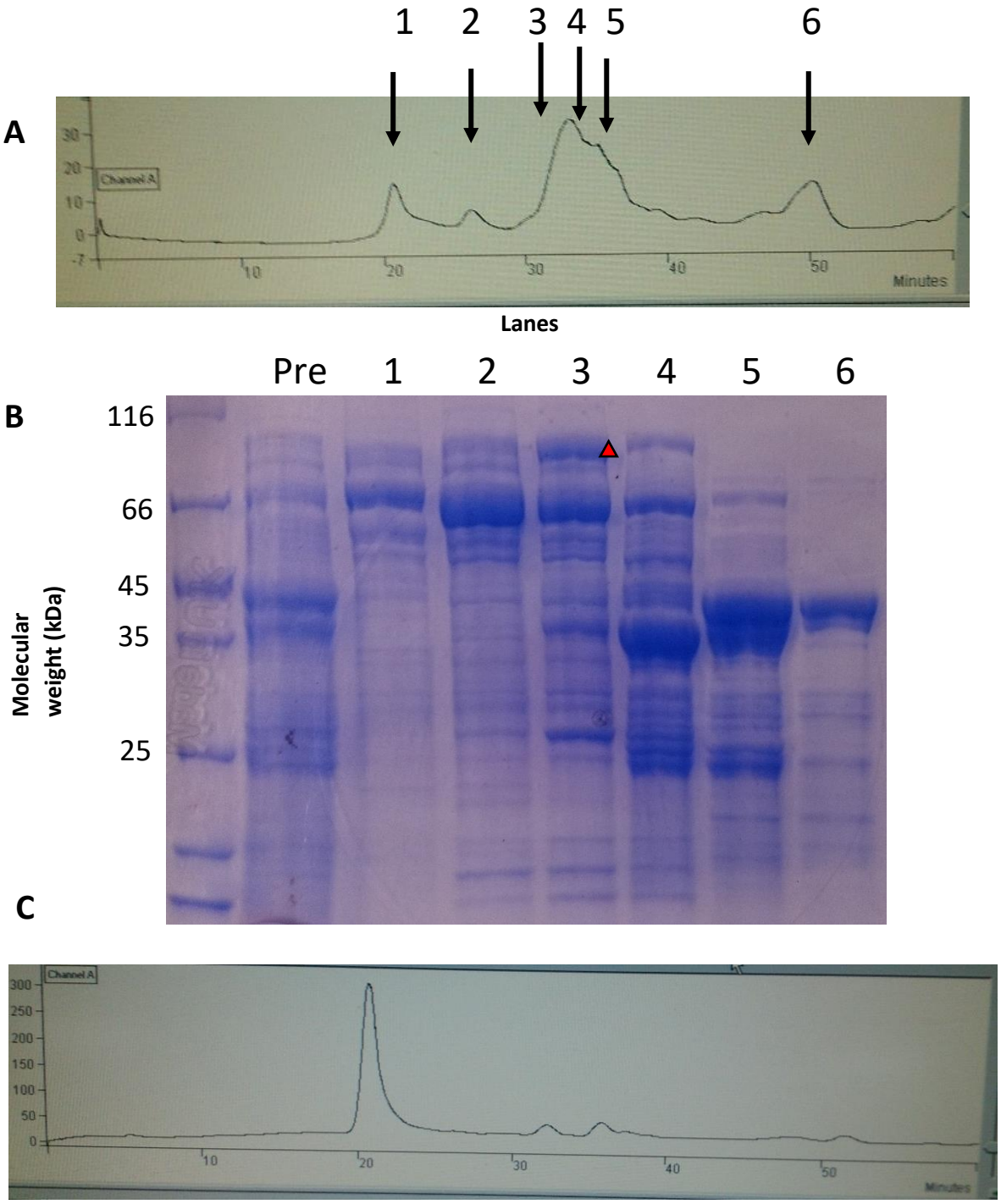


Figure 3.29 Size exclusion chromatography of PNPLA3-MBP

A) Trace of absorbance against time after injection on to the Superdex 200 size exclusion column
B) SDS-PAGE analysis of each fraction eluted from the SEC column
C) Rerun of the first peak eluted from the previous SEC run. trace of absorbance against time after injection on to the Superdex 200 size exclusion column.

3.5.3.5 Pnpla3 proteolytic cleavage of TF tag

Cleavage of the pnpla3-TF sample with thrombin was effective at very low concentrations. However, while a large band corresponding to TF was seen on SDS-PAGE very little pnpla3 was detected. A much higher level of protein cleavage was obtained with 3C protease with maintenance of the pnpla3 to TF ratio (Figure 3.30). However, neither thrombin nor 3C protease were able to cleave the TF tag in buffers containing detergent (results not shown).

A significant amount of cleavage occurred after incubation with either enzyme. However, the elution profile from size exclusion chromatography was unchanged (Figure 3.31).

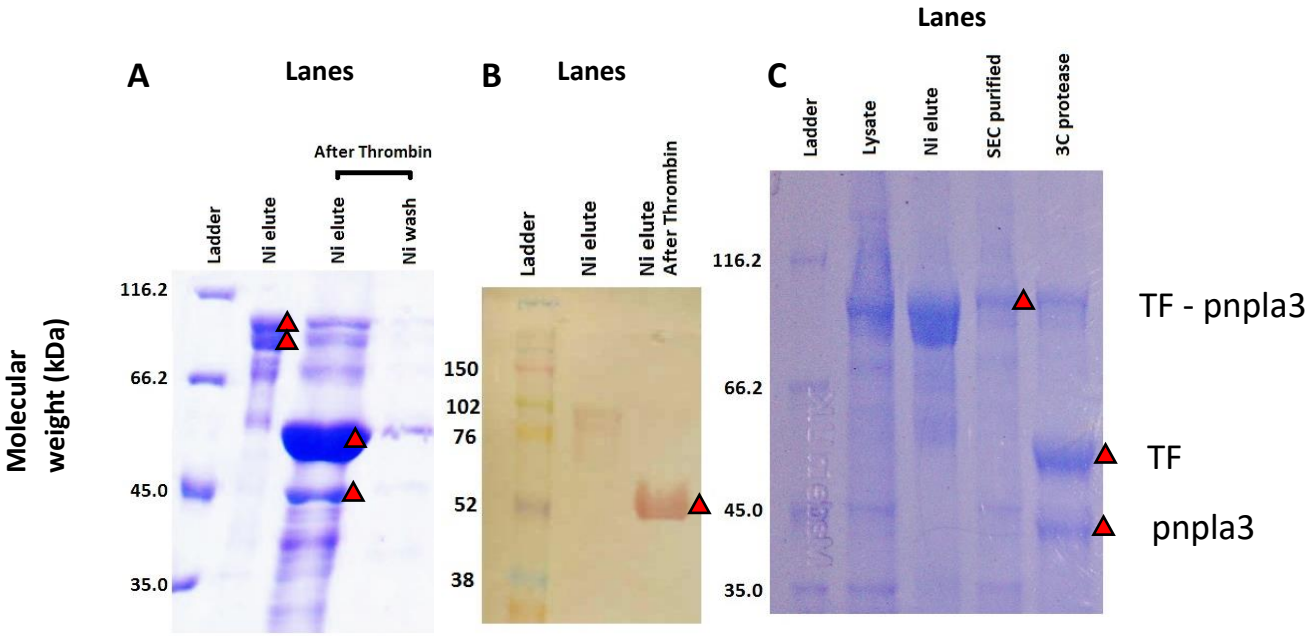


Figure 3.30 SDS-PAGE and corresponding western blot analysis of cleaved pnpla3-TF recombinant protein

Red triangles indicate key expression bands of interest
A) SDS-PAGE of pnpla3-TF after cleavage with thrombin
B) The corresponding western blot with his-tag antibody
C) SDS-PAGE of cleavage with 3C protease.

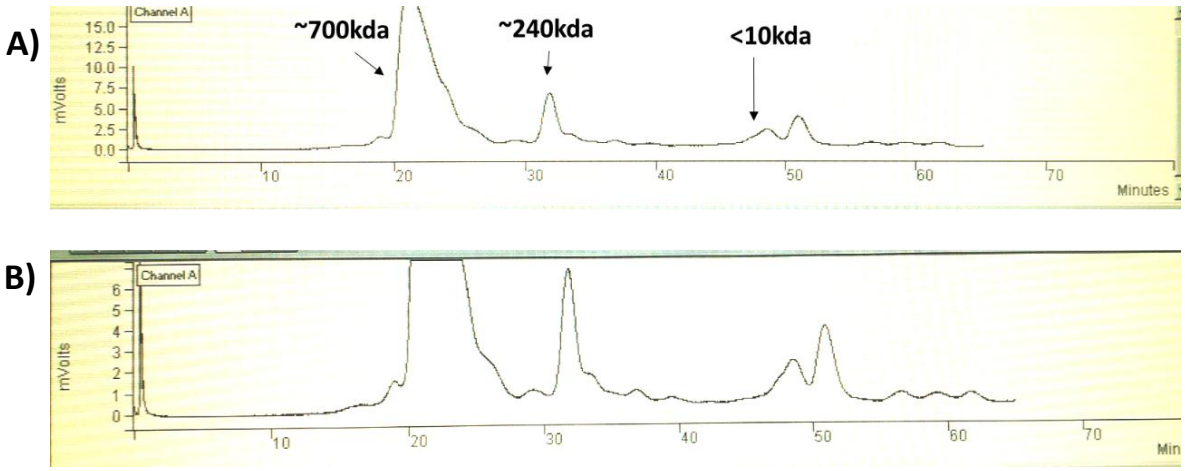


Figure 3.31 Size exclusion chromatography of cleaved pnpla3 and TF solution on a Superdex 200 column

A) Trace of absorbance against time after injection on to the size exclusion column; the predicted molecular weights for each large peak indicated in bold
B) Enlarged section of the trace to more clearly show impurities within the sample

3.5.3.6 PNPLA3 proteolytic cleavage of MBP tag

Cleavage of the PNPLA3-MBP sample with 3C-protease was effective at low concentrations (Figure 3.32). The levels of MBP far exceeded those of PNPLA3 and the solution was highly unstable. The elution profile from size exclusion chromatography was also unchanged, as with pnpla3 (results not shown).

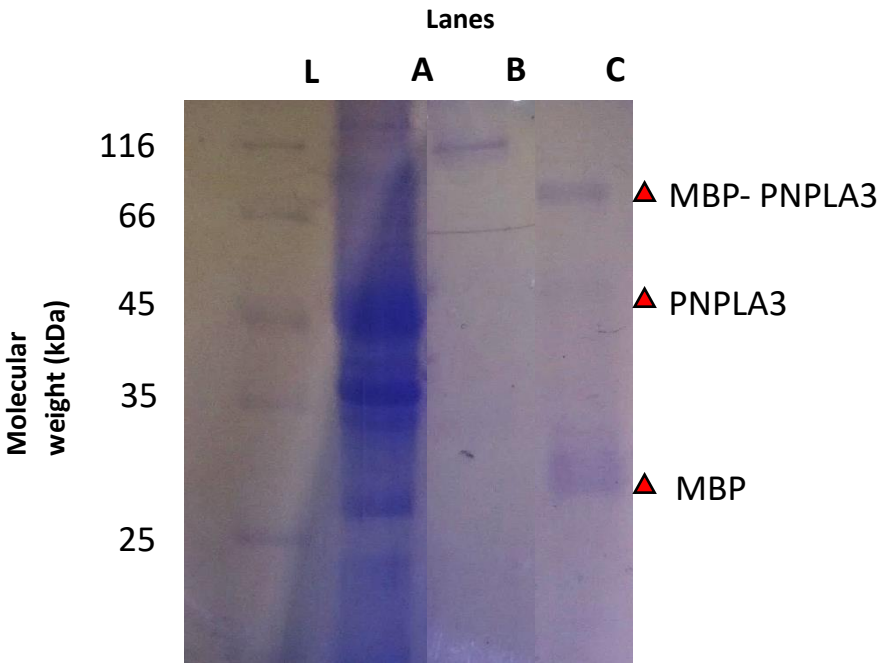


Figure 3.32 SDS-PAGE analysis of cleaved PNPLA3-MBP recombinant protein

L) Molecular marker **A)** Cell lysate; **B)** Purified protein sample; **C)** After cleavage with 3C protease.

3.5.6 Lipase activity assays

The initial results of the DGGR lipase activity assay show that crudely purified pnpla3 from the size exclusion chromatography is active and that increasing DGGR concentration increases the rate of reaction in a seemingly linear fashion (Figure 3.33). When compared, crudely purified PNPLA3 exhibited nearly three times the activity of pnpla3 (Figure 3.34).

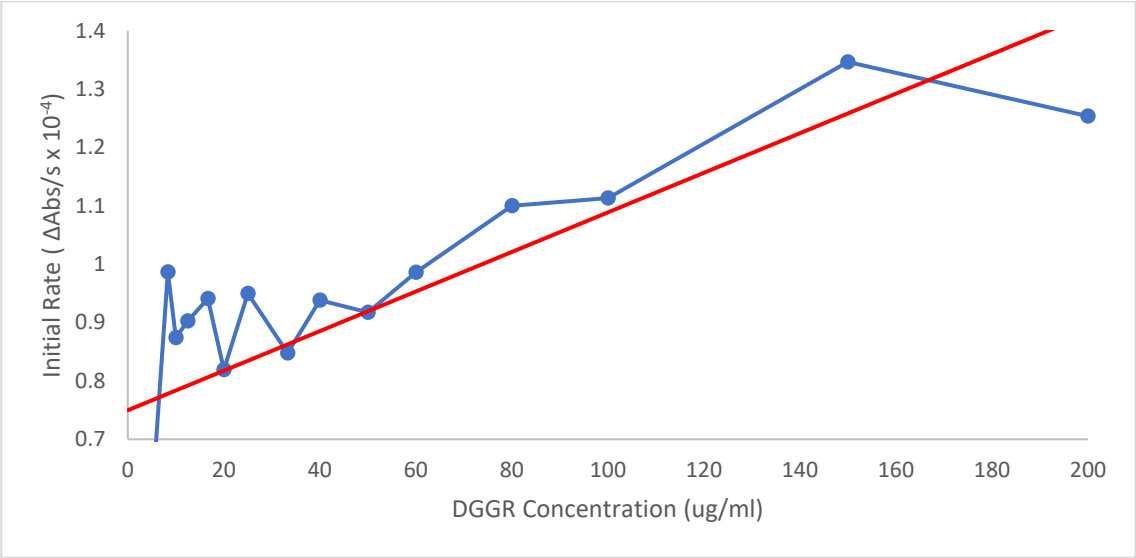


Figure 3.33 Lipase activity of pnpla3

Blue: line between actual absorbance data points. Red: line representing linear regression

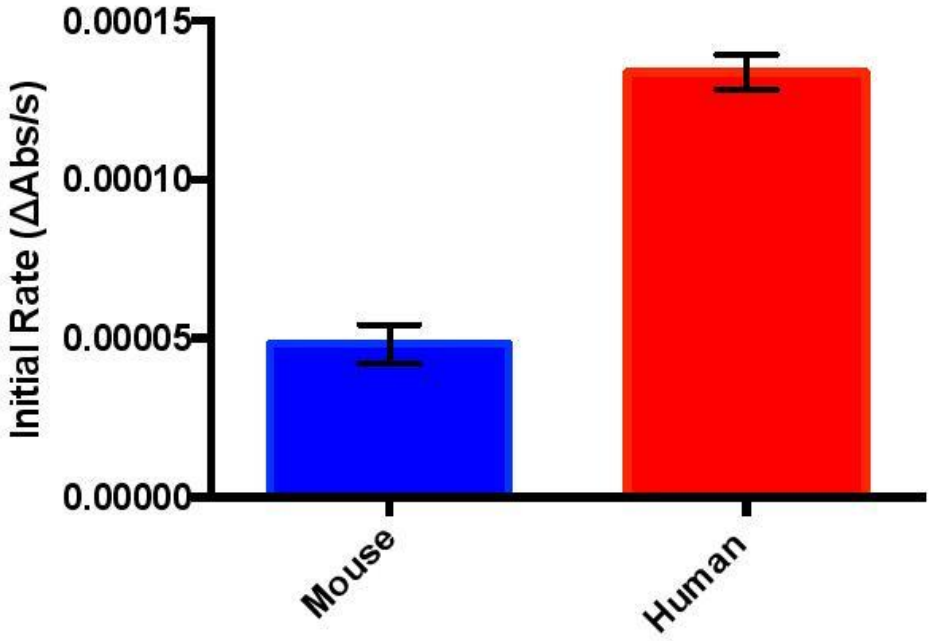


Figure 3.34 Comparative lipases activities of pnpla3 and PNPLA3

The bar graph represents the mean, with upper and lower 95% confidence intervals shown as error bars.

3.5.7 Crystallization screening

None of the conditions screened with the Mosquito generated crystals. Approximately half of the conditions resulted in protein precipitation, some of which had a white crystalline appearance which appears to be more ordered in structure (Figure 3.35).

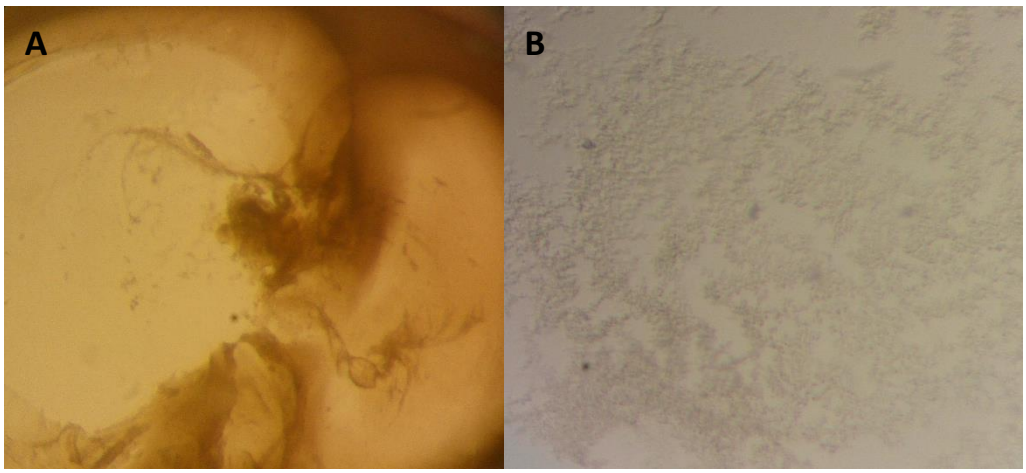


Figure 3.35 Photograph of example crystal screening results
A) Precipitated protein sample.
B) White ordered precipitate.

3.5.8 Pnpla3 ammonium sulfate precipitation

After ammonium sulfate precipitation and subsequent re-dissolving into 8M urea, 6M guanidine-HCl and 2% SDS, there were 50%, 60% and 40% respective yields of recovered protein.

Size exclusion chromatography showed a disruption of the multimeric state, and the majority of the protein had an apparent molecular weight of below 10kDa. Dialysis into folding buffer of the precipitated sample led to an alteration of the size exclusion chromatography trace but not a recovery of the high weight multimeric state (Figure 3.36).

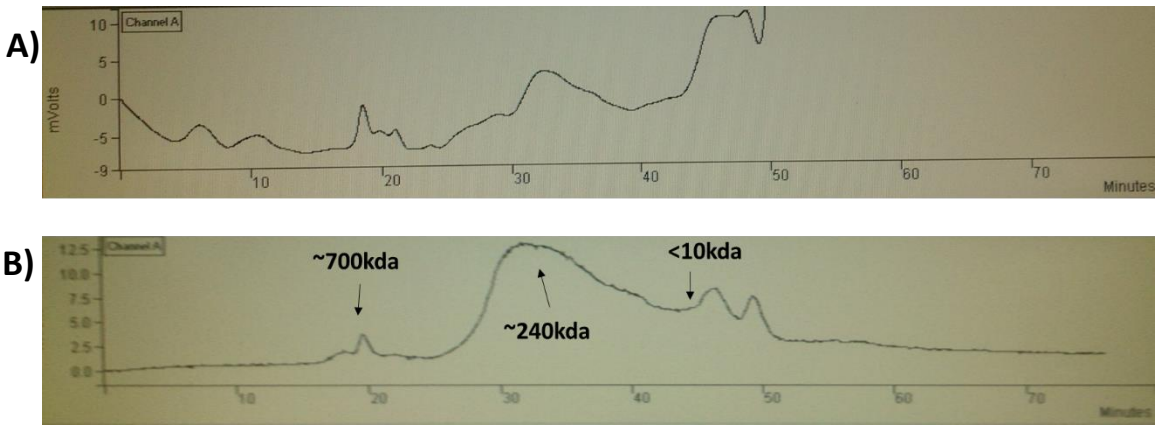


Figure 3.36 Size exclusion chromatography of pnpla3-TF

The traces show absorbance against time after injection on to the size exclusion column
A) Ammonium sulfate precipitated sample re-dissolved in urea
B) Re-dissolved sample after dialysis into refolding buffer.
Predicted molecular weights for each large peak indicated in bold.

3.5.9 Pnpla3 circular dichroism

The pnpla3-TF complex showed peak absorbance around 195nm. The buffer also showed some absorbance around this region. Predictions of the secondary structure indicate that 40% of the protein is unordered, and an equal proportion is in a beta sheet formation. Approximately 17% is predicted to be turns with very little α- helices (Figure 3.37).

The cleaved pnpla3 TF mixture showed the exact same result (Figure 3.38).

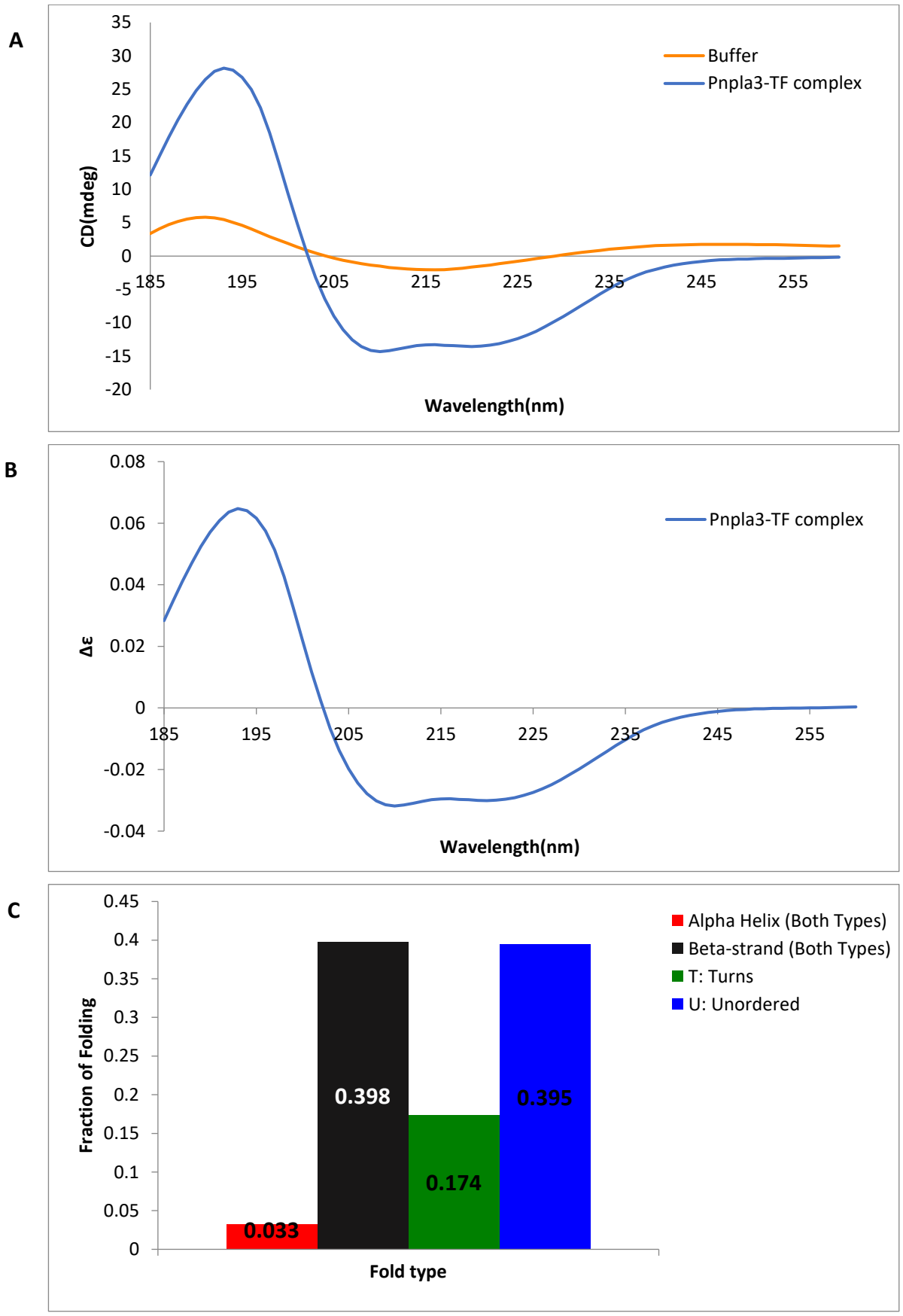


Figure 3.37 Circular dichroism of pnpla3-TF

A) Circular dichroism absorbance spectrum of pnpla3-TF sample and buffer
B) Extinction coefficient spectrum of pnpla3-TF complex corrected for buffer absorption.
C) CDApps predicted secondary structure pnpla3-TF complex.

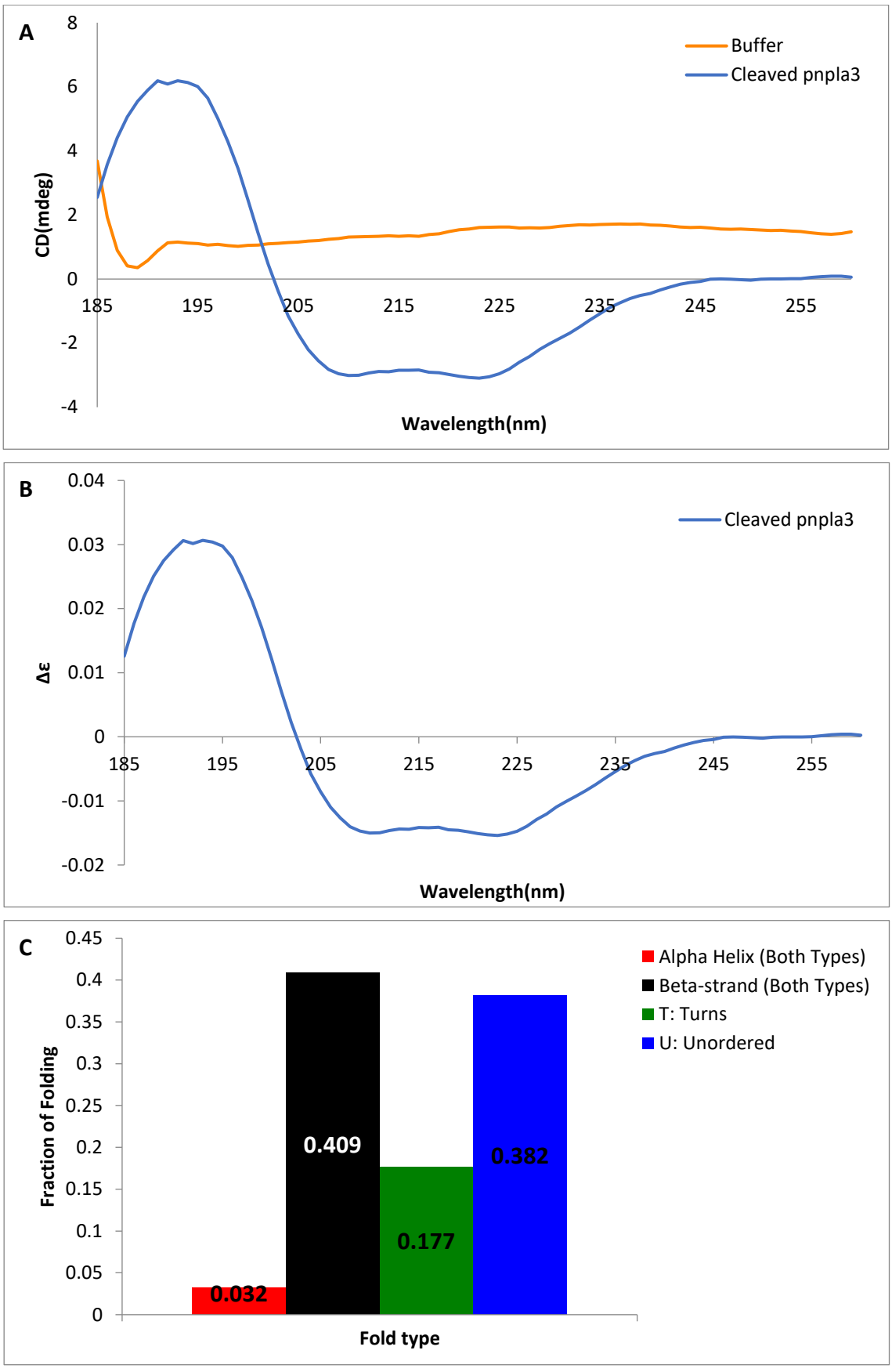


Figure 3.38 Circular dichroism of cleaved pnpla3-TF sample

A) Circular dichroism absorbance spectrum of cleaved pnpla3 sample and buffer
B) Extinction coefficient spectrum of cleaved pnpla3 sample corrected for buffer absorption
C) CDApps predicted secondary structure cleaved pnpla3 sample.

3.5.10 Pnpla3 mass spectroscopy

Following multiple rounds of repeated Ni-affinity chromatography and size exclusion chromatography, a purified band of around 75 kDa was successfully purified from the degradation product in a stable form (Figure 3.39).

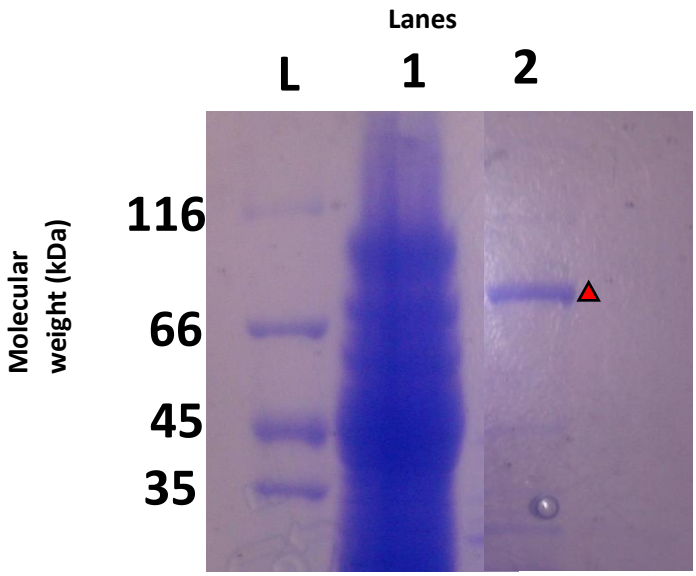


Figure 3.39 SDS-PAGE of purified pnpla3 70kDa band.

The red triangle represents the band used for mass spectrometry
L) Protein ladder **1)** Concentrated sample before purification **2)** Final eluted purified sample

This was confirmed to be pnpla3 using western blotting (data not shown); however, more detailed analysis of this band with mass spectrometry revealed that it consisted primarily of a native *E. coli* protein ArnA (Table 3.7).

Table 3.7 Top hits in mass spectrometry analysis

Accession	Score	Mass	No. of matches
ARNA_ECOBW	7591	74869	428
ARNA_ECO24	7511	74867	426
ARNA_ECOL6	7018	74974	376
ARNA_ECO81	6953	75031	368
ARNA_ESCF3	2650	74772	149
ARNA_SALA4	1428	74075	74
GLMS_ECOLI	1300	67081	40
K2C1_HUMAN	869	67170	23

3.6 Discussion

3.6.1 Expression of PNPLA3

3.6.1.1 Expression trials

A range of expression constructs was used. Nevertheless, the proteins did not consistently appear at the correct molecular weight and expression levels were extremely low. The most encouraging expression clones were pnpla3 and clone A4. These both expressed protein at the expected weight molecular weight, but the results were often unreliable between repeats and the quantity of protein expressed still consistently low.

The two clones which performed best in the expression trials (pnpla3 clone and clone A4) were both full-length clones with a large solubility tag (TF and MBP). This implies that the smaller constructs may not fold successfully, may be deleterious to the cell or else that the protein may not be soluble without an appropriate tag when expressed at high levels in *E. coli*.

The most successful expression conditions were those maintaining a lower temperature during induction, likely because this facilitates slower folding. However, the difference between expression conditions were generally minimal.

3.6.1.2 Large scale expression

The expression of both constructs improved when using larger volume systems. This could be due to improved aeration of the samples, which were grown in large 2L baffled conical flasks, as opposed to 50ml falcon tubes in the expression trials.

Millimolar quantities of both pnpla3 and PNPLA3 were produced from only a litre of LB medium, which implies that PNPLA3 is not toxic to *E. coli* when present at high concentration although there was notable protein degradation present even within the cell. The degradation was evidenced by large smearing and broad peaks on SDS-PAGE.

Attempts to reduce degradation through cold expression, maintaining a cold temperature during lysis and using protease inhibitors, was unable to effectively reduce degradation.

Further, in addition to traditional degradation patterns, both pnpla3 and PNPLA3 appeared as two closely positioned or ‘double’ bands on SDS-PAGE gels. This suggests the presence of a single significant degradation product. However, increasing the concentration of protease inhibitors in the cell lysis buffer, and maintaining cold temperatures during purification did not seem to

alter the ratio of these bands as would be expected by simple degradation; similarly, the band ratios did not change overtime when stored.

The fact that this double band pattern was observed in experiments of both pnpla3 and PNPLA3 proteins, suggests this may be an intrinsic property of the patatin domain. The double band pattern has been observed in previous expression of PNPLA3 in *E. coli*. At this time, both bands were determined to be full length PNPLA3 using mass spectroscopy.[313]

This suggests that this double banding may not represent degradation of the protein but rather anomalous running on the SDS-PAGE gels caused by uneven binding with SDS during sample preparation. Proteins with sharp hydrophobic elbows may not fully unfold in Laemmli buffer and are known to bind SDS unevenly, resulting in anomalous running on the gel.[321]

High weight molecular multimers were also observed with SDS-PAGE and western blots;[161] This suggests that the protein forms multimers *in vitro* which are resistant to reducing and denaturing agents and therefore must form highly stable interactions.

3.6.2 Purification of PNPLA3

In previous studies, PNPLA3 often had to be purified from the membrane fraction,[159] and in one study undertaken in *E. coli* using a TF tag half the protein was found in the membrane fraction after cell lysis.[166] In the present studies, protein was found in the membrane fraction when a non-mechanical technique was used to lyse the cells, but not when the cells were sonicated. Thus, the reported presence of PNPLA3 in the membrane fraction most likely reflects incomplete lysis of the cells with the result that lipid droplets remain associated to the cell membranes.

Both pnpla3 and PNPLA3 proteins behaved the same throughout the purification process. Only crude levels of purification was achieved for both. The protein-tag complexes were initially purified by nickel affinity chromatography. However, when the protein was passed through a size exclusion chromatography column, large multimers were identified as the predominant constituents of the sample which were unable to be separated.

It is common for overexpression of a protein in an *E. coli* expression system to lead to poor folding of the protein and the formation of insoluble inclusion bodies and other high molecular weight aggregates; It is possible that the multimers observed with both pnpla3-TF and PNPLA3-MBP due to these factors.[322]

While using an *E. coli* expression system could be causing poor folding of the protein, similar high weight multimers having previously been identified from expression in insect cells, HEK cells and purified from human blood. When combined with the fact that both homologues behave in the same fashion despite having different solubility tags; This suggests that the multimers may be a natural phenomenon of the protein.[161,166] Indeed, the homologous protein patatin, is also often found *in vivo* in aggregated denatured states.[323161]

To date there is very little information or discussion of the multimeric form of the protein in the literature. The lack of characterisation of these multimers may be a confounding factor which underpins inconsistencies observed in the activities of the protein, as it could influence the activity of the protein.

These large molecular weight multimers are highly resistant to breakdown, showing no shift in composition under a wide range of denaturing conditions, for example after buffer exchange into urea or guanidine-HCl and are stabilised even further by use of detergents.

However, ammonium precipitation and re-dissolving the pellet into a denaturing agent did result in some break down of the multimers into smaller weight fragments. These fragments appeared to have a molecular weight of only 10 kDa but did run anomalously through a size exclusion column most likely due to the high molarity in the salt. Attempts were made to refold the protein using buffer exchange but the large multimer did not reform and most of the protein was lost to precipitation.

After purification with nickel affinity chromatography and several rounds of size exclusion chromatography, at the protein purity approached 95%, if it is assumed that both bands on the gel were the same protein. Purification of a single band remained challenging under any conditions, but a protein sample was obtained at a size which corresponded to the lower weight band of PNPLA3. However, mass spectroscopy confirmed that this purified band was contaminating *E. coli* protein ArnA (*vide infra* section 3.6.5).

Thus, it would seem that PNPLA3 was removed from solution during the size exclusion chromatography. Whether this was because of the presence of large inclusion bodies whose passage was impeded, interactions with the matrix, or loss of solubility under pressure is not known. However, this suggests that size exclusion is not a good approach for the purification of PNPLA3, or that additional steps are needed to better stabilise the protein in solution.

3.6.3 Cleavage of fusion protein tags

Both pnpla3 and PNPLA3 were able to be cleaved from their tags with minimal precipitation. A larger quantity of the tag remained after cleavage, which does suggest some loss of solubility. However, contrary to this, when the samples underwent size exclusion chromatography, the multimers persisted and separation of the target proteins and tags was not possible.

The pnpla3 sample underwent circular dichroism which confirmed that there was no change in the structure in solution after cleavage occurred. This makes the sample a poor candidate for crystallisation, since tags of this size are unlikely to readily crystallise as binding partners.

3.6.4 Activity assays

An activity assay was successfully devised to measure lipase activity using DGGR as a substrate was able to detect changes in protein activity. This will be a useful technique for future work on purified protein samples and to facilitate ligand screening.

Since the CD spectra of pnpla3 had very low levels of ordered secondary structure, it was expected the protein was not well folded. However, both PNPLA3 and pnpla3 samples did exhibit lipase activity *in vitro*, showing the purified samples were at least in part correctly folded.

Human PNPLA3 had roughly three fold higher activity than the murine protein, which is consistent with findings from Kumari *et al.*.[166]

The precise function of PNPLA3 is not certain and so it is unclear whether the low activity observed is in fact a true reflection of PNPLA3 activity, or because of improper folding, lack of yet identified post-translational modifications or missing binding partners.

There is also the possibility that this activity was in fact due to contaminating ArnA within these samples.

3.6.5 Contamination

ArnA has a range of activities which overlap with PNPLA3 (Table 3.8), and a sample purified from untransformed *E. coli* displayed similar activity levels to those observed with the PNPLA3 samples.[121]

Table 3.8 Comparison of properties between ArnA and PNPLA3-MBP

	ArnA	PNPLA3-MBP
Molecular weight	75	105
Apparent molecular weight on gels	76	100 and 76 (doublet)
Multimer	Heximeric	Heximeric equivalent mass
Activities	Hydrolase Formyltransferase Decarboxylase Dehydrogenase	Hydrolase Acyltransferase

The native activity of PNPLA3 is uncertain; a propensity to bind contaminants could impact on activity assays at low levels and may explain some of the contrary results reported in the literature to date. Notably, the study which provides the strongest support for LPAAT activity used protein expressed within an *E. coli* system and included only one nickel affinity chromatography purification step, which is unlikely to have provided a purified sample.[166] These results must therefore be treated with caution as the change in activity ascribed by the workers to the I148M variant may simply reflect changes in ArnA levels, or the ability of PNPLA3 to interact with binding partners such as ArnA.

Any further expression within *E. coli* should be performed in strains lacking ArnA such as LOBSTER.[324]

3.6.6 Crystallisation screening

After several rounds of purification, the protein samples were not of high enough purity to make crystal formation probable and neither pnpla3 nor PNPLA3 formed crystals in any of the explored conditions.

A white ordered precipitate formed under certain conditions, but the quality of these precipitates was too low to be used for further seeding experiments. There was no consistency in the conditions that resulted in the formation of these ordered precipitates which could be used to inform future screens.

The broad peak observed in the size exclusion column implies that the large multimeric forms of PNPLA3 are not uniform and so would not be amenable to any of the structural determination methods.

3.7 Conclusion

Key findings:

Support for previous findings:

- Expression of both PNPLA3 and pnpla3 was achieved using an *E. coli* expression system.
- PNPLA3 was shown to form large multimers *in vitro*.
- Crudely purified PNPLA3 was observed to have lipase activity *in vitro*.

Novel findings:

- A major contaminant from *E. coli,* the protein ArnA, was identified.
- A wide range of buffer conditions were shown not to impact PNPLA3 stability.
- *E. coli* was identified as a poor expression system for PNPLA3.

The expression and purification of PNPLA3 has proved extremely challenging. Large quantities of PNPLA3 has been successfully expressed. However, problems with degradation, insolubility and anomalous behaviour on SDS-PAGE remain unresolved.

ArnA was identified as a significant contaminant of PNPLA3 after Ni-affinity chromatography, which co-purified with PNPLA3. Multiple characteristic similarities with PNPLA3 made this a difficult contaminant to remove, viz: The formation of high weight molecular multimers in solution, similar molecular weight as degradation products of PNPLA3, appearing on western blots using His-tagged antibodies and having low levels of lipase activity. This highlights the need to verify *E. coli* obtained samples are not contaminated with more precise analytical techniques in any future studies.

Both the human and murine homologues of the protein behave similarly *in vitro* and form high molecular weight multimers of around 670kDa which could not be separated with non-destructive techniques.

A homogenous protein sample is needed for crystallisation and the fact that PNPLA3 expressed in *E. coli* was so unstable, presented a range of multimeric forms and constant degradation makes this a poor starting point for future structural work. Ultimately, the inability to separate these large molecular weight multimers, and remove the large solubility tags from the protein, show that an alternative expression system is likely needed for further structural investigation.

Similar observations of high weight molecular multimers both on SDS-PAGE gels and using size exclusion chromatography suggest that this may be intrinsic behaviour of PNPLA3 rather than

simply a side effect of expression in *E. coli*. This may mean that the PNPLA3 itself is not readily amenable to experimental structure determination through traditional methods.

3.7.1 Reflections and Next Steps

The clinical importance and significance of the I148 variant in *PNPLA3* was relatively well established by the outset of this project. However, relatively little had been published on the expression and purification of PNPLA3. There was some published work in which an *E. coli* expression system had been used to express PNPLA3.[151,166] As this is a well-established expression system, which when scaled up, is capable of producing the large amounts of protein needed for structural studies, it seemed logical to choose this expression system. It is now clear that several changes could have been made to the initial approach to this experimentation which might have improved the outcome.

First, a more efficient method of testing conditions for expression and purification should have been employed. Experimental factors were changed one at a time based on previous experience with other proteins. However, this becomes inefficient when the number of conditions tested increased substantially due to difficulties in expression. These initial experiments were not scalable, and a better approach would have been to perform smaller expressions using a fractional factorial design to test a wider range of conditions and interacting factors.

Second, mass spectrometry should have been used at each step of the purification process rather than relying on Western blotting to identity the protein fractions. This would have identified the contamination with ArnA at a much earlier stage.

Third, a switch or parallel inclusion of another expression system should have been made at a much earlier stage. Subsequent publications suggest that this may have been no more effective than the current protocols.

Finally, exploring the possibility that this protein needs a binding partner. Other homologous proteins, ATGL/PNPLA2, as well as patatin, and ExoU, all have important activation roles for binding partners. To date there have been no studies on human isolated PNPLA3 looking for binding partners, which could be pivotal in the purification and isolation of this protein.

None of the above modifications or further explorations is guaranteed to produce quantities of purified PNPLA3. In addition, despite an exponential increase in the number of publications in this field, the structure of PNPLA3 has still not been solved. Therefore, an *in-silico* approach was adopted in the hope of gaining further insights into the structure of PNPLA3.

Chapter 4

Homology modelling of PNPLA3

"If you run into a wall and pretend it doesn't exist, you'll never make progress. The wall will never change, so you're the one who has to change."

Hijikata Toushirou

4.1 Overview

The three-dimensional structure of a protein gives insight into both the function of the protein as well as any potential disruption to this structure which may be caused by sequence variation.

Although experimentally derived structures are the gold standard, in many cases, the structure of a protein is not amenable to investigation using standard experimental investigations such as X-ray crystallography or NMR. To date, attempts to solve the PNPLA3 structure experimentally have failed due to difficulty encountered in obtaining millimolar quantities of highly purified protein.

Difficulties have also arisen in obtaining the structure of PNPLA3 via structural modelling due to insufficient sequence homology with other proteins of known structure. Two structural models have been predicted, spanning the first 179 residues of the protein in 2006 and 2010. However, these did not achieve high enough confidence for detailed structural predictions on the effect of the I148M variation.

In this chapter, the three-dimensional structure of PNPLA3 is investigated through a combination of *in silico* homology modelling techniques.

Nine novel structural models of PNPLA3 were developed; including the first full-length homology model and also an improved model of the patatin domain based on a novel sequence alignment with an *E. coli* protein, ExoU.

These models provide insight into 1) The structural architecture of PNPLA3 as a whole. 2) The relationship between the catalytic residues. 3) The properties of the active site. The information gained, in particular on the active site, allows the potential impact of the I148M variant on the local structure to be predicted.

4.2 Introduction

It is crucial that the tertiary structure of the protein be known to an atomistic resolution in order to gain mechanistic insight into the function of the protein. Knowledge of the structure will also allow the assessment of the impact of functional genetic variants; both through the impact on both the local and global architecture of the protein.[325]

A range of both experimental and in silico approaches have been developed to obtain high-resolution structural information on protein of interest. In this chapter, the focus will be on *in silico* methodologies; however, traditional experimental methods of structure determination will first be discussed to better contextualise the advantages and limitations of the *in silico* approach.

4.2.1 Experimental structure determination

To date the primary methods used to determine protein structures to high resolution experimentally, have been X-ray crystallography[326] and nuclear magnetic resonance (NMR) spectroscopy.[327] More recently, advancements in the field of cryo-electron microscopy has seen this emerge as a potential third primary approach for structural studies.[328]

4.2.1.1 X-ray crystallography

X-ray crystallography determines the atomistic detail of a three dimensional protein structure based on the regular diffraction pattern of an incident X-ray beam through a protein crystal lattice.[326]

Of all the experimental structural determination techniques, X-ray crystallography is capable of producing the highest resolution structures, often greater than 1Å, and hence is particularly valuable for structure-based drug design. The majority of previously unknown protein structures have been solved using this approach.

While X-ray crystallography is a powerful technique, it is not without limitations. First, it relies on a proteins ability to form protein crystals, and crystal growth is often the limiting step when using this technique. Second, as any structure solved through crystallography will be in a crystal form, the extent to which the stable crystal form resembles the active protein in solution is often unknown; However, in many cases this has been experimentally validated through inhibitor

design. Third, there is little dynamic information provided regarding the protein structure; loops which may be flexible can be predicted, however, these can be difficult to differentiate from poorly modelled regions. [326]

4.2.1.2 Nuclear magnetic resonance (NMR)

NMR takes advantage of the fact that changes in chemical environment of isotopically labelled nuclei, show a change in absorbance of electromagnetic energy. This can be measured under a large magnetic force and used to mathematically build protein structures.[327]

The advantages that solution NMR can offer over other techniques, are that you can investigate the dynamics of proteins experimentally; facilitating investigation into the protein-protein interactions and protein-ligand interactions under close to physiological conditions. Although NMR often produces a lower resolution when compared with X-ray crystallography, the structures can still have a resolution within the 1A range.

NMR has proven to be particularly useful when examining proteins which have high levels of intrinsic disorder, and can often be challenging to obtain quality data using other techniques, for example Zinc finger containing proteins.[329–332]

Despite these advantages, elucidating information on large proteins of above 50kDa can remain challenging using this technique, and structure determination still relies on the ability to produce high concentration protein samples with simple buffer components. The size limit for NMR has significantly increased over the last decade with the development of transverse relaxation optimised spectroscopy (TROSY), but remains a challenging aspect of the technique.[333]

4.2.1.3 Cryo-electron microscopy (cryo-EM)

Cryo-EM is a form of transmission electron microscopy in which an electron beam is transmitted through an ultrathin grid at cryogenic temperatures. Embedding large proteins on the surface of the surface of the grid, allows the 3-dimensional structure of the protein to be investigated.[328,334]

The advantages of cryo-EM are that the samples can be produced from native protein in solution, meaning that the proteins are trapped in their native solution states and do not require crystallisation. From here different conformations can be extracted experimentally.

In contrast to NMR spectroscopy, which works optimally with structures ≤ 50KD cryo-EM can only be applied to large structures and even with advancements in the technology the current minimal recommended size is ≥ 200kDa. Currently cryo-EM has the lowest resolution of the three experimental techniques; the resolution expected is between 10 and-30Å. Advancements in the technology, such as the introduction of direct electron detectors in 2012, have shown the potential for high resolution structures to be obtained, for example the first structure of E. coli beta-galactosidase which was solved to a resolution of 2.2 A using this technique. However, it is still rare to achieve structures with a resolution this high using cryo-EM.[334,335]

4.2.1.4 *in silico* approaches to structural determination

Advancements in all three experimental techniques, are constantly improving our ability to determine higher resolution structures of challenging proteins, including very large proteins and proteins which form complex quaternary structures or interact with membranes. However, all these techniques are still reliant on the intrinsic ability to be able to produce large milligram quantities of homogeneous protein in a soluble and stable form.

While some proteins are easily isolated and purified, others are not, including PNPLA3. Even when proteins can be isolated and purified subsequent experimental structural determination often proves time and resource expensive. In consequence, computational investigation of protein structure provides an attractive additional or alternative approach.

Historically, structural prediction using *in silico* computational approaches have been unreliable and difficult to accomplish because of high computational expense and limited structural knowledge to inform the modelling process. This has meant that the predominant method to investigate protein structure has remained experimental techniques.[336]

Rapid advancements in the computational capacity available to researchers, the development of software which can take advantage of novel GPU accelerated code and a significant increase in the number of experimentally solved structures, has led to rapid improvements to *in silico* structure determination; which can now be used to overcome some of the limitations of experimental approaches.[336]

4.2.2 Structural modelling of proteins

Structural modelling is the general technique used to predict the three-dimensional structure of a protein *in silico*. This is a crucial element of any *in silico* investigation of a protein and is the first step to all further computational approaches within this research framework (Figure 4.1).

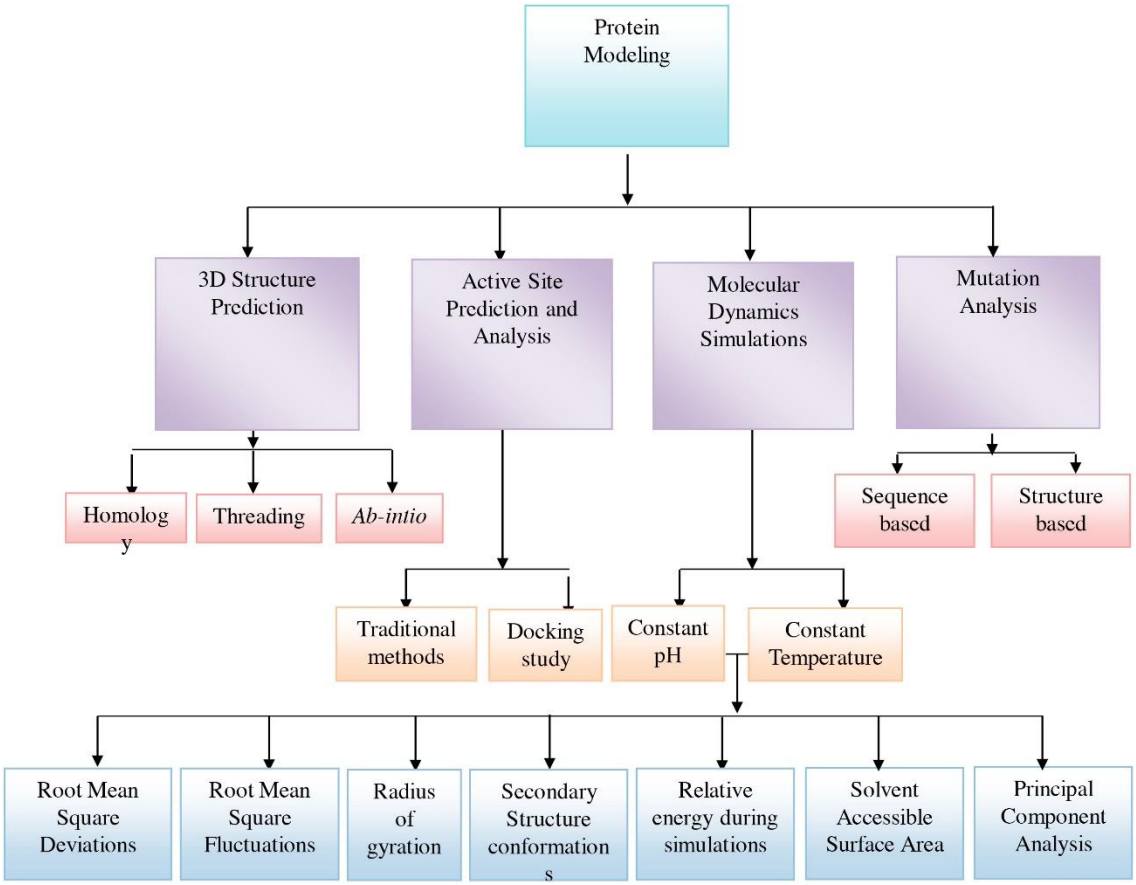


Figure 4.1 The work flow of *in silico* investigation into the structural properties of proteins

(Adapted from Khan et al. 2015).[337]

Structural modelling can be broken down into various specific approaches which use either homology modelling or *ab initio* based methods to create a three-dimensional replica of a protein without an experimentally determined structure.

Homology modelling, also known as comparative based modelling, involves predicting a structure by means of sequence similarities to another known protein structure. This remains the most accurate approach to modelling protein sequences.

By minimally disturbing existing solutions to protein structures, it facilitates the exploration of conformational space while relaxing the stringent need for accurate force fields and the extensive spatial and temporal searching involved in *ab initio* investigations.[338]

The accuracy obtained using homology modelling is surprisingly high, in part due to the fact that the conformation of a protein is more conserved than that of its amino acid sequence resulting in conservation of critical functional motifs.[339] The use of homology has underpinned many of the most accurate modelling approaches shown in the bi-annual critical assessment of techniques for protein structure prediction (CASP).[340]

In order to undertake accurate homology modelling there needs to be significant detectable likeness between the known structural template and the target protein sequence, and an ability to create a significant correct alignment between them.[337]

The ever-increasing number of experimentally solved structure adds to the current database of known folds, which due to the large number of analogous structures is assumed to be limited. This allows the approach to continuously improve until all folds theoretically become known.[341,342]

The most accurate structural models come from close homologues with high levels of sequence identity. While the requirement for sequence identity between the structural template and target protein sequence is decreasing each year along with methodological and computational advances, in general, a minimum identity of 25% is required for the generation of a model which can be utilised at high confidence. However, when carefully applied, even low sequence identity of around 20% can provide suitable models.[343]

When the level of structural identity is insufficient to allow homology modelling, *ab initio* methods can be implemented to increase modelling accuracy. These approaches generally rely on an initial manifold threading based alignment between structural portions of the protein, followed by the conformational sampling of the remaining flexible regions and full spatial exploration of regions in which there is insufficient structural homology through *de novo* sampling.[337] As threading still relies on template information, there is clear overlap with homology-based approaches.

More recently, the combination of both homology modelling and ab initio techniques has produced the most successful modelling procedures. Increased usage of *ab initio* conformational sampling in model refinement has begun to blur the lines between the traditionally disparate disciplines.

In particular, *ab initio* approaches can enhance the structural model of flexible loop regions, which are intrinsically difficult to define through homology because of the huge variation and lack of structural information in crystallographic structures, as well as additional regions of the protein which cannot be explored through homology alone.

4.2.3 Generating protein models

Despite variations in algorithmic approach, most successful modelling processes rely on a chronological implementation of 5 key methods (Figure 4.2):[344]

(1) Identification of the template;

(2) single or multiple sequence alignments;

(3) model building for the template;

(4) model validation;

(5) model refinement.

Full *de novo* modelling begins at step 3, and generates models without information from any templates; although this is rare particularly when working with larger target proteins.

4.2.3.1 Template search and selection

The first step in generating a homology-based model is to identify structures with similarity to the target protein and select those which are most structurally similar for modelling. This usually involves exploration of the large range of databases containing protein structural information such as CATH[345], DALI[346], PDB[347] and SCOP[348].

Basic Local Alignment Search Tool (BLAST) has traditionally been the approach to finding protein homologues, and applies an independent pairwise comparison of the query sequence against each sequence within database.[349]

However, when homology falls below 30% the hits often become unreliable. This has resulted in the development of Position Specific Iterative – Basic Local Alignment Search Tool (PSI-BLAST), which enhances detection of homologues by applying an iterative array search of databases.[350] Use of PSI-BLAST has been shown to identify similar structures for twice as many proteins as the traditional BLAST approach.[351]

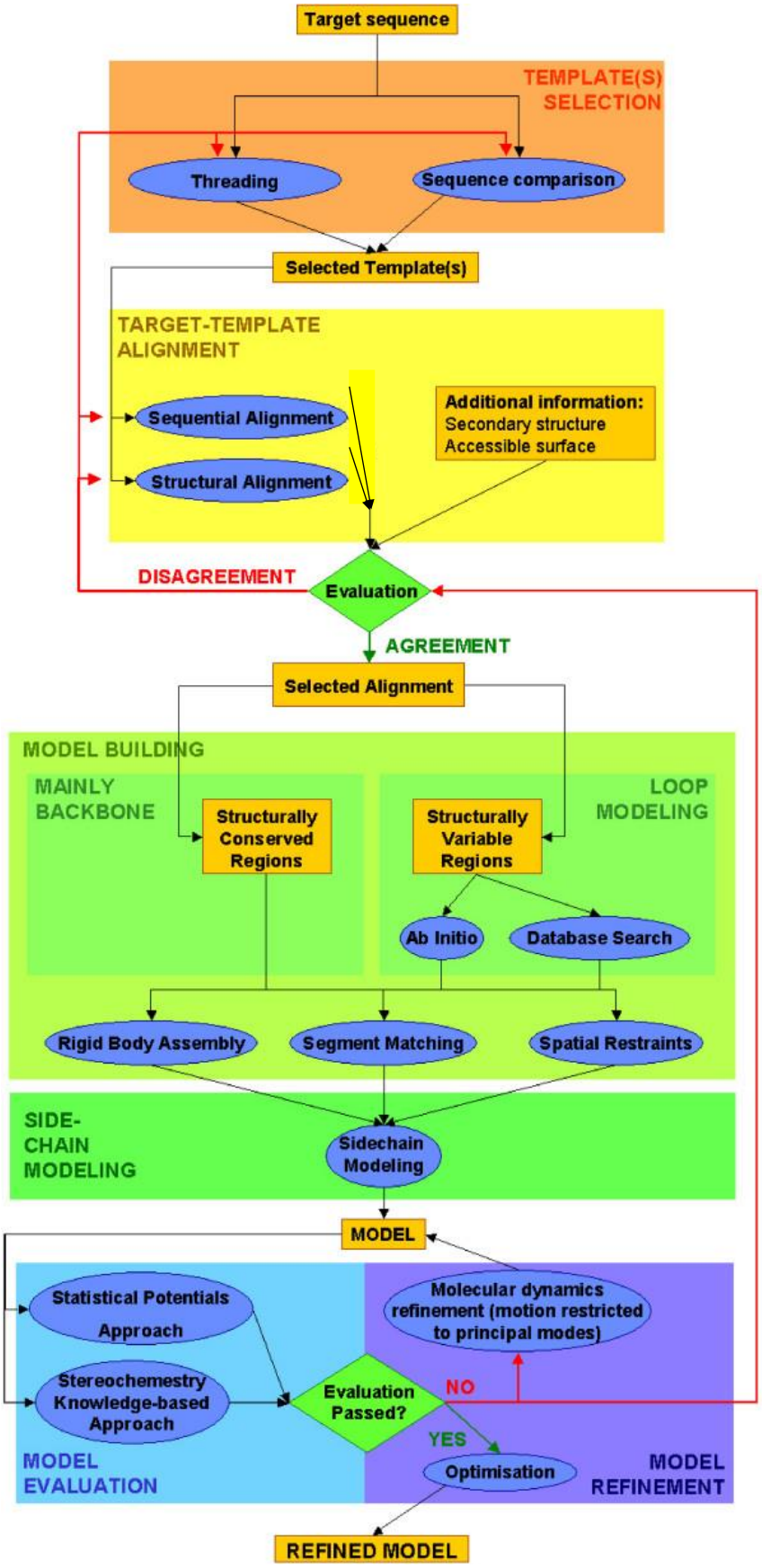


Figure 4.2 The work flow for the structural modelling of proteins

(Adapted from Centeno *et al.* 2005).[344]

Hidden Markov models (HMMs) can also be applied for profile comparison across databases such as HMMER.[352,353] These methods are particularly useful for sequences with low homology and can detect regions of low homology based on a pairwise comparison of HMMs. HMMs often underpin the methods utilised for threading of small protein fragments and are used for example by the HHPred server.

Selection of each template to be included, is of course primarily based on the sequence similarity between the target and template. However, it is also important to include additional information to select a template which will lead to the most accurate model. This includes the quality of the experimental data of the template structure, the similarity of the experimental conditions between query and template, for example the presence or absence of a ligand, and the phylogenetic similarity. Each algorithm will often apply different weighting to each characteristic, although the desired use of the model can help to steer weighting. For example if ligand based docking is planned then the resolution of experimental data will play a more pivotal role in template selection.[344]

When threading is being used for more disparate homologues, relevant templates must undergo more advanced matching criteria to assess the putative structural environment of each template and their relevant compatibility.[354]

4.2.3.2 Template-target alignment

Once the templates have been selected, in order to appropriately construct a three-dimensional model based upon the query sequence, optimal alignment between the query and template must be achieved.

Several strategies can be used for template-target alignment based either on iterative search-based methods, HMMs, or alignments with integrated structural information. No one methods is superior to the others; it is simply important to select a well-established method.[337]

Iterative search-based methods rely on progressive alignments of the related sequences. This is the basis of the frequently used CLUSTALW program, in which additional weighting applied to include residue specific gap penalties.[355,356]

HMMs provide a probabilistic approach which is highly effective at detecting conserved patterns in multiple sequences.[357] This is used in the SATCHMO algorithm, which relies on the construction of a similarity tree of multiple sequence alignments that can be compared at each internal node of the tree.[358]

Finally, methods such as PRALINE can be used to incorporate not only alignment of the sequences, but any inherent secondary structure information that might be present.[359]

4.2.3.3 Three-dimensional model building

The prediction of an entire protein model cannot be achieved using a single approach, because of inherent differences in the characteristics across the protein chain. In order to effectively utilise homology modelling, the main protein backbone is first modelled which forms the fundamental structure of the model. This structure can then be used to enhance the prediction of flexible loop positioning and side chain orientations (Figure 4.3).

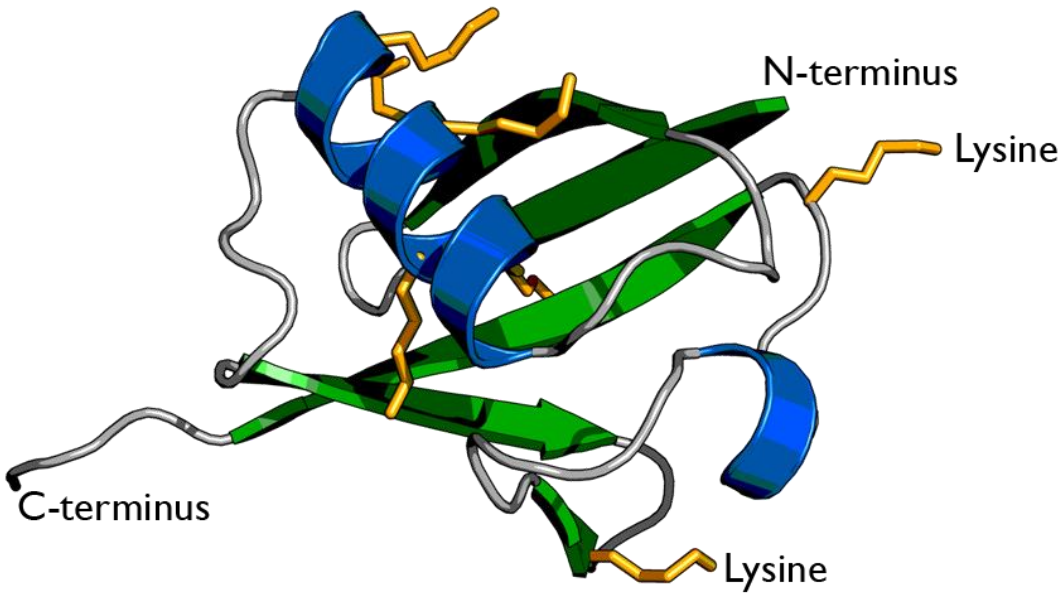


Figure 4.3 Example 3D protein structure

The protein shown is ubiquitin. Helices are shown in blue. Sheets are shown in green. The flexible loops are represented in grey. The lysine residues are highlighted in yellow as an example of an amino acid side chain.

Protein backbone modelling

Modelling the protein backbone is the most challenging step in structure prediction as it will be used to inform all subsequent steps. This step is generally achieved through use of rigid body assembly or spatial restraints. Both approaches can successfully generate structural models; however, will result in different biasing based on the similarity to the original structural template.

The first step in rigid body assembly is the formulation of structural alignments within the query structure and template structure. In a traditional homology modelling approach, this will be done with large protein segments, separated by flexible loops. These alignments are then treated as rigid bodies, and spatially rearranged to minimise energy and align with the template to form a structural model. A similar approach can be taken in de novo approaches. However, the structural alignments are based on much smaller secondary structural folds, which then necessitates a very wide spatial search to complete the model.[360]

Assembly via spatial restraints is performed by generating restraints on each atomic position based primarily on the template protein. These restraints function to constrain the model and prioritise structural similarity with the template without absolute positioning relying on the assumption that the homologous structures are likely to have similar structural positions and intramolecular interactions. To improve the models further, traditional molecular mechanics restraints are additionally applied which act to conserve good bond geometry and avoid steric clashes which could otherwise occur. Additional restraints can be added to account for other contributing factors to the protein core such as hydrophobicity considerations.[361]

Loop modelling

The loop regions link together other core secondary structural elements within the protein. It is in these regions that most sequence variation occurs, and changes which modify the specificity and activity of the protein are found.[362] This means that modelling loops with high accuracy is pivotal to the quality of the model and to investigating the effects of many common genetic variants.

Loops, unlike the protein backbone, are inherently flexible and so can exhibit multiple confirmations. This means that they are often missing in crystal structures, hence constructing loops from homology alone is very limited and normally not possible.

Loop modelling can be achieved by either database searching and aligning, or *ab initio* methods relying on Monte Carlo simulations. Loops are likely to be much shorter than the protein backbone, and when anchored to key secondary structure points within the template, allows much greater accuracy than would be achieved from free *ab initio* modelling.

The modelling of loops can generally be assumed accurate when the regions are less than 14 residues in length. However, when loops are longer or interact with one another, the confidence in the loop environment is significantly reduced.

Side chain modelling

Side chain modelling is generally based on energetic and steric environmental conditions. By minimising the energy of the local region, it is possible to acquire favourable side chain positions while detecting side chain couplings and side chain – main chain interactions.[363]

This means the modelling of side chains is generally based on the template structure, without adjusting the backbone. A flexible rotamer based search is applied, during which common rotamers are compared and selected based on the lowest energy conformation.[364]

The modelling side chains can be very accurate. however, it relies heavily on the correct positioning of the protein backbone. In models with low homology and low confidence in the secondary structure, side chain packing cannot be accurately achieved

4.2.3.4 Model validation

The assessment of model quality is crucial. Where multiple models are generated accuracy predictions must be used to determine which model to select to take forward.

There are two main ways to validate a model system. The first is to compare the predicted and template structure to determine the degree of similarity. Proteins highly similar to the template are likely to be more conserved and so of higher quality. Comparison can also be made against a range of protein structures deposited in the PDB , to find similarity with other known protein families.[337]

The second approach is to judge the quality of the structure based on the physiological properties of the residues within the protein. This approach looks more closely at the stereochemistry of the model based on the assumption that the better the stereochemistry the greater the confidence in the model. One example which focuses solely on this approach is the PROtein Structure Evaluation Suite and Server (PROSESS).[365]

The methods for judging the quality of a model often overlap with approaches used to determine confidence in experimental structures. However, without experimental data the results remain theoretical.

Over recent years, predicting the quality of structural models has become of increasing interest as models have become accurate enough to be used for real world applications. This is an area which is still undergoing development and therefore is under frequent review by CASP.[366]

4.2.3.5 Model refinement

The final step in producing the completed structural model is to perform a refinement step to fine tune the alignment of the protein backbone, side chains and loops. This will usually begin with energy minimisation of the predicted structure under a molecular mechanics forcefield, followed by some additional Monte Carlo sampling focused on regions which are predicted to be erroneous.[367,368]

It is now possible to further refine homology models by applying an additional step of molecular dynamic simulation. This attempts to sample the native phase space of the protein under physiological conditions and subsequently minimise the energy of the system. This approach has been used to improve the quality of many models; however, because of the complexity applying this technique to different proteins, which would often each need unique simulation parameters (particularly large proteins with metal ion groups), and the high computational expense of running the simulation, it has not been widely integrated into current modelling software packages.[369]

4.2.4 Software variations

It has become common for modelling software to be packaged together to provide full end to end modelling solutions. This has the benefit of not requiring user input on each step of the modelling process and opens up the field to non-specialist users.

A large range of software has been developed which can perform structural modelling, each with unique algorithms implementing different degrees of homology and *ab initio* modelling. Popular software includes: HHpred[370], M4T[371], Modeller[372], ModWeb[373], PHYRE2[374], SWISS-MODEL[375] (predominantly homology based), Rosetta[376] (predominantly *de novo* based), I-Tasser[377], IMP[378], Robetta[379] (predominantly mixed); however, this list is far from exhaustive.

The I-TASSER and SWISS-MODEL were both used in the present study. I-TASSER has performed consistently well in CASP assessments as one of the highest scoring approaches to structural modelling using a combined homology *ab initio* approach, while SWISS-MODEL performs well using a more traditional homology based approach.[380]

4.2.5 SWISS-MODEL

SWISS-MODEL was the first fully automated server for homology modelling of proteins and has been consistently maintained and updated. The most common implementation of SWISS-MODEL is provided as a Webserver, which can be accessed without the need to download any software or database information.

As a homology only based modelling server, the SWISS-MODEL process conforms to a four-step modelling approach as described above, without significant model refinement.

4.2.5.1 SWISS-MODEL modelling process

A range of template data is regularly extracted from PDB with release of new structures and added to SWISS-MODELs own database, the SWISS-MODEL template library (SMTL). This allows efficient access to a range of template information by their search algorithms and removes low quality structural data.

The target sequence is used in a step-wise search using PSI-BLAST to identify initial high sequence similarity templates, followed by HMM-HMM profile methods and HHblits to search the database for both near and more distant homologous templates.

The best template match is then determined based on modelling with a probability density function (PDF) which incorporates not only sequence identity and similarity between the template and target sequence, but a range of other descriptive predictions; such as predicted solvent accessibility and predicted secondary sequence agreement.

The model is constructed based on a best fit alignment of a monomeric subunit of the structure, by default using ProMod3. In cases where ProMod3 does not provide adequate modelling, particularly in loop regions, an alternative is produced using MODELER.

Side chains are reconstructed based on weighted positions of the residues in template structures, beginning with the most conserved residues. The side chains are adjusted by replacing template side chains with most feasible side chain conformation selected from a backbone dependent rotamer library.

A final energy minimisation is applied with the Gromos 96 forcefield to improve stereochemistry and remove possible steric clashes.

Oligomeric structures are not directly modelled. However, in the case a quaternary structure exists within the template structure, the probability of conservation of the interface is predicted.

A random forest is generated based on the properties of the interfaces between polypeptide chains, for example the sequence similarity, identity, interface hydrophobicity and consensus occurrence of the interface within the structural database. If the results are in a desired range, the quaternary structure will be said to match that of the template. A similar process is used to address ligands which exist within the templates.[375,381–385]

4.2.5.2 Quality assessment

Quality assessment is a focus of the SWISS-MODEL implementation and is undertaken by reference to a wide range of statistics. The local quality plot is calculated as a per residue confidence of similarity between the template and the model structure.

The QMEAN is the composite scoring function based on the geometric properties of the model.[386] It is a weighted formula used to assess the agreement of predicted and calculated secondary structure and solvent accessibility.

The QMEAN Z-score is a statistic which is used to assess the general normalised QMEAN against those for high resolution crystal structures in the PDB.

Finally, a Global Model Quality Estimation (GMQE) is calculated which takes into account both the confidence of the template target alignment and the geometric properties of the protein. This is currently used as the most accurate quality estimation statistic by SWISS-MODEL.

4.2.6 I-TASSER software suite

The I-TASSER software suite combines homologous threading and *ab initio* methods to create the most confident protein model from an input protein sequence. Like SWISS-MODEL, this is also made available online as a fully automated webserver.

The I-TASSER modelling process uses a hierarchical multistep approach to help model even sequences with low homology (Figure 4.4). The way the software is constructed allows modular improvement to each stage of the modelling system, and because the algorithm is under frequent development, specific elements within the algorithm are likely to change over time.

4.2.6.1 Modelling process

The target sequence is first threaded using variations of secondary structure enhanced profile-profile threading alignments with multiple alignment algorithms (vis. PSI-BLAST, Hidden Markov

model, Needleman-Wunsch and Smith-Waterman). The alignments are excised and used to construct a model based on the α carbons and side chain centre of masses of the structure. Parts of the protein without adequate alignments are then predicted with ab initio methods.

The conformational space of the model is searched with replica-exchange Monte-Carlo simulations and clustered with SPICKER. Cluster centroids are generated by averaging all structures in the centroid and these are used to search for further structural alignments in the PDB as a starting point, and to generate restraints for a second round of fragment assembly and conformational exploration.

The resulting structures are clustered once again and the lowest energy structure from each cluster taken forward; whereby backbone and side chain atoms are then built using Pulchra and Scwrl_3.0 respectively.

While I-TASSER is not purposefully designed to deal with multi-domain proteins, when regions longer than 80 amino acids have no structural alignment, the large gaps are treated as domain boundaries. These regions are then modelled, both separately and as a full sequence. To mitigate a biasing influence of separate "domains" on one another within the protein, the separate structures are treated as more structurally confident, and the full sequence is used to orient the various domains together, while minimising steric clashes.[203,387,388]

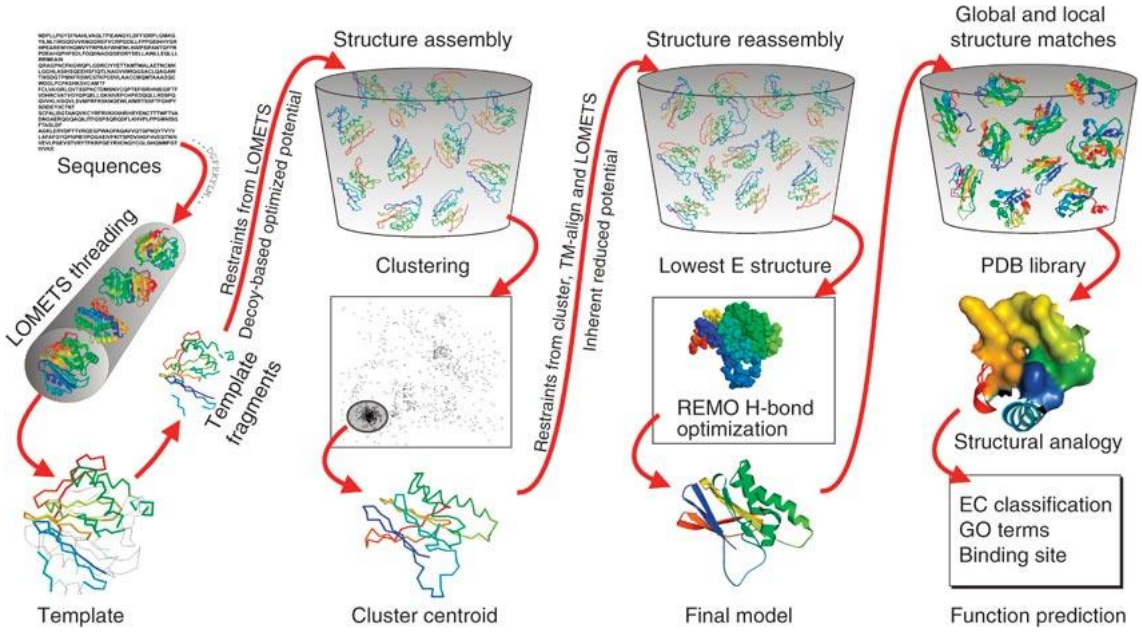


Figure 4.4 A schematic representation of the I-TASSER protocol for protein structure and function predictions

(Adapted from Roy *et al.* 2010).[389]

4.2.6.2 Quality assessment

The quality assessment in I-TASSER is based primarily on three scores *viz.* the template modelling score (TM-Score) and the confidence score (C-score), which offer information on the overall protein structural quality, and the normalised simulated temperature factor (B-factor) used to provide a measure of local confidence.

The TM-score is a description of the similarity between two protein structures utilising a size dependent algorithm which assesses the pairwise alignment assessment of each residue within the sequence.[390]

The C-score is a unique scoring method used by I-TASSER, which consists of two key terms. The first is a description of the convergence of the clustering between all models clustered within a centroid; this assesses the degree of consistency with the model that will be used to determine the restraints applied by the I-TASSER potential in the second round of assembly. The second is a description of a normalised Z-score between the model and the best template alignment. This score has been shown to correlate strongly with model quality and is the score used to determine the most confident model.[391]

A normalised simulated B-factor is used to provide an estimate of local confidence in the model. This is generated using a combination of coverage of threading alignments, divergence of I-TASSER simulation decoys, and sequence-based secondary structure and solvent accessibility predictions; by which higher threading coverage, α-helix and β-strand regions as well as buried residues are determined to be more confident. This successfully aligns with local confidence of the model, but must be differentiated from the absolute values of the classical B-factor.[392]

4.2.7 Limitations of *in silico* structural modelling

Homology modelling provides an extremely useful approach to predict structure when experimental techniques are not available. However, there are still limitations to the application of the technique.

The success of homology modelling relies on the conservation of structural motifs between similar sequences. However, as the similarity between templates and the protein of interest decreases, so too does the probability that these structures will be conserved. This means that to varying degrees, structural distortions from the template will be improperly imposed through the model.[393]

In addition, assigning restraints on the model and assessing the quality of the resulting conformations also relies on information from known structures. This is inherently an incomplete dataset, and while successful for the majority of structures could introduce bias, particularly in structures with no known homologues which could have unique or undiscovered protein folds.

Incorrect template assignment and alignment will generate another source of error. When the template and query share above 40% sequence identity, incorrect assignment and alignment will occur only rarely. However, with lower degrees of sequence homology errors can occur and if they do this will have a significant effect on the subsequent steps; this source of error is not open to amendment in later stages. Thus, template assignment should be very carefully monitored especially in the when sequence homology is low.

Dealing with insertions and deletions between the template and protein of interest is particularly challenging. This is more so the case when the insertions and deletions are long as there will be a mismatch between regions, and this can make mapping difficult. If insertions force the stem residues out of alignment, the model is likely to be inaccurate.[394]

Proteins are subject to small alterations in structure which do not affect the protein backbone, but the more detailed sidechain packing. There are often errors in the prediction of this detail, as even rotamers of identical residues may be changed based on alterations in the biochemical surroundings of other nearby groups.[395] This error is the most likely to go unnoticed, but when it impacts critical residues, such as key functional residues in the active site can severely hinder the interpretation of the model.

Generating tertiary structural models of proteins has become a simple task to perform practically using the range of software suites and online servers which have been created requiring only an amino acid sequence input. This facilitates access to structural modelling of proteins relevant to a wide range of users who need not be experts in structural biology or modelling. This is largely of benefit to the scientific community but comes with an inherent downside of any 'black box' approach, namely that users not fully aware of the underlying processes might misinterpret the data. In addition, the system is essentially 'closed' so that individual necessary manipulation of the steps of the modelling process cannot be made.

A further limitation of homology modelling is that the introduction of some degree of error into the model is inevitable, and each step in the modelling process will naturally introduce some degree of error; this is due to both the intrinsic nature of the homology modelling process and the additional likelihood that parameters will be incorrectly assigned.[396] However, while most of

the software suites will generate an ensemble of models and use estimates of model quality to select best models, the detail of model quality is often not passed on to the end user, who generally only receive an overall summary statistic and so are not able to access specific information on model quality.

Quality assessment remains one of the most crucial and challenging aspects of structural modelling. Protein models can be created which provide reliable generic insight into the overall backbone of a protein and can be used to investigate functional implications and even putative ligand binding for drug discovery. However, without the ability to adequately assess the accuracy of the model, its interpretation will be hampered. This is one of the strongest limitations of homology modelling and one of the biggest challenges to overcome.[397]

4.2.8 Models of PNPLA3

The structure of PNPLA3 has not been solved experimentally despite the fact that it is a protein of significant biological interest. In addition, very little effort appears to have been expended in attempting to structurally model the protein. This is likely due to the fact that PNPLA3 exhibits very low levels of homology with any known protein, so that in silico modelling is likely to be challenging

Two models of PNPLA3 have been reported, both based on a native patatin, Pat17 using a strictly homology-based approach. The first, produced in 2006, was a model of the first 179 amino acids of the protein at the N terminal; it has been used for all further in silico investigations (Figure 4.5).[149,158]

The second, produced in 2010 also modelled the first 179 amino acids but for both the wild type (Ile148) and the variant (Met148) proteins. Based on these models the suggestion was made that the I148M variant is associated with reduced catalytic activity, and hence a loss of function, as the methionine substitution was postulated to cause steric hindrance at the active site (Figure 4.6).

No further investigation of this N terminal model haves been undertaken. In addition, no attempts have been made to model the whole protein. Significant advancements in the performance of modelling software make it likely that more accurate modelling will be possible and may shed new light on the structure and possible function of PNPLA3 and its important variant.

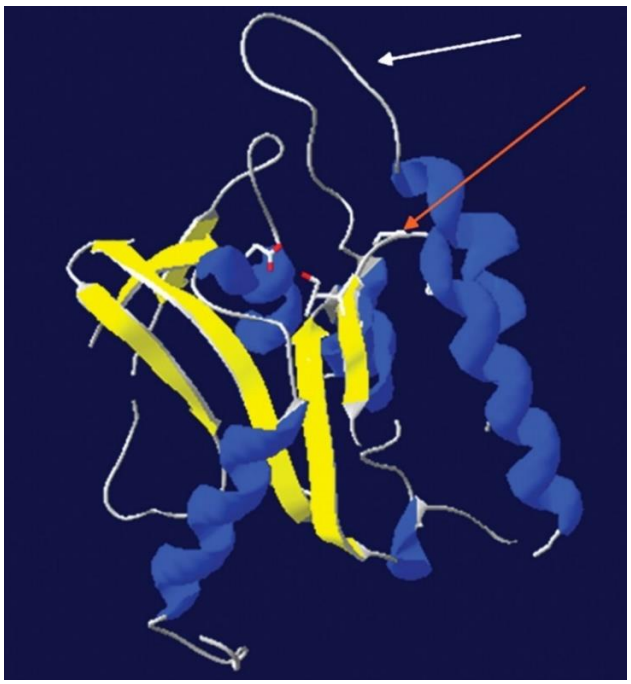


Figure 4.5 Low-resolution homology model of the human PNPLA3 patatin-like domain

This was generated using the DeepView/Swiss-PDB Viewer. Regions predicted to fold as β-sheets and α-helices are shaded yellow and blue, respectively. Side chains of the catalytic aspartate and serine residues are rendered in stick format and the atoms coloured using standard Corey, Pauling & Kultun (CPK). The glycine-rich region is indicated by the red arrow and the putative 'lid' region by the white arrow (Adapted from Wilson *et al.* 2006).[149]

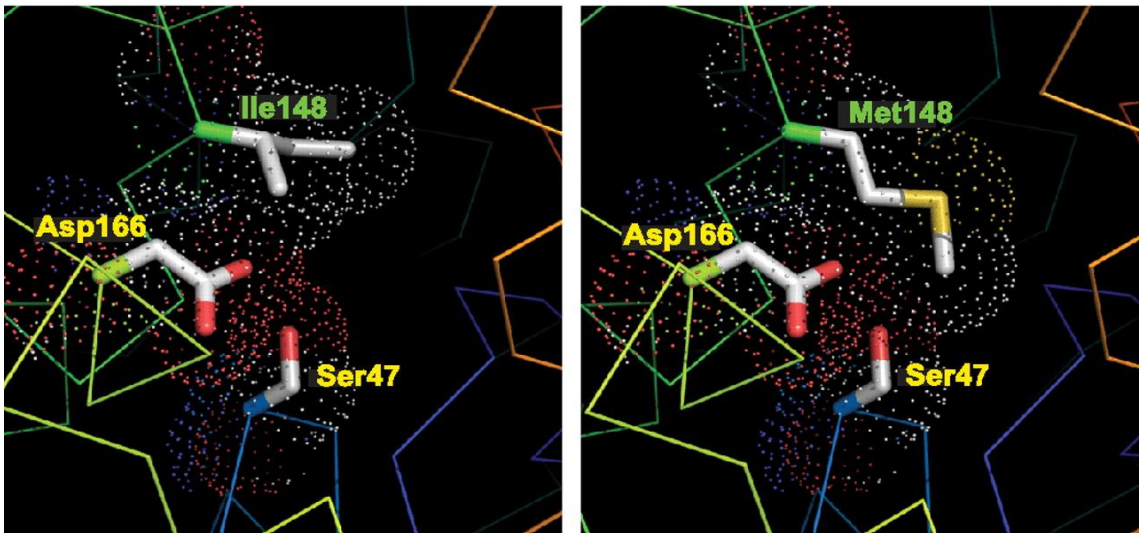


Figure 4.6 Structural model of wild type and mutant (I148M) PNPLA3

The domain structure of PNPLA3, showing the patatin-like domain (black) and locations of the catalytic dyad (Ser47 and Asp166) and the I148M substitution is shown. Structural models of normal (Ile148) and the variant (Met148) PNPLA3 are shown in the left and right panels, respectively. Protein traces are rainbow-colored from N to C terminus (blue to red) with the side chains of the catalytic dyad residues (positions 47 and 166) shown. The dots indicate a space-filling model corresponding to van der Waals atomic radii with oxygen and sulphur atoms coloured red and yellow, respectively (Adapted from He *et al.* 2010).[158]

4.3 Aims

The aim of this chapter is to achieve a model of PNPLA3 with the highest possible accuracy, through the generation of an ensemble of homology models. This will help to predict the overall structure and domain architecture of the protein, as well as providing insight into the location of the key I148M variation.

If adequate quality is achieved, these models can be subjected to molecular dynamic simulations to assess the potential impacts of the I148M variation on the dynamic system, and for the potential identification of a novel construct which shows increased stability *in silico*.

4.4 Methods

4.4.1 Generating models of PNPLA3

The protein sequence of human PNPLA3 (Uniprot Accession: Q9NST1) was retrieved from the Uniprot database.[121] This sequence was used as the basis for all initial computational modelling investigations (Figure 4.7).

MYDAERGWSLSFAGCGFLGFYHVGATRCLSEHAPHLLRDARMLFGASAGALHCVGVLSGIPLEQTLQVLS
DLVRKARSRNIGIFHPSFNLSKFLRQGLCKCLPANVHQLISGKIGISLTRVSDGENVLVSDFRSKDEVVD
ALVCSCFIPFYSGLIPPSFRGVRYVDGGVSDNVPFIDAKTTITVSPFYGEYDICPKVKSTNFLHVDITKL
SLRLCTGNLYLLSRAFVPPDLKVLGEICLRGYLDAFRFLEEKGICNRPQPGLKSSSEGMDPEVAMPSWAN
MSLDSSPESAALAVRLEGDELLDHLRLSILPWDESILDTLSPRLATALSEEMKDKGGYMSKICNLLPIRI
MSYVMLPCTLPVESAIAIVQRLVTWLPDMPDDVLWLQWVTSQVFTRVLMCLLPASRSQMPVSSQQASPCT
PEQDWPCWTPCSPKGCPAETKAEATPRSILRSSLNFFLGNKVPAGAEGLSTFPSFSLEKSL

Figure 4.7 PNPLA3 protein sequence retrieved from Uniprot database. Each letter corresponds to a single amino acid within the protein sequence.

Models were generated based on homologous template sequences of Patatin, VipD, ExoU and PlpD (Table 4.1; Figures 4.8 - 4.11). Templates were aligned using Clustalω[398] and the structural similarity of the templates compared using mTm-align.[399]

Table 4.1 Summary of template protein structures used for homology modelling

PDB ID	Protein	Host species	Multimeric form	Resolution (A)
1OXW	patatin-17	*Solanum cardiophyllum*	homo-3-mer	2.2
4AKF	VipD	*Legionella pneumophila*	homo-3-mer	2.9
3TU3	ExoU	*Pseudomonas aeruginosa*	hetero-2-mer complexed SpcU	1.92
4KYI	VipD	*Legionella pneumophila*	hetero-2-mer complexed human GTPase Rab5	3.08
4AKX	ExoU	*Pseudomonas aeruginosa*	hetero-4-mer complexed SpcU	2.94
5FYA	PlpD	*Pseudomonas aeruginosa*	homo-2-mer	2.14
4PK9	Patatin	*Solanum cardiophyllum*	monomer	1.96

MATTKSFLILIFMILATTSSTFAQLGEMVTVLSIDGGGIRGIIPATILEFLEGQLQEMDNNADARLADYF
DVIGGTSTGGLLTAMISTPNENNRPFAAAKEIVPFYFEHGPQIFNPSGQILGPKYDGKYLMQVLQEKLGE
TRVHQALTEVVISSFDIKTNKPVIFTKSNLANSPELDAKMYDISYSTAAAPTYFPPHYFVTNTSNGDEYE
FNLVDGAVATVADPALLSISVATRLAQKDPAFASIRSLNYKKMLLLSLGTGTTSEFDKTYTAKEAATWTA
VHWMLVIQKMTDAASSYMTDYYLSTAFQALDSKNNYLRVQENALTGTTTEMDDASEANMELLVQVGENLL
KKPVSEDNPETYEEALKRFAKLLSDRKKLRANKASY

Figure 4.8 Patatin-17 protein sequence retrieved from Uniprot database.

Each letter corresponds to a single amino acid within the protein sequence.

MKLAEIMTKSRKLKRNLLEISKTEAGQYSVSAPEHKGLVLSGGGAKGISYLGMIQALQERGKIKNLTHVS
GASAGAMTASILAVGMDIKDIKKLIEGLDITKLLDNSGVGFRARGDRFRNILDVIYMMQMKKHLESVQQP
IPPEQQMNYGILKQKIALYEDKLSRAGIVINNVDDIINLTKSVKDLEKLDKALNSIPTELKGAKGEQLEN
PRLTLGDLGRLRELLPEENKHLIKNLSVVVTNQTKHELERYSEDTTPQQSIAQVVQWSGAHPVLFVPGRN
AKGEYIADGGILDNMPEIEGLDREEVLCVKAEAGTAFEDRVNKAKQSAMEAISWFKARMDSLVEATIGGK
WLHATSSVLNREKVYYNIDNMIYINTGEVTTTNTSPTPEQRARAVKNGYDQTMQLLDSHKQTFDHPLMAI
LYIGHDKLKDALIDEKSEKEIFEASAHAQAILHLQEQIVKEMNDGDYSSVQNYLDQIEDILTVDAKMDDI
QKEKAFALCIKQVNFLSEGKLETYLNKVEAEAKAAAEPSWATKILNLLWAPIEWVVSLFKGPAQDFKVEV
QPEPVKVSTSENQETVSNQKDINPAVEYRKIIAEVRREHTDPSPSLQEKERVGLSTTFGGH

Figure 4.9 VipD protein sequence retrieved from Uniprot database.

Each letter corresponds to a single amino acid within the protein sequence.

MHIQSLGATASSLNQEPVETPSQAAHKSASLRQEPSGQGLGVALKSTPGILSGKLPESVSDVRFSSPQGQ
GESRTLTDSAGPRQITLRQFENGVTELQLSRPPLTSLVLSGGGAKGAAYPGAMLALEEKGMLDGIRSMSG
SSAGGITAALLASGMSPAAFKTLSDKMDLISLLDSSNKKLKLFQHISSEIGASLKKGLGNKIGGFSELLL
NVLPRIDSRAEPLERLLRDETRKAVLGQIATHPEVARQPTVAAIASRLQSGSGVTFGDLDRLSAYIPQIK
TLNITGTAMFEGRPQLVVFNASHTPDLEVAQAAHISGSFPGVFQKVSLSDQPYQAGVEWTEFQDGGVMIN
VPVPEMIDKNFDSGPLRRNDNLILEFEGEAGEVAPDRGTRGGALKGWVVGVPALQAREMLQLEGLEELRE
QTVVVPLKSERGDFSGMLGGTLNFTMPDEIKAHLQERLQERVGEHLEKRLQASERHTFASLDEALLALDD
SMLTSVAQQNPEITDGAVAFRQKARDAFTELTVAIVSANGLAGRLKLDEAMRSALQRLDALADTPERLAW
LAAELNHADNVDHQQLLDAMRGQTVQSPVLAAALAEAQRRKVAVIAENIRKEVIFPSLYRPGQPDSNVAL
LRRAEEQLRHATSPAEINQALNDIVDNYSARGFLRFGKPLSSTTVEMAKAWRNKEFT

Figure 4.10 ExoU protein sequence retrieved from Uniprot database.

Each letter corresponds to a single amino acid within the protein sequence.

MRRLLLVLLLLLPLSALAAEARPKIGLVLSGGAARGLAHIGVLKALDEQGIQIDAIAGTSMGAVVGGLYA
SGYTPAELERIALEMDWQQALSDAPPRKDVPFRRKQDDRDFLVKQKISFRDDGTLGLPLGVIQGQNLAMV
LESLLVHTSDNRDFDKLAIPFRAVSTDIATGEKVVFRKGHLPQAIRASMSIPAVFAPVEIDGRLLVDGGM
VDNIPVDVARDMGVDVVIVVDIGNPLRDRKDLSTVLDVMNQSITLMTRKNSEAQLATLKPGDVLIQPPLS
GYGTTDFGRVPQLIDAGYRATTVLAARLAELRKPKDLNSEALDVARTPNQRKPVIDAIRVENNSKVSDEV
IRHYIRQPLGTRLDLGRLQDDMSTLYGLDYFDQVQYRVVKEKKLNTLVIHATGKKGGTDFLRLGLNLSDD
MRGESTFNLGGSYRMNGLNRLGAEWLTRVQLGDRQELYSEFYQPLDVGSRYFVAPFLFHEAQNVDVTEDN
DPLLRYRLERYGYGLNVGRQIANNGEIRLGAVQAYGKADVRIGDPSLPDIDFTEGYYELKYSFDTVDDVN
FPHEGEEIGLTMRRYDKSLGSDDSYRQWDLRLNKALSFGADTWVFGGGYGRTLDDAEVVTSSFTLGGARE
LSGFRRDALSGQNYSLGRIVYYRRLTERSFLPLDFPLYLGGSIERGRIWNNDNEYDSGYINAASLMIGFD
TPLGPLTFSYGINDENFKAFYLNLGQNF

Figure 4.11 PlpD protein sequence retrieved from Uniprot database.

Each letter corresponds to a single amino acid within the protein sequence.

Models 1-4 were generated using SWISS-MODEL.[381–383,400]

Model 1: Homology model constructed using the crystal structure of patatin-17 from *Solanum cardiophyllum* as a template (PDB Id: 1OXW; homo-3-mer - resolution 2.2A).[401]

Model 2: Homology model constructed using the crystal structure of VipD from *Legionella pneumophila* as a template (PDB Id: 4AKF; homo-2-mer - resolution 2.9A).[402]

Model 3: Homology model constructed using ExoU from *Pseudomonas aeruginosa* as a template (PDB Id: 3TU3; hetero-2-mer complexed SpcU - resolution 1.92A).[403]

Model 4: Homology model constructed using VipD from *Legionella pneumophila* as a template (PDB Id: 4KYI; hetero-2-mer complexed human GTPase Rab5 - resolution 3.08A).[404]

C-terminal domain modelling was attempted using SWISS-MODEL; however, no templates of sufficient quality were found which allowed modelling of more than a short 30 residue fold (results not shown).

Models 5-9 were generated using the I-TASSER program suite.[203,389,405]

Models 5-8 were run with default settings on the web server, while model 9 was generated with a manually set template.

Model 5: Homology model constructed based on PNPLA3 full length sequence. (highest similarity to ExoU from *Pseudomonas aeruginosa* (PDB Id: 4AKXB; hetero-4-mer complexed ExoU with SpcU - resolution 2.94A).[406]

Model 6: Homology model constructed based on PNPLA3 N-terminal 239 residues. (highest similarity to PlpD from *Pseudomonas aeruginosa* (PDB Id: 5FYAA; homo-2-mer - resolution 2.14A).[407]

Model 7: Homology model constructed based on PNPLA3 N-terminal 179 residues. (highest similarity to native patatin from *Solanum cardiophyllum* (PDB Id: 4PK9A; monomer - resolution 1.96A).[408]

Model 8: Homology model constructed based on PNPLA3 C-terminal 242 residues. (highest similarity to ExoU from *Pseudomonas aeruginosa* (PDB Id: 3TU3B; hetero-2-mer complexed SpcU - resolution 1.92A).[403]

Model 9: Homology model constructed based on PNPLA3 full length sequence. The template was strictly set to patatin-17 from *Solanum cardiophyllum* as a template (PDB Id: 1OXW; homo-3-mer - resolution 2.2A).[401]

4.4.2 Model accuracy assessment

Each model was assessed initially based on the direct quality predictions of the modelling software used (SWISS-MODEL or I-TASSER). These statistics incorporate an assessment of confidence based on homology and threading alignments. Following this, model accuracy was evaluated based on structural quality using the PROtein Structure Evaluation Suite and Server (PROSESS).[365]

Secondary structure was compared with predicted secondary structure using Phyre2. [374]

4.4.3 Image generation

Images were generated from PDB files using the PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC. and the visual molecular dynamics (VMD) software suite.[409]

Two-dimensional Richardson diagrams were generated using Pro-origami,[410] using the following the following settings: 1) Secondary structure was selected using STRIDE. 2) A distance matrix used rather than heuristics for placing helices. 3) Helices were prevented from being placed between neighbouring sheets with separate numbering for helices and strands.[410]

4.5 Results

4.5.1 Template assessment

The templates used to generate models of PNPLA3 were all of low homology, below 15% sequence identity. The novel bacterial alignments improved on the patatin-17 previously used for modelling of PNPLA3, increasing the coverage from 66.7% to 92.7% and identity from 10.3% to 14.1% (Table 4.1; Figures 4.10 – 4.13).

Table 4.1 Alignment statistics of template protein sequences and PNPLA3

Protein	Coverage (%)	Identity (%)
patatin-17	66.7	10.3
VipD	89.0	13.9
ExoU*	92.7	14.1
PlpD	89.0	11.5

*The protein with the highest coverage and identity with PNPLA3

Each template was similar, deriving from the Patatin-like superfamily of proteins, however, were sufficiently different to achieve unique alignments with a shared core of only 13 residues in length (Table 4.2). While the shared core is highly conserved, the overall TM-score between the templates was only 0.447 a value expected of distantly related proteins within a protein family.

Visual alignment of the templates demonstrates a similar patatin region shared between proteins, consistent with the pairwise RMSD difference of only 2.02A (Figure 4.14). The C-terminus varies more widely.

Figure 4.10 PNPLA3 and Patatin-17 sequence alignment.

The top row represents the PNPLA3 amino acid sequence and the middle row the Patatin-17 sequence, with numbers at the end of the row designating the position of the amino acid in the respective protein. Each letter corresponds to a single amino acid within the protein sequence and gaps in the alignment represented with dashes (-). The bottom row denotes the degree of conservation between the sequences in which:

*** (asterisk)** represents an identical conserved amino acid.
: (colon) represents conservation of a residue with strongly similar properties.
. (period) represents conservation of a residue with weakly similar properties.

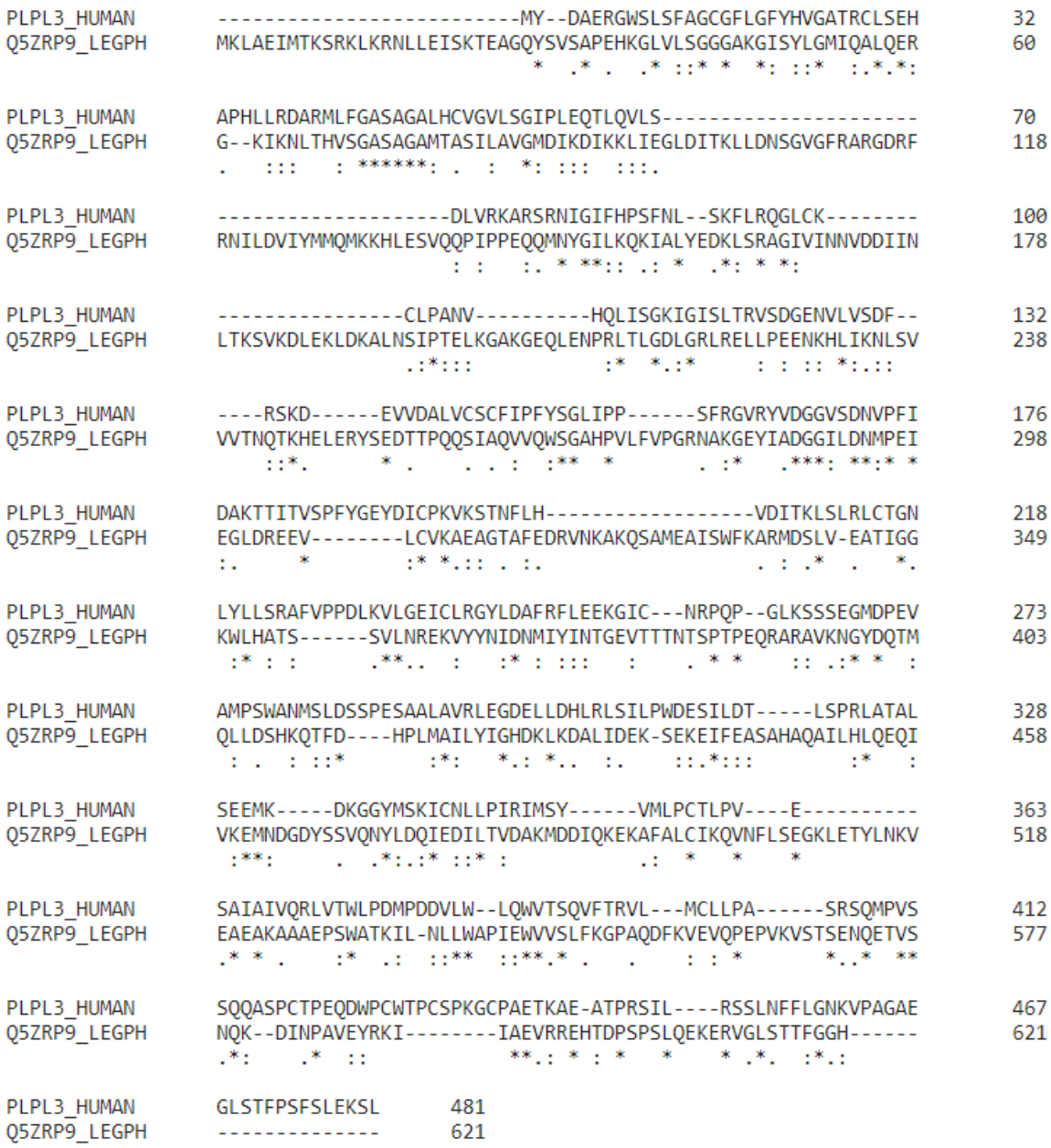


Figure 4.11 PNPLA3 and VipD sequence alignment.

The top row represents the PNPLA3 amino acid sequence and the middle row the VipD sequence, with numbers at the end of the row designating the position of the amino acid in the respective protein. Each letter corresponds to a single amino acid within the protein sequence and gaps in the alignment represented with dashes (-). The bottom row denotes the degree of conservation between the sequences in which:

*** (asterisk)** represents an identical conserved amino acid.
: (colon) represents conservation of a residue with strongly similar properties.
. (period) represents conservation of a residue with weakly similar properties.

```
PLPL3_HUMAN       ------------------------------------------------------------        0
O34208_PSEAI      MHIQSLGATASSLNQEPVETPSQAAHKSASLRQEPSGQGLGVALKSTPGILSGKLPESVS        60


PLPL3_HUMAN       -----------------------------------MYDAERGWSLSFAGCGFLGFYHV        23
O34208_PSEAI      DVRFSSPQGQGESRTLTDSAGPRQITLRQFENGVTELQLSRPPLTSLVLSGGGAKGAAYP        120
                                                    .     ** ::* *  *  :

PLPL3_HUMAN       GATRCLSEHAPHLLRDARMLFGASAGALHCVGVLSGIPLEQTLQVLSD---LVRKA--RS        78
O34208_PSEAI      GAMLALE--EKGMLDGIRSMSGSSAGGITAALLASGMSPAA-FKTLSDKMDLISLLDSSN        177
                  **  .*.     :* . * : *:***.: .. : **:     ::.***   *:      .

PLPL3_HUMAN       RNIGIFH-PSFNLSKFLRQGLCKCL-------------------------PANVHQLISG        112
O34208_PSEAI      KKLKLFQHISSEIGASLKKGLGNKIGGFSELLLNVLPRIDSRAEPLERLLRDETRKAVLG        237
                  ::: :*:  * ::.  *::** : :                              :.:: : *

PLPL3_HUMAN       KIG------------ISLTRVSDGENVLVSDFRSKDEVVDAL------------------        142
O34208_PSEAI      QIATHPEVARQPTVAAIASRLQSGSGVTFGDLDRLSAYIPQIKTLNITGTAMFEGRPQLV        297
                  :*.               :*:..*..* ..*:   .  :  :

PLPL3_HUMAN       ---------VCSCFIPFYSGLIPPSFRG--------------VRYVDGGVSDNVPFIDAK        179
O34208_PSEAI      VFNASHTPDLEVAQAAHISGSFPGVFQKVSLSDQPYQAGVEWTEFQDGGVMINVPVPEMI        357
                           :  .   . ** :*  *:               ..: ****  ***. :

PLPL3_HUMAN       TTITVSPFYGEYDICPKVKSTNFLHVDITKLSLRLCTGNLYLLSRAFVPPDLKVLGEICL        239
O34208_PSEAI      D--------KNFDSGPLRRNDNLIL-E-----FEG--------EAGEVAPDRGTRG-GAL        394
                            ::*  *  :. *::  :     :.         . . * **  . *  .*

PLPL3_HUMAN       RGYLDAFRFLE------EKGICNR-------P---QPGLKSSSEGMDPEVAMPSWANMSL        283
O34208_PSEAI      KGWVVGVPALQAREMLQLEGLEELREQTVVVPLKSERGDFSGMLGGTLNFTMPDEIKAHL        454
                  :*:: ..  *:       :*: :        *   : *  *.  *   :.:**.  :  *

PLPL3_HUMAN       DSSPE---SAALAVRLEGD--ELLDHLRLSILPWDESILDTLSPRLATALSEEMKDK-GG        337
O34208_PSEAI      QERLQERVGEHLEKRLQASERHTFASLDEALLALDDSMLTSVAQQNP-----EITDGAVA        509
                  :.  :   .  *  **:..  . :  *  ::*  *:*:* ::: :       *:.*   .

PLPL3_HUMAN       YMSK---ICNLLPIRIMSYVMLPCTLPVESAIAIVQRLVTWLPDMPDDVLWLQWVTSQVF        394
O34208_PSEAI      FRQKARDAFTELTVAIVSANGLAGRLKLDEAMRSALQRLDALADTPERLAWLAAELNHAD        569
                  : .*     . * : *:*   *   * ::.*:  . : :  * * *: : **    .:.

PLPL3_HUMAN       TRVLMCLLPASRSQ---MPVSSQQA------------SPCTPEQDWPC-WTPCSPKGCPA        438
O34208_PSEAI      NVDHQQLLDAMRGQTVQSPVLAAALAEAQRRKVAVIAENIRKEVIFPSLYRPGQPDSNVA        629
                  .     ** * *.*    ** :               .    *  :*. : * .*..  *

PLPL3_HUMAN       E-------TKAEATPRSILRSSLNFFLGNKVPAGAEGLSTFPSF------SLEKSL----        481
O34208_PSEAI      LLRRAEEQLRHATSPAE-INQALNDIVDNYS---ARGFLRFGKPLSSTTVEMAKAWRNKE        685
                           :  ::* . :..:** ::.*     *.*:  * .       .: *:

PLPL3_HUMAN       --  481
O34208_PSEAI      FT  687
```

Figure 4.12 PNPLA3 and ExoU sequence alignment.

The top row represents the PNPLA3 amino acid sequence and the middle row the ExoU sequence, with numbers at the end of the row designating the position of the amino acid in the respective protein. Each letter corresponds to a single amino acid within the protein sequence and gaps in the alignment represented with dashes (-). The bottom row denotes the degree of conservation between the sequences in which:

*** (asterisk)** represents an identical conserved amino acid.
: (colon) represents conservation of a residue with strongly similar properties.
. (period) represents conservation of a residue with weakly similar properties.

```
PLPL3_HUMAN       -----------------MYDAERGWSLSFAGCGFLGFYHVGATRCLSEHAPHLLRDARML       43
Q9HYQ6_PSEAE      MRRLLLVLLLLLPLSALAAEARPKIGLVLSGGAARGLAHIGVLKALDEQGIQIDAIAGTS       60
                                :*.   .* ::* .  *: *:*. :.*.*:. ::   *

PLPL3_HUMAN       FGASAGALHCVGVLS----GIPLEQTLQ-VLSDLVRKARSRNIGIFHPSFNLSKFLRQGL       98
Q9HYQ6_PSEAE      MGAVVGGLYASGYTPAELERIALEMDWQQALSDA----PPRKDVPFRRKQDDRDFLVKQ-       115
                  :** .*.*:. *        * **   * .***       *:   *: . :  .** :

PLPL3_HUMAN       CKCLPANVHQLISGKIGISLT-----------------------------------RVS       122
Q9HYQ6_PSEAE      ------KISFRDDGTLGLPLGVIQGQNLAMVLESLLVHTSDNRDFDKLAIPFRAVSTDIA       169
                        ::    .*.:*: *                                      ::

PLPL3_HUMAN       DGENVLVSDFRSKDEVVDALVCSCFIPFYSGLIPPSFRGVRYVDGGVSDNVPFIDAKTTI       182
Q9HYQ6_PSEAE      TGEKVVFR----KGHLPQAIRASMSI--PAVFAPVEIDGRLLVDGGMVDNIPVDVARDMG       223
                   **:*:.     *..: :*: .*  *   : : * .: *   ****: **:*.  *:

PLPL3_HUMAN       TV--------SPFYGEYDICPKVKSTNFLHVDITKLSLRLCTGNLYLL--SRAFVPPDLK       232
Q9HYQ6_PSEAE      VDVVIVVDIGNPLRDRKDLSTVLD---VMNQSITLMTRKNSEAQLATLKPGDVLIQPPLS       280
                  .         .*: .. *:.  :.   .:: .** :: : . .:*  *  . .:: * *.

PLPL3_HUMAN       VLGE--------ICLRGYLDAFRFLEEKGICNRPQPGLKSSS--EGMDPEVAMPSWANMS       282
Q9HYQ6_PSEAE      GYGTTDFGRVPQLIDAGYRATTVLAARLAELRKPK-DLNSEALDVARTPNQRKPVIDAIR       339
                    *         :   **  :  :  . .  .:*: .*:*.:   .  *:   *    :

PLPL3_HUMAN       LDSSPES----------AALAVRLE------------GDELLDHLRLSILPWDESILDTL       320
Q9HYQ6_PSEAE      VENNSKVSDEVIRHYIRQPLGTRLDLGRLQDDMSTLYGLDYFDQVQYRVVK--EKKLNTL       397
                  ::.. :             *..**:            * : :*:::  ::   *. *:**

PLPL3_HUMAN       ------------SPRLATALSEEMKDKGGYMSKICNLLPIRIMSYVMLPCTLPVESAIAI       368
Q9HYQ6_PSEAE      VIHATGKKGGTDFLRLGLNLSDDMRGESTFNLG---------GSYRM-------N---GL       438
                               **.  **::*:.:. :            ** *       :   .:

PLPL3_HUMAN       VQRLVTWLPDMPD---DVLWLQWVTS-QVFTRVLMCLLP-ASRSQMPVSSQQASPCTPEQ       423
Q9HYQ6_PSEAE      NRLGAEWLTRVQLGDRQELYSEFYQPLDVGSRYFVAPFLFHEAQNVDVTE-DNDP---LL       494
                   :  . **  :     : *: ::    :* :* ::. :   . .:: *:. : .*

PLPL3_HUMAN       DWPCWTPCSPKGCPAETKAEAT-PRSILRSSLNFFLGNKVPAGAEGLSTFPSFSLEKSL-       481
Q9HYQ6_PSEAE      RYR----LERYGYGLNVGRQIANNGEIRLGAVQAYGKADVRIG---DPSLPDIDFTEGYY       547
                   :      .  *   :.  : :   .*  .::: :   .*  *     ::*.:.: :.

PLPL3_HUMAN       ------------------------------------------------------------       481
Q9HYQ6_PSEAE      ELKYSFDTVDDVNFPHEGEEIGLTMRRYDKSLGSDDSYRQWDLRLNKALSFGADTWVFGG       607

PLPL3_HUMAN       ------------------------------------------------------------       481
Q9HYQ6_PSEAE      GYGRTLDDAEVVTSSFTLGGARELSGFRRDALSGQNYSLGRIVYYRRLTERSFLPLDFPL       667

PLPL3_HUMAN       ------------------------------------------------------------       481
Q9HYQ6_PSEAE      YLGGSIERGRIWNNDNEYDSGYINAASLMIGFDTPLGPLTFSYGINDENFKAFYLNLGQN       727

PLPL3_HUMAN       -   481
Q9HYQ6_PSEAE      F   728
```

Figure 4.13 PNPLA3 and PlpD sequence alignment.

The top row represents the PNPLA3 amino acid sequence and the middle row the PlpD sequence, with numbers at the end of the row designating the position of the amino acid in the respective protein. Each letter corresponds to a single amino acid within the protein sequence and gaps in the alignment represented with dashes (-). The bottom row denotes the degree of conservation between the sequences in which:

*** (asterisk)** represents an identical conserved amino acid.
: (colon) represents conservation of a residue with strongly similar properties.
. (period) represents conservation of a residue with weakly similar properties.

Table 4.2 Summary of the template similarity assessment using mTM-align

Metrics	Value
L_{core}	13
ccRMSD	1.78
ccTM-score	0.065
L_{ali}	132
RMSD	2.02
TM-score	0.447

Analysis included protein template structures: 1oxw 3tu32 4akf 4akx 4kyi 4pk9 5fya.
Lcore: length of common core;
ccRMSD: average pairwise RMSD in common core;
ccTM-Score: Average pairwise TM-score in common core;
Lall: Average pairwise length;
RMSD: Average pairwise RMSD;
TM-score: average pairwise TM-Score;

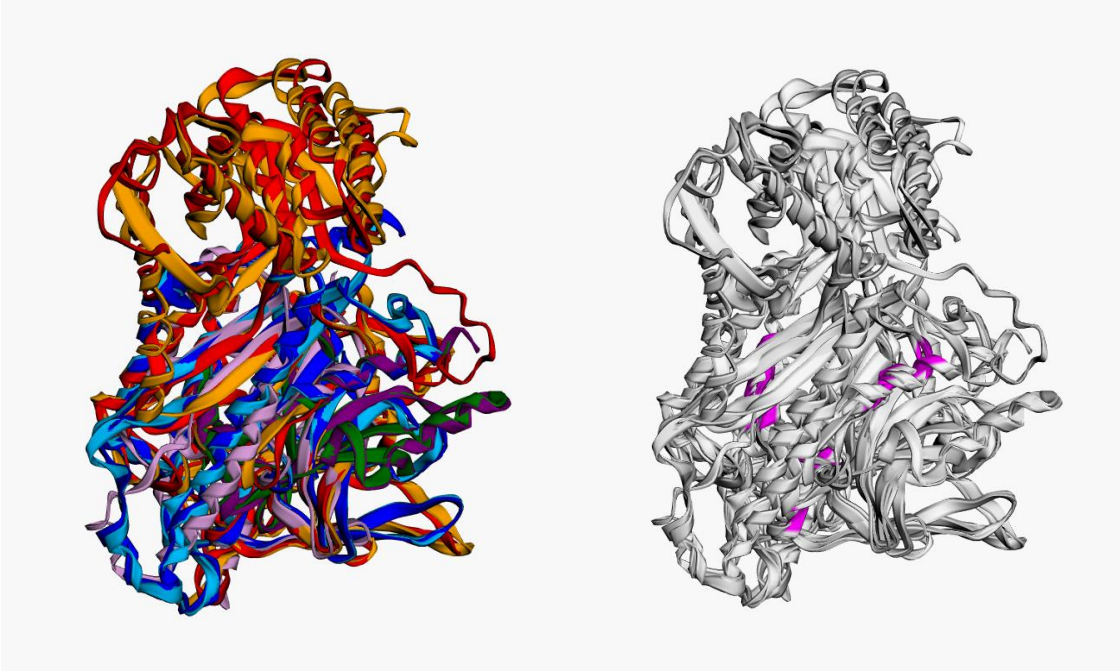


Figure 4.14 Superimposed alignment of modelling template structures

Analysis included protein template structures: 1oxw 3tu32 4akf 4akx 4kyi 4pk9 5fya
Left: overlapping templates each coloured individually.
Right: overlapping structures with common core highlighted in magenta.

4.5.1 Secondary structure composition of the PNPLA3 models

Nine structurally distinct models were successfully created with SWISS-MODEL and I-TASSER. Five models span the initial 179 residues (models 1, 2, 3, 4 and 7); two models span the full length of the protein (models 5 and 9); one model spans an extended N-terminal range to residue 239 (model 6); one model spans the C-terminal residues 239-481 (model 8); (Table 4.3; Figures 4.15 - 4.23).

Table 4.3 length of predicted PNPLA3 models

Models	Software	Residue range
1	SWISS-MODEL	5-179
2	SWISS-MODEL	6-179
3	SWISS-MODEL	6-178
4	SWISS-MODEL	7-179
5	I-TASSER	1-481
6	I-TASSER	1-239
7	I-TASSER	1-179
8	I-TASSER	239-481
9	I-TASSER	1-481

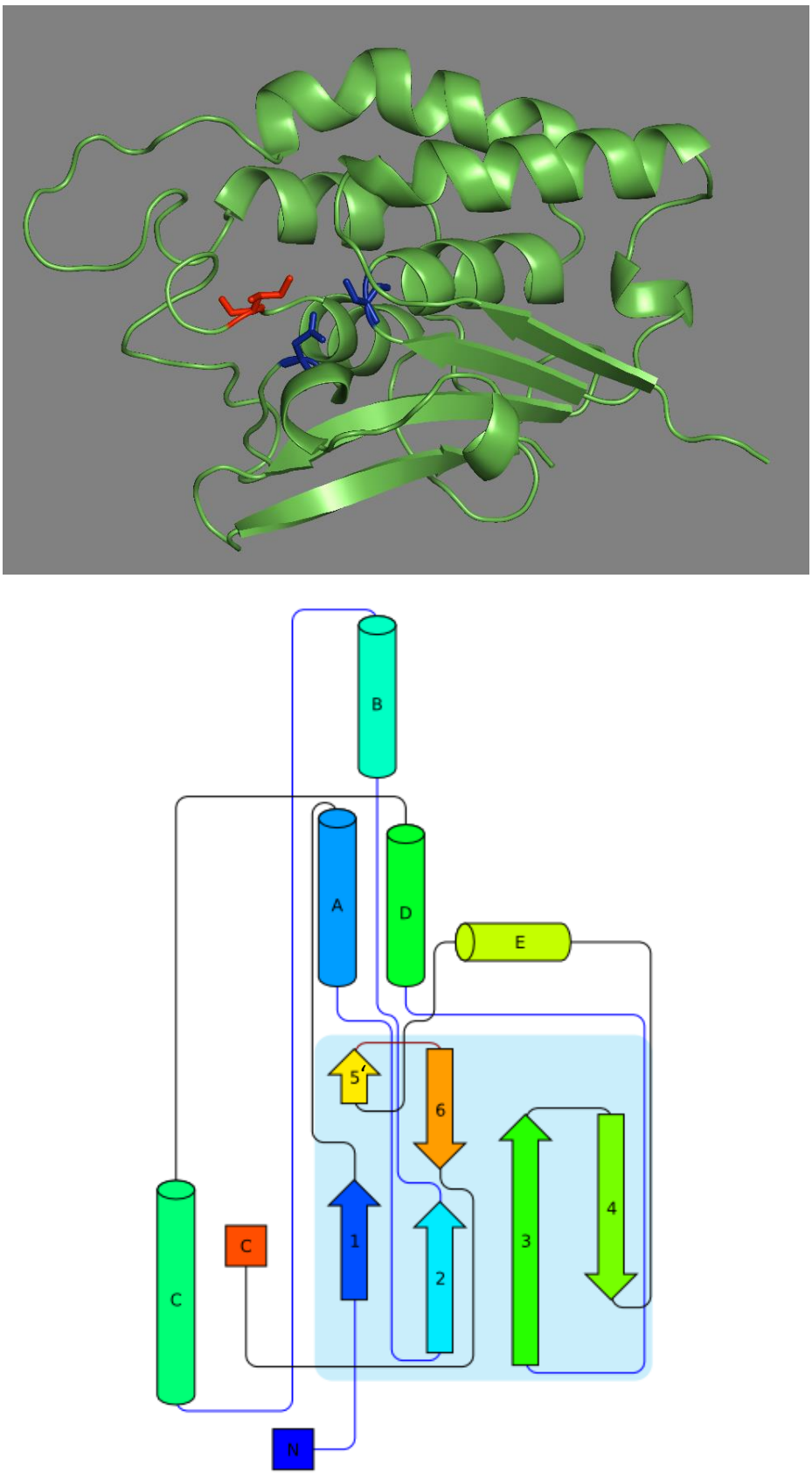


Figure 4.15 Three-dimensional structure of Model 1

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

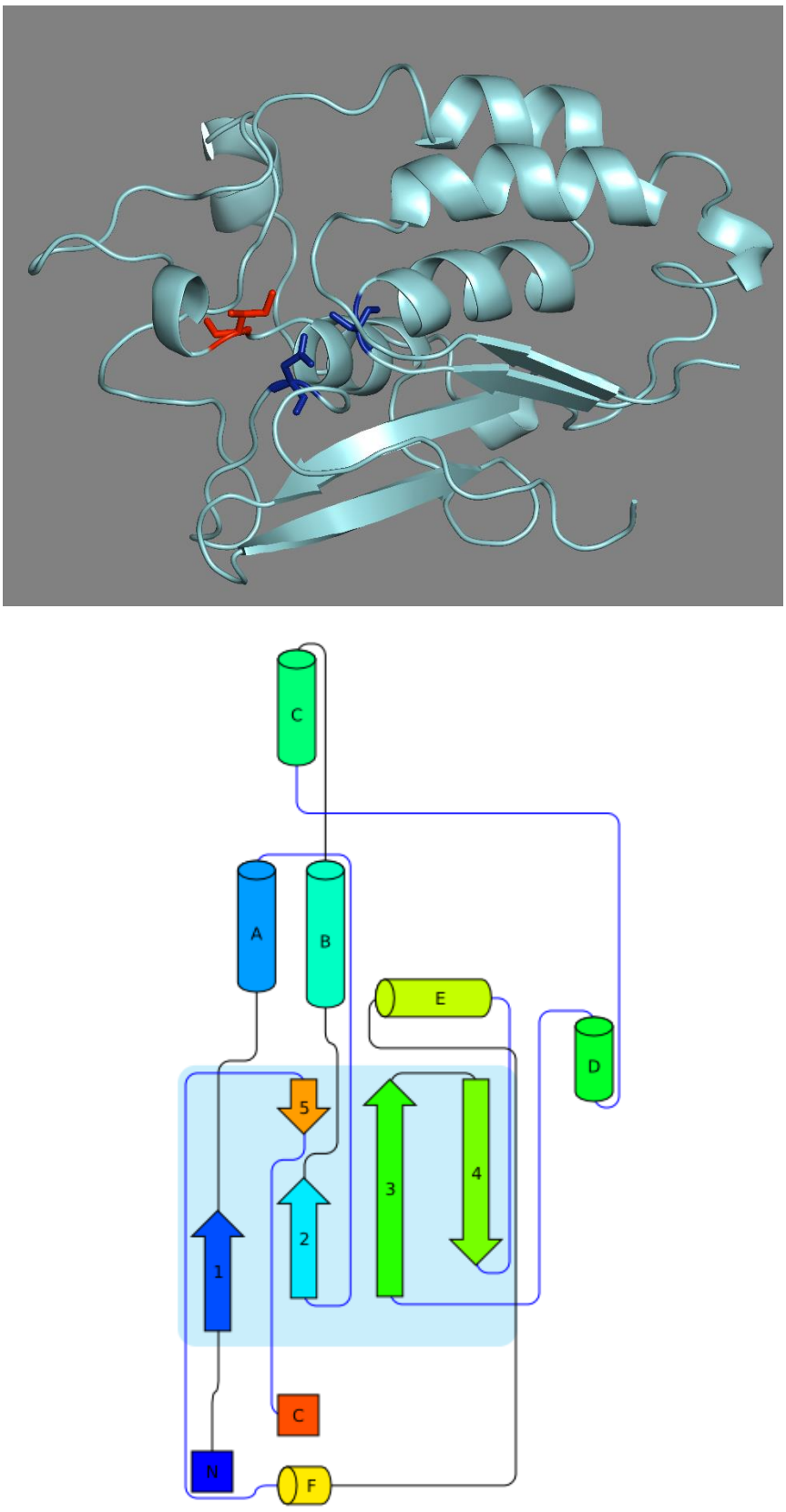


Figure 4.16 Three-dimensional structure of Model 2

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

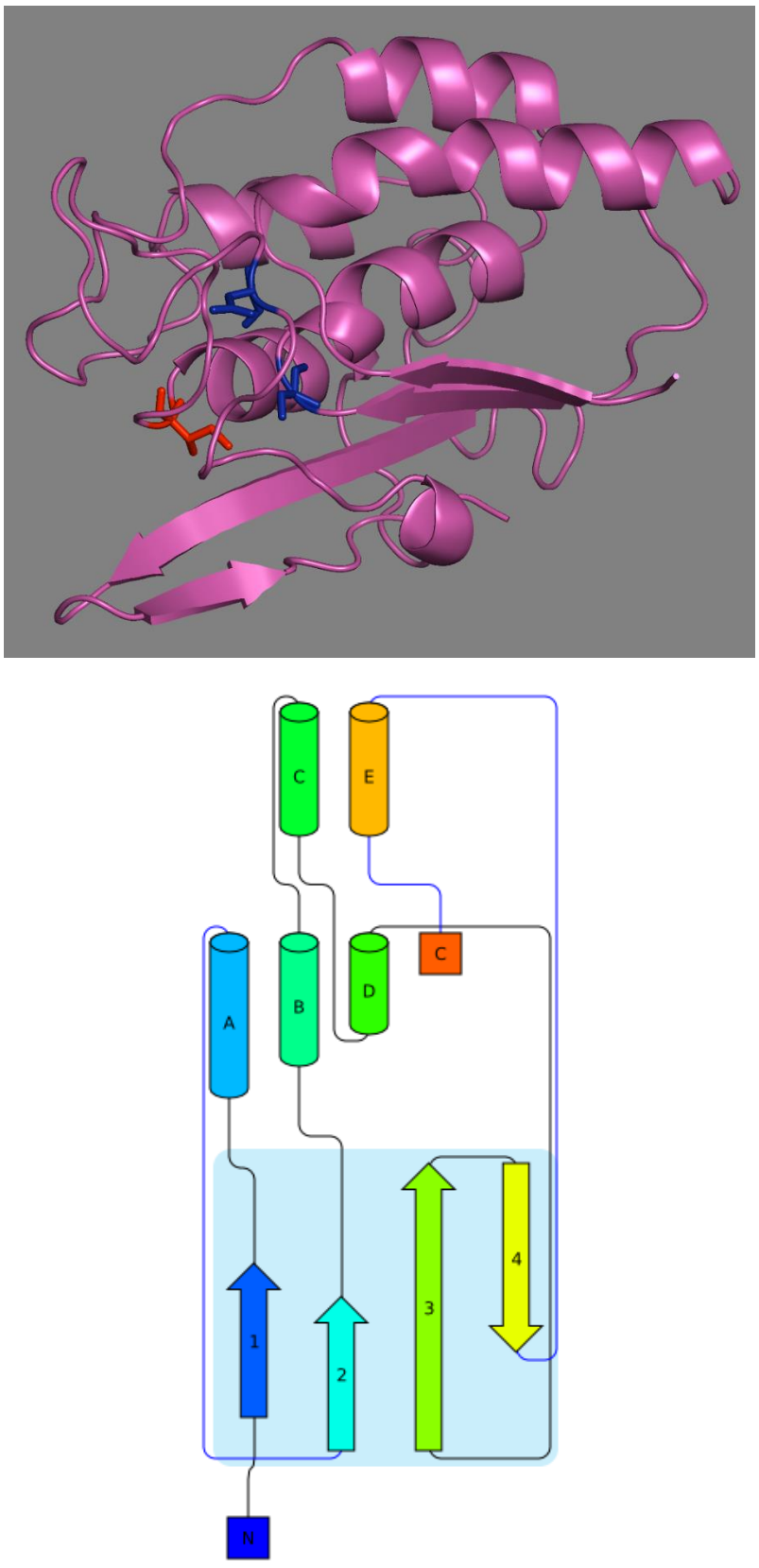


Figure 4.17 Three-dimensional structure of Model 3

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

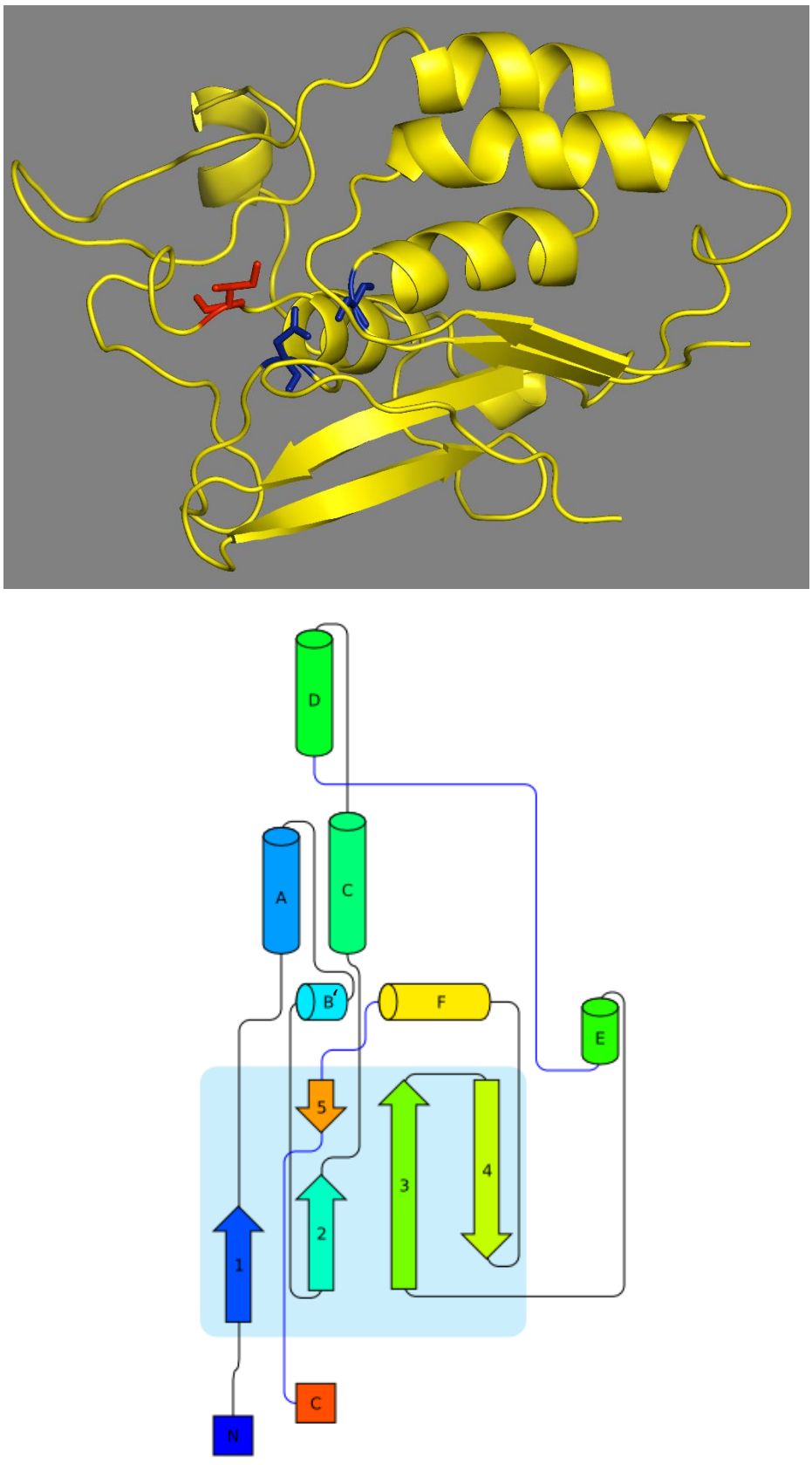


Figure 4.18 Three-dimensional structure of Model 4

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

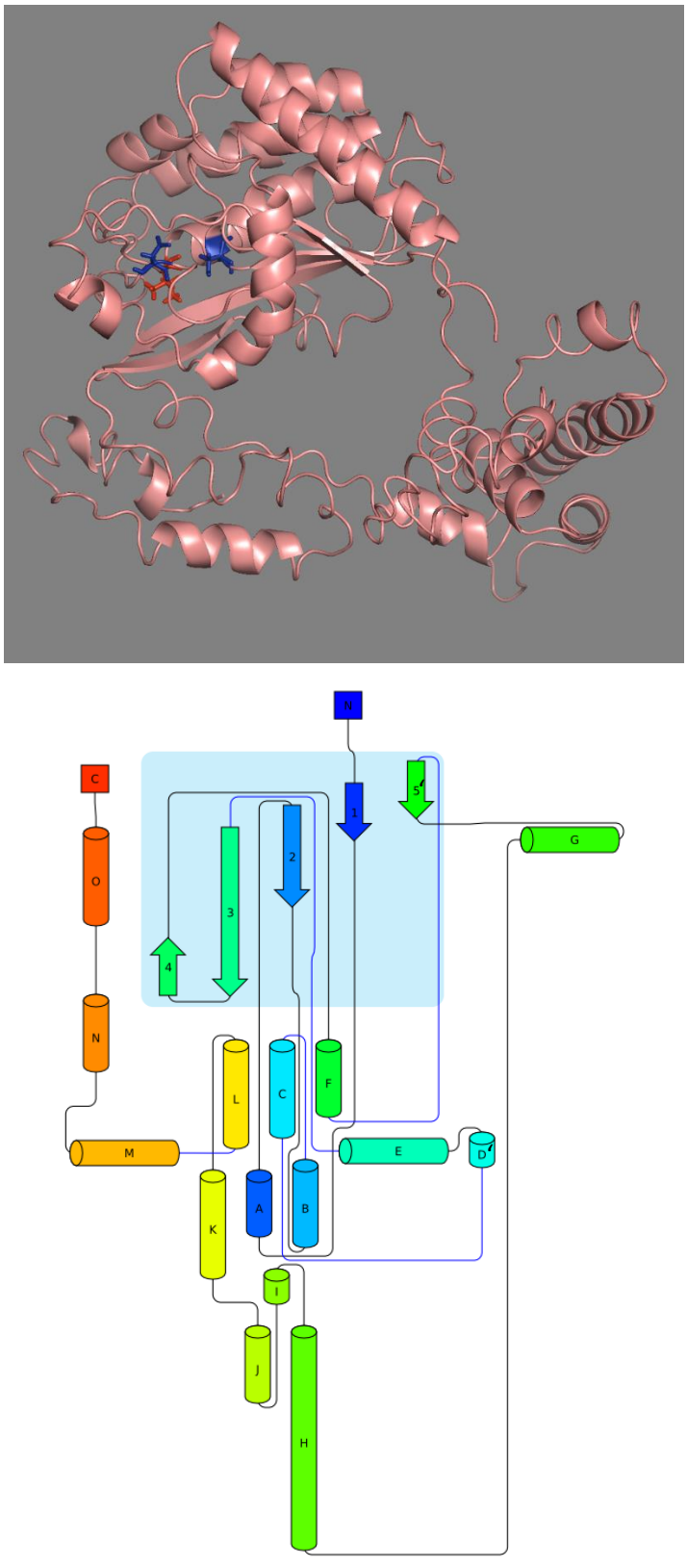


Figure 4.19 Three-dimensional structure of Model 5

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

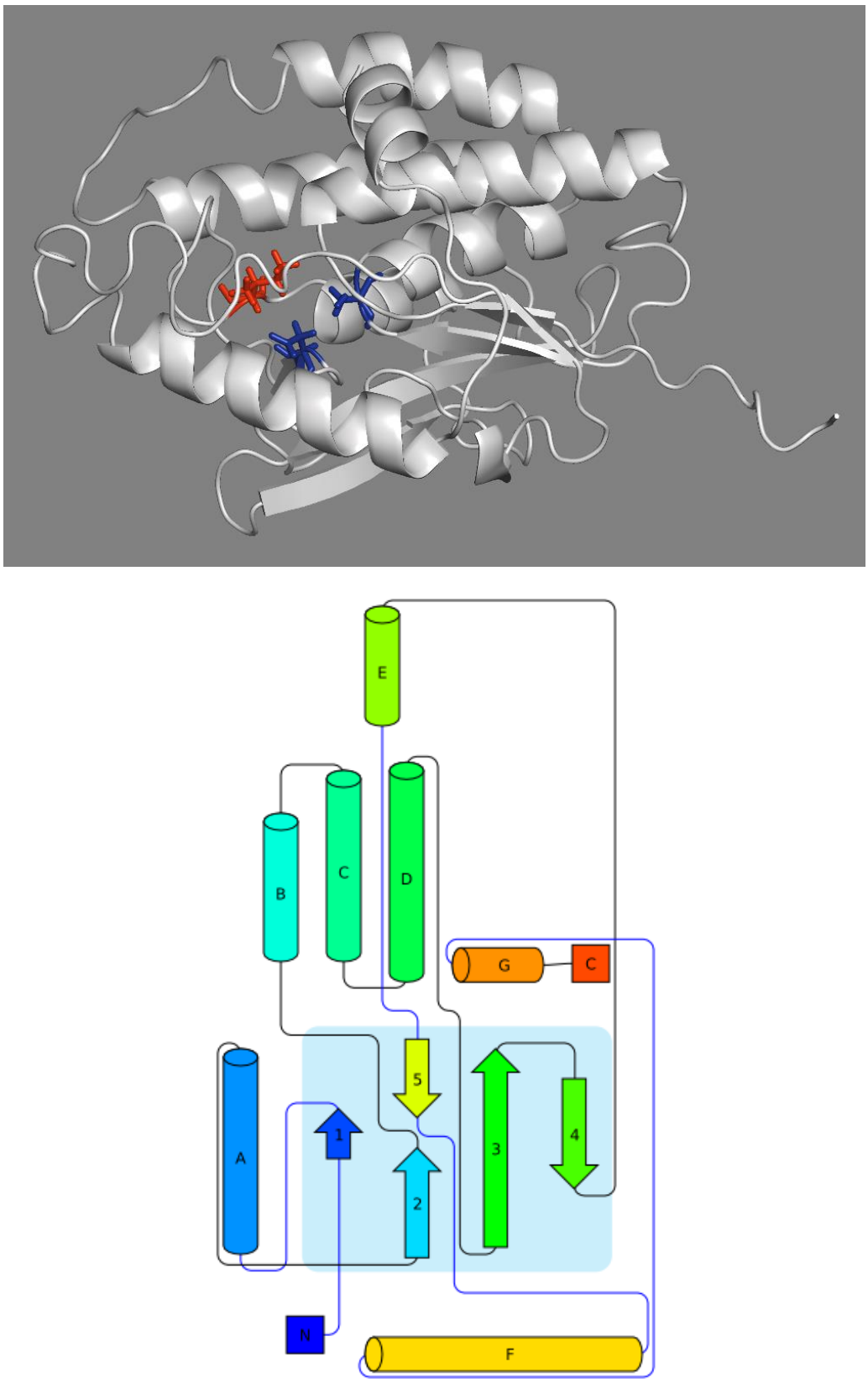


Figure 4.20 Three-dimensional structure of Model 6

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

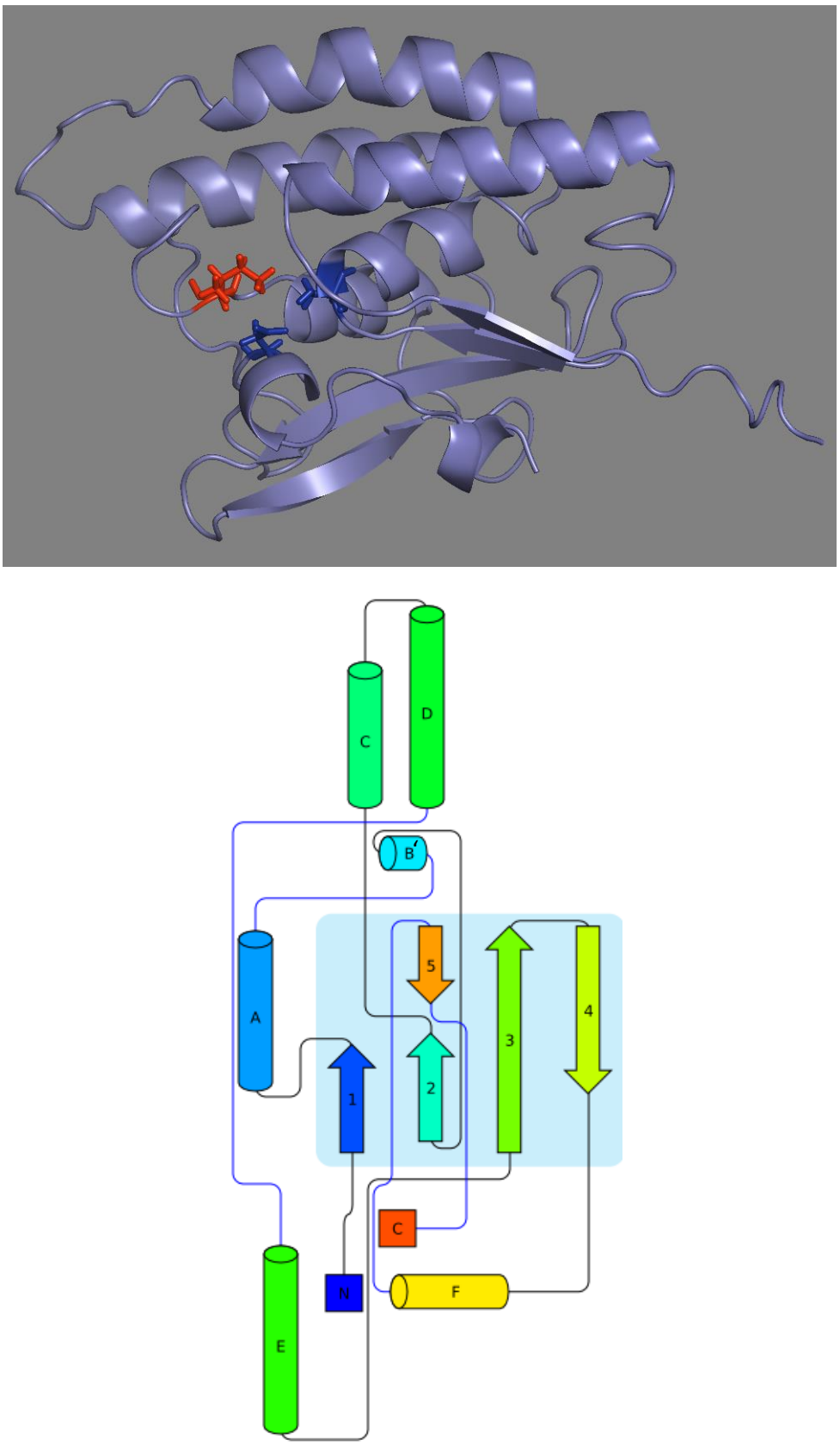


Figure 4.21 Three-dimensional structure of Model 7

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

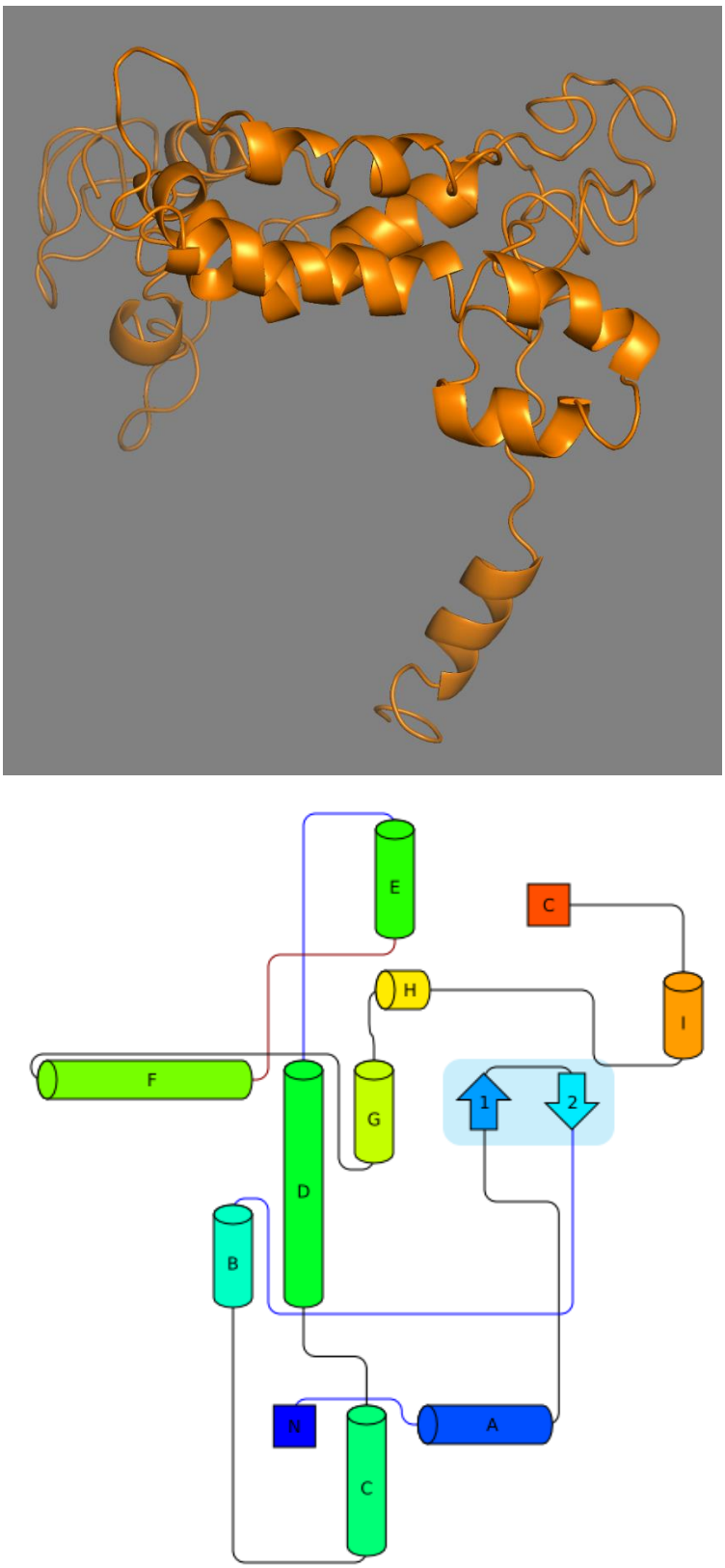


Figure 4.22 Three-dimensional structure of Model 8

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

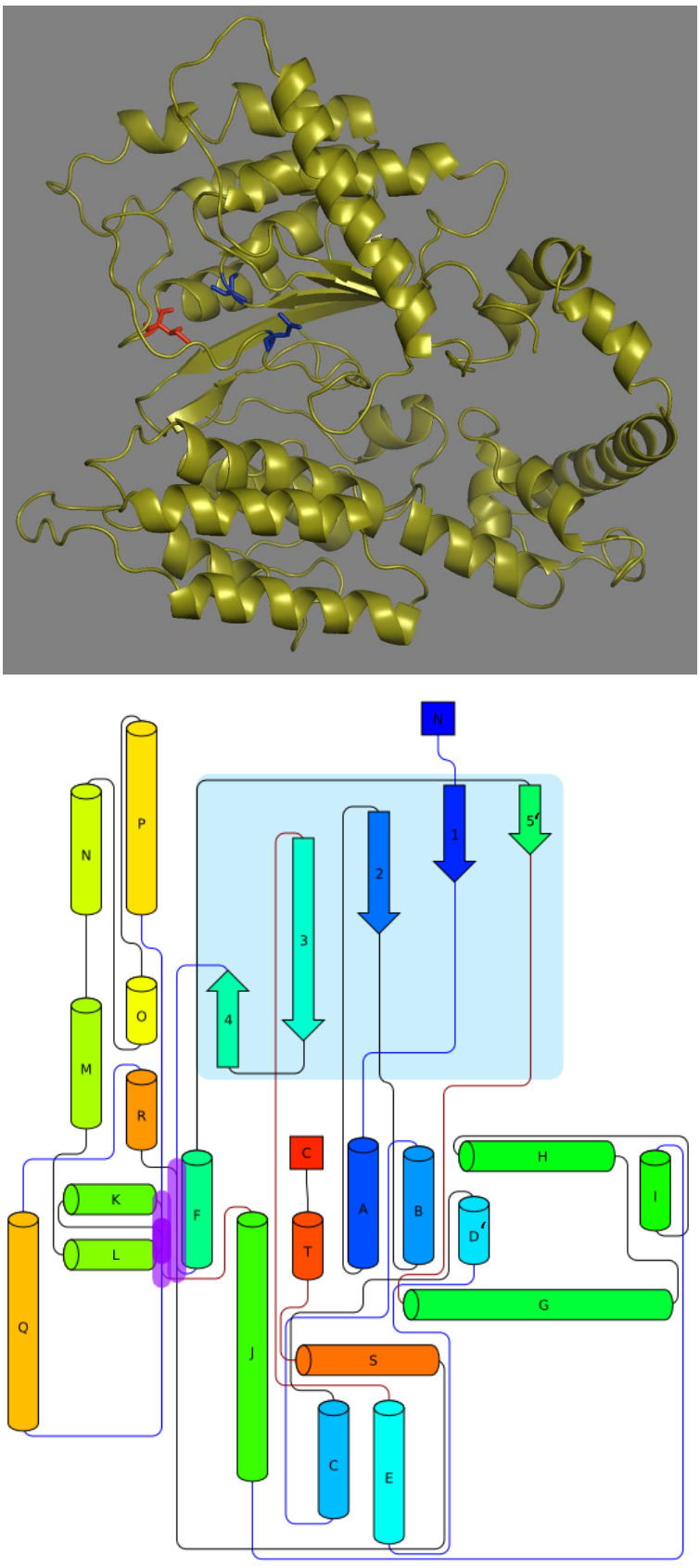


Figure 4.23 Three-dimensional structure of Model 9

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

Models spanning only the N-terminal patatin domain (1, 2, 3, 4 and 7), exhibit between 3 to 15% more α-helix compared with β-strand in the structure, suggesting that the most structured part of these models is the N-terminal (Table 4.4). While the most significant secondary structure in all the models is random coil.

The full-length models (5 and 9) also have a significantly lower proportion of β-strands, suggesting the C-terminal region consists of mainly α-helix and coil with little β-strand structure. This is supported by the fact that the model of the C-terminal (8) has the highest level of random coil (69%) and the lowest β-strand (1%), although the α-helix level remains roughly the same at 30%.

Table 4.4 Secondary structure of the PNPLA3 homology models

Model	α-helix (%)	β-strand (%)	Turn (%)	Coil (%)	Protein length
1	36	26	27	38	175
2	25	24	27	51	174
3	30	18	20	52	173
4	27	24	23	49	173
5	32	9	14	59	481
6	37	19	10	44	239
7	38	23	13	39	179
8	30	1	16	69	242
9	50	7	13	43	481
1OXW	41	21	24	38	360
1CJY	33	26	22	41	633

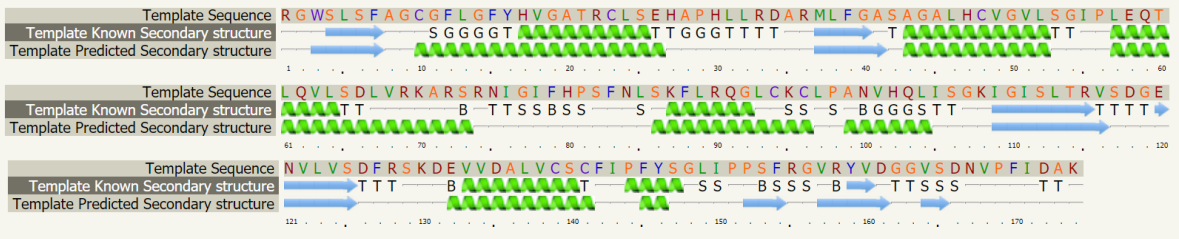
*10XW and 1CJY correspond to the PDBid of the crystal structures of pat-17 and cPLA2 respectively.

All models have secondary structures which support predictions, with almost all key secondary structural elements occurring within a three-residue window (Figure 4.24). Where the secondary structures are not present, similar structures remain; for example, predicted α-helices being replaced by 3-turn helices within the model.

The exception to this occurs in the C-terminal of the protein; where a predicted β-strand spanning residues 203-215 does not occur in either full length model (5 and 9) suggesting this low confidence region is poorly modelled and should be treated with caution.

Model 1

Model 2



Model 3

Model 4

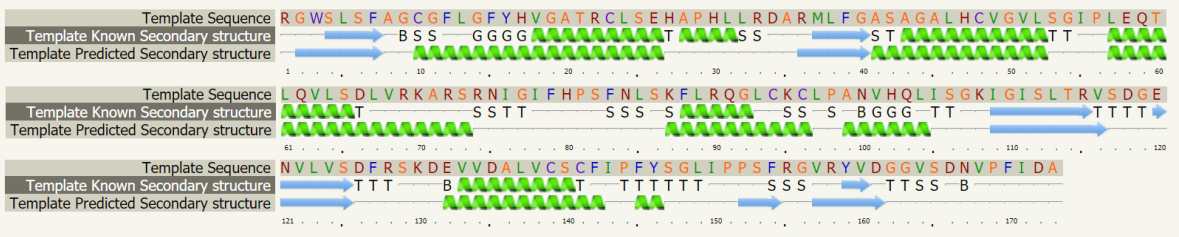
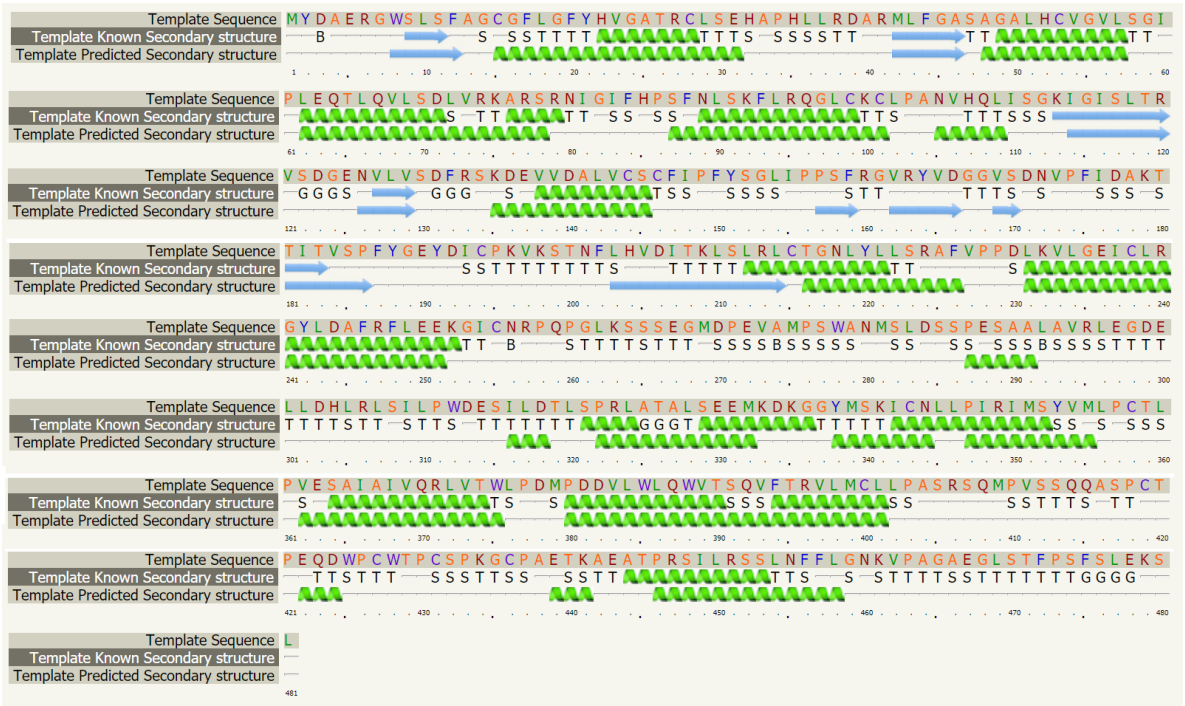


Figure 4.24 Secondary structure against predicted secondary structure of models (Cont'd next page)

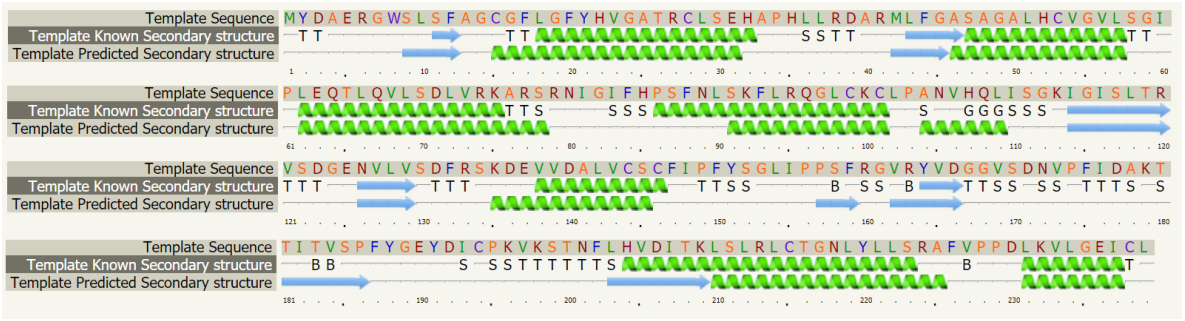
For each model the sequence is represented along the top row with single letter abbreviations representing each amino acid in the protein chain. The "known" template in the figure refers to the respective molecular model and predicted secondary structures of the region are shown below, with the β-sheets represented as blue arrows; the α-helices are represented as green helices. Additional assigned secondary structures by the program DSSP are represented by single letters as follows:

G: 3-turn helix (310 helix);
I: 5-turn helix (π helix);
T: hydrogen bonded turn;
B: residue in isolated β-bridge;
S: bend.

Model 5



Model 6



Model 7

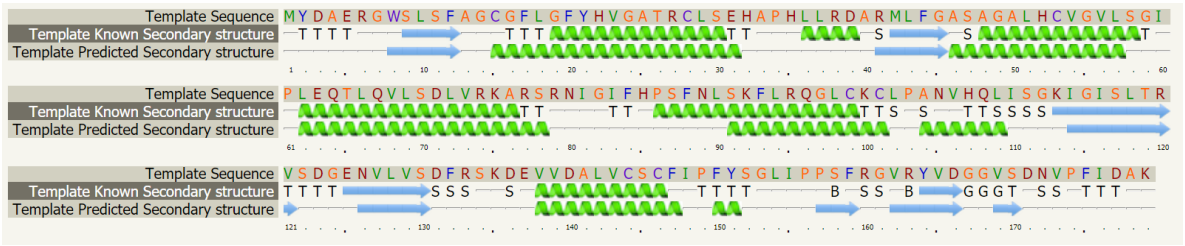
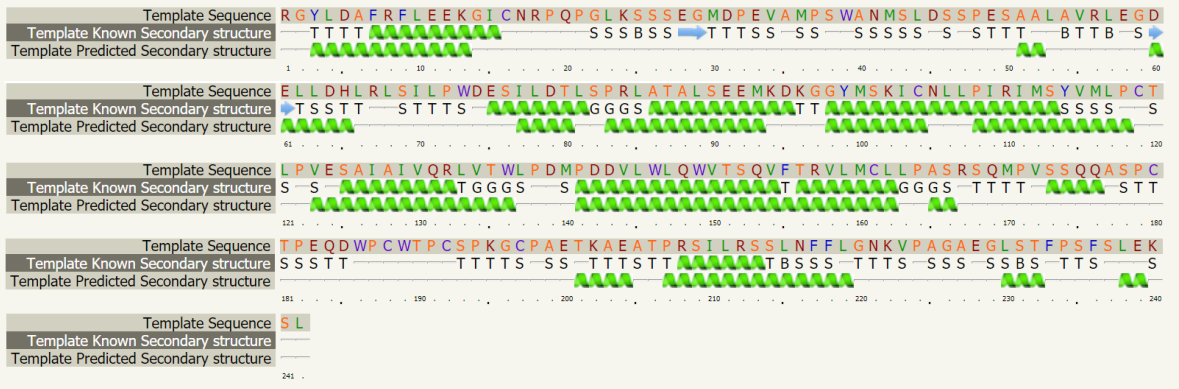


Figure 4.24 (continued)

Model 8



Model 9

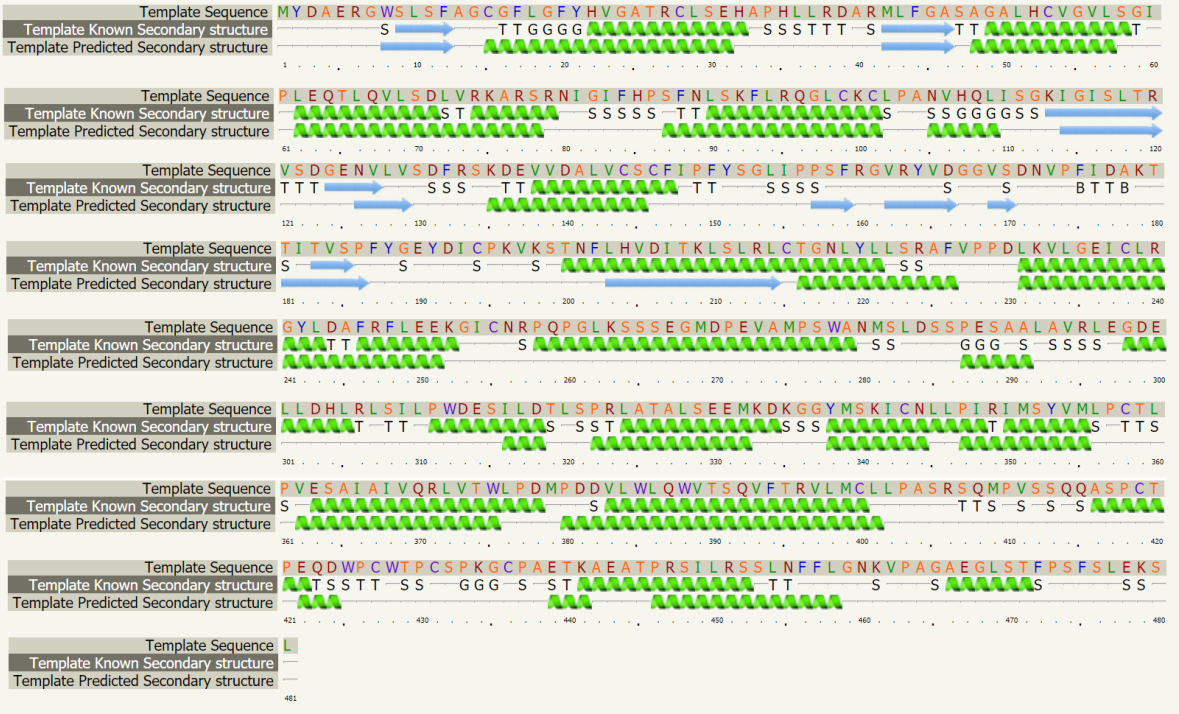


Figure 4.24 (continued)

4.5.2 Detailed structural architecture of the models

The patatin domain of each of the models is comprised of a cluster of β-strands and a cluster of α-helices, forming a classical α/β confirmation. While it is artificial to do this, for simplicity of discussion, the results will be presented separately based on secondary structure.

4.5.2.1 Patatin domain β-strands

All the models have a consistent β-sheet core consisting of 3 parallel β-strands (Strands 1, 2, 3) and one additional β-strand (Strands 4), which runs anti-parallel to strands 1, 2 and 3 (Table 4.5).

Models 1, 2, 4, 6 and 7 have an additional β-strand (Strand 5), which also runs anti-parallel to strands 1, 2 and 3.

These 5 main strands are highly conserved across these models, varying only slightly in strand length. In each model strand 3 is the longest and lies along the entire length of the central core of the protein. Strands 1, 2 and 4 are of similar lengths, with strand 5 being markedly shorter. This varies slightly in model 5, 6 and 9 in which strand 1 is shorter than the other models.

On top of the core 5 strands, model 1 has an additional strand (strand 5') which runs parallel with strands 1, 2 and 3. This is a short strand which forms a small anti-parallel pair with strand 6.

Models 5 and 9 also have an additional strand (Strand 5') which runs parallel to strand 1 in the structure. This is a short additional strand starting at residue 179 forms an additional stable structural element.

Table 4.5 β-strands within the patatin domain of all models

Model	Strands 1,2,3,4	Strand 5	Additional strand information
1	1,2,3,4	6	Additional strand 5' (short complement to 6)
2	1,2,3,4	5	
3	1,2,3,4		
4	1,2,3,4	5	
5	1,2,3,4		Additional strand 5 running parallel to strand 1. Strands 1 and 4 are very short
6	1,2,3,4	5	Strand 1 is very short
7	1,2,3,4	5	
8	NA		C-terminal only
9	1,2,3,4		Additional strand 5 running parallel to strand 1

*Strands 1-5 comprise of a consistent core of the protein.

4.5.2.2 Patatin domain α- helices

In these models, there are five main conserved α-helices, which cluster together spatially. Helix A lies between the first and second β-strands. Then three α-helices (helices B, C and D) lie between strand 2 and 3. There are no helices between the two anti-parallel strands 3 and 4, but one between strand 4 and 5 (helix E; Table 4.6).

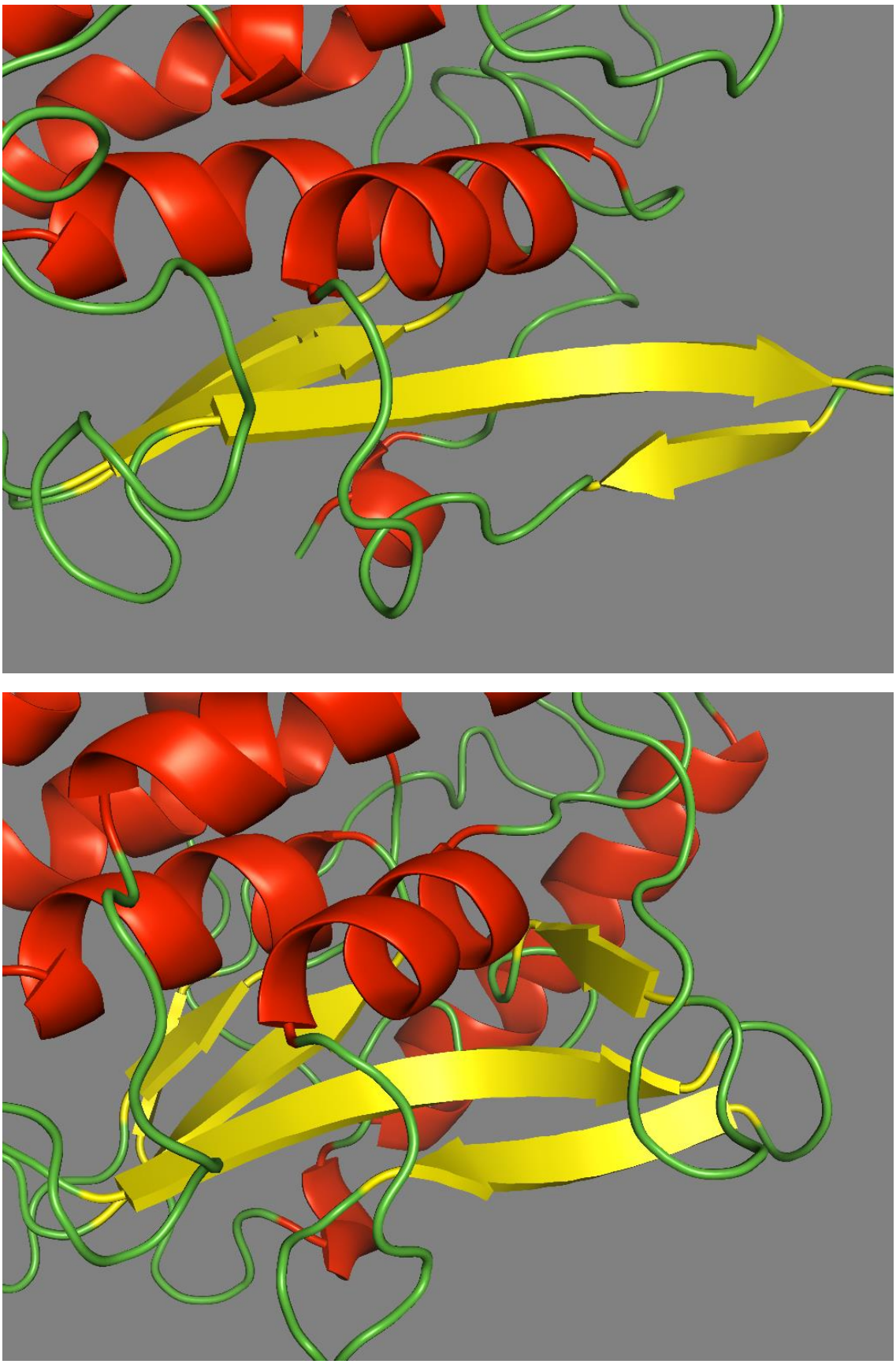


Figure 4.25 β-sheet core of models 3 and 6

β-sheets represented as arrows coloured yellow, α-helices represented as spiral turns in red and flexible loops in green.

Top panel; model 3.

Bottom panel; model 6.

Table 4.6 α- helices within the patatin domain of PNPLA3 models

Model	α-helix A	α-helix B,C and D	α-helix E	Additional helices
1	A,	B,C,D	E	
2	A,	B,C,D	E	F small terminal
3	A	B,C,D	E	
4	A	C,D,E	F	Additional helix B'.
5	A	B,C,E	F	D'
6	A	B,C,D	E	F and G after patatin
7	A	C,D,E	F	Additional small helix B'.
8	NA			C-terminal only
9	A	B,C,E	F	D'

*Helices A-E comprise of a consistent core of the protein.

In addition to the 5 conserved helices, a small additional helix (Helix B') is predicted to lie between helix A and strand 1 in models 4 and 7. Model 2 also has and additional helix (helix F) between helix E and strand 5.

Model 6 extends an additional 60 residues beyond the other models of the patatin domain. The initial 179 residues are structurally conserved with models 1,2,4 and 7. The additional residues form 2 additional helices (helices F and G) between strand 5 and the C-terminal. Helix F forms a long helix which positions on the opposite side of the β-strand cluster than the other helices; while strand G is a small strand which sits on the periphery of the protein.

4.5.2.3 N-terminal structure

Models 5 and 9 extend across the full protein sequence. Like model 6 they have an additional helix (helix h), which extends across the opposite side of the β-strands from the main helix cluster; however, this is shorter and less pronounced.

The N-terminal region consists of an additional seven helices in model 5 (helix I, J, k, L, M, N and O) and 11 helices in model 9 (helix I, J, K, L, M, N, O, P, Q, R, S and T); which do not interact with the core patatin domain in any obvious way.

The model of the N-terminal domain (Model 8) is primarily a helical structure and displays moderate levels of disorder. There is one initial helix (helix A) at the N-terminal of the protein followed by two short anti-parallel β-strands (strands 1 and 2). The strands are then followed by 8 further helices, six of which (Helices B, C, D, E, F and G) are spatially linked, forming a helical cluster.

4.5.2.4 Structural alignments

In addition to the conservation of the secondary structure elements, models 1, 2, 4, 5 and 7 align to an almost identical conformation (Figure 4.26).

Contrary to this model 3 has significant changes in the alignment, in particular with large conformational shifts in loop and coil regions, which is more similar to the conformation exhibited in model 9 (Figure 4.27).

There is no spatial conservation of the C-terminal domain in any of the models, despite similar secondary structure, and model 8 cannot easily align with the two full length models (Figure 4.28).

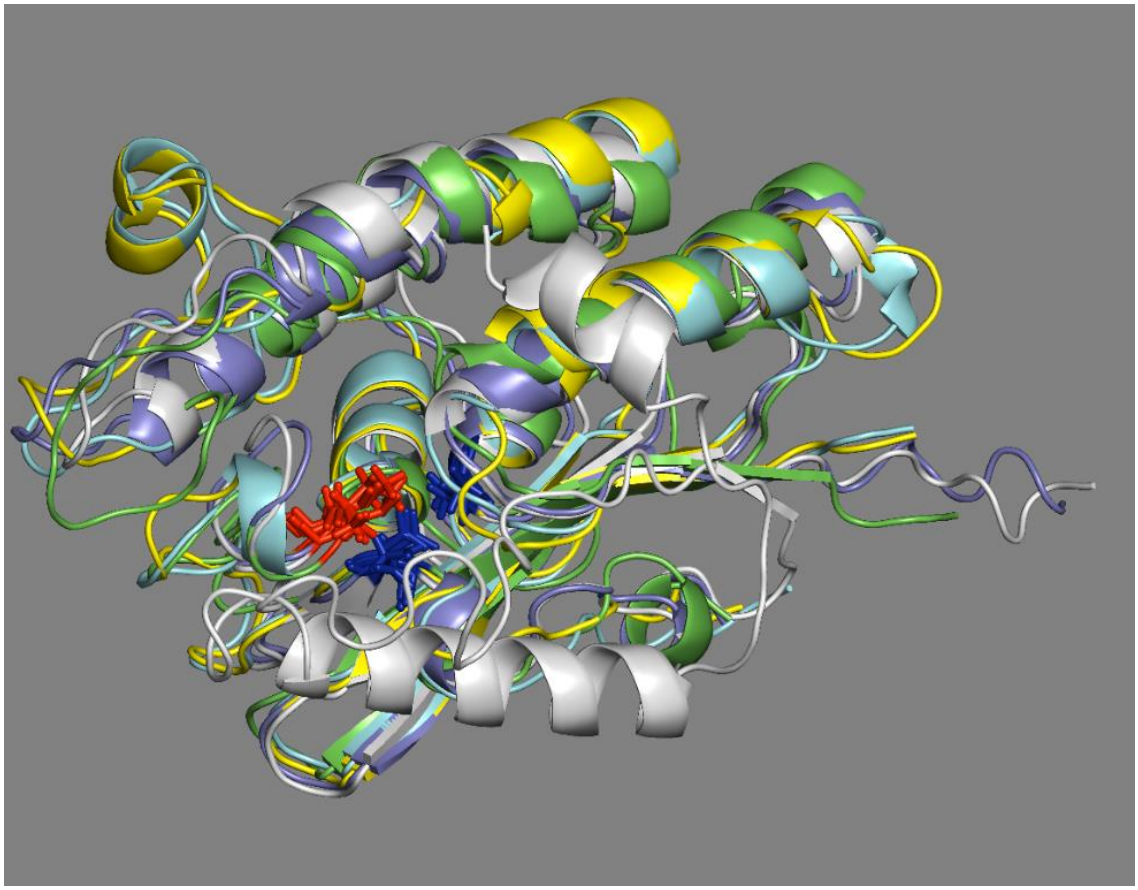


Figure 4.26 Aligned three-dimensional structures of Models 1, 2, 4, 6 and 7

β-sheets are represented as arrows, helices are represented as spiral turns. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue. Model 1 - green; Model 2 - light blue; Model 4 - yellow; Model 6 - white; Model 7 - light purple.

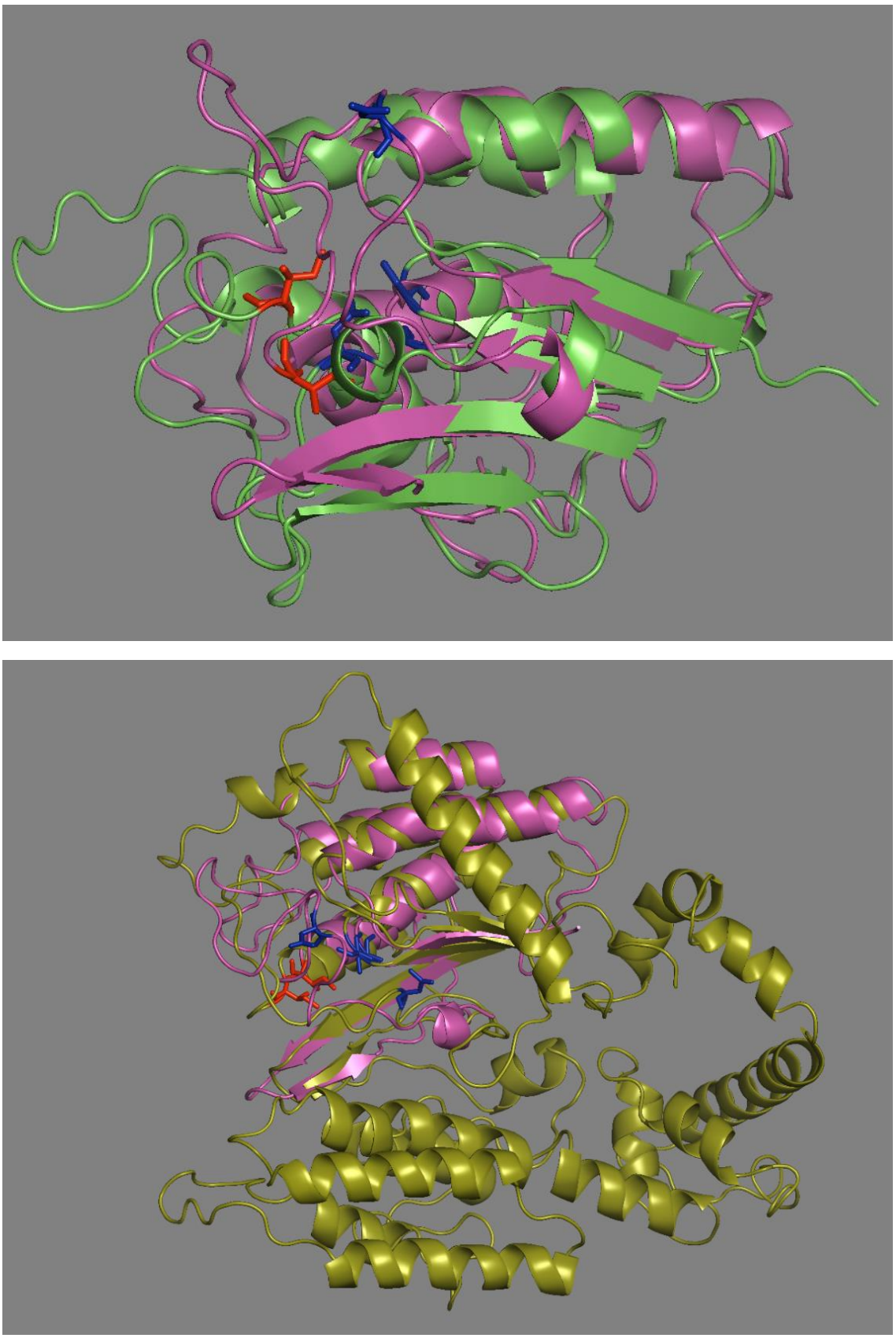


Figure 4.27 Aligned three-dimensional structures of Models 1, 3 and 9

β-sheets are represented as arrows, helices are represented as spiral turns. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.

Top panel; Model 1: green; Model 3: pink.

Bottom panel; Model 3: pink; Model 9: dark yellow.

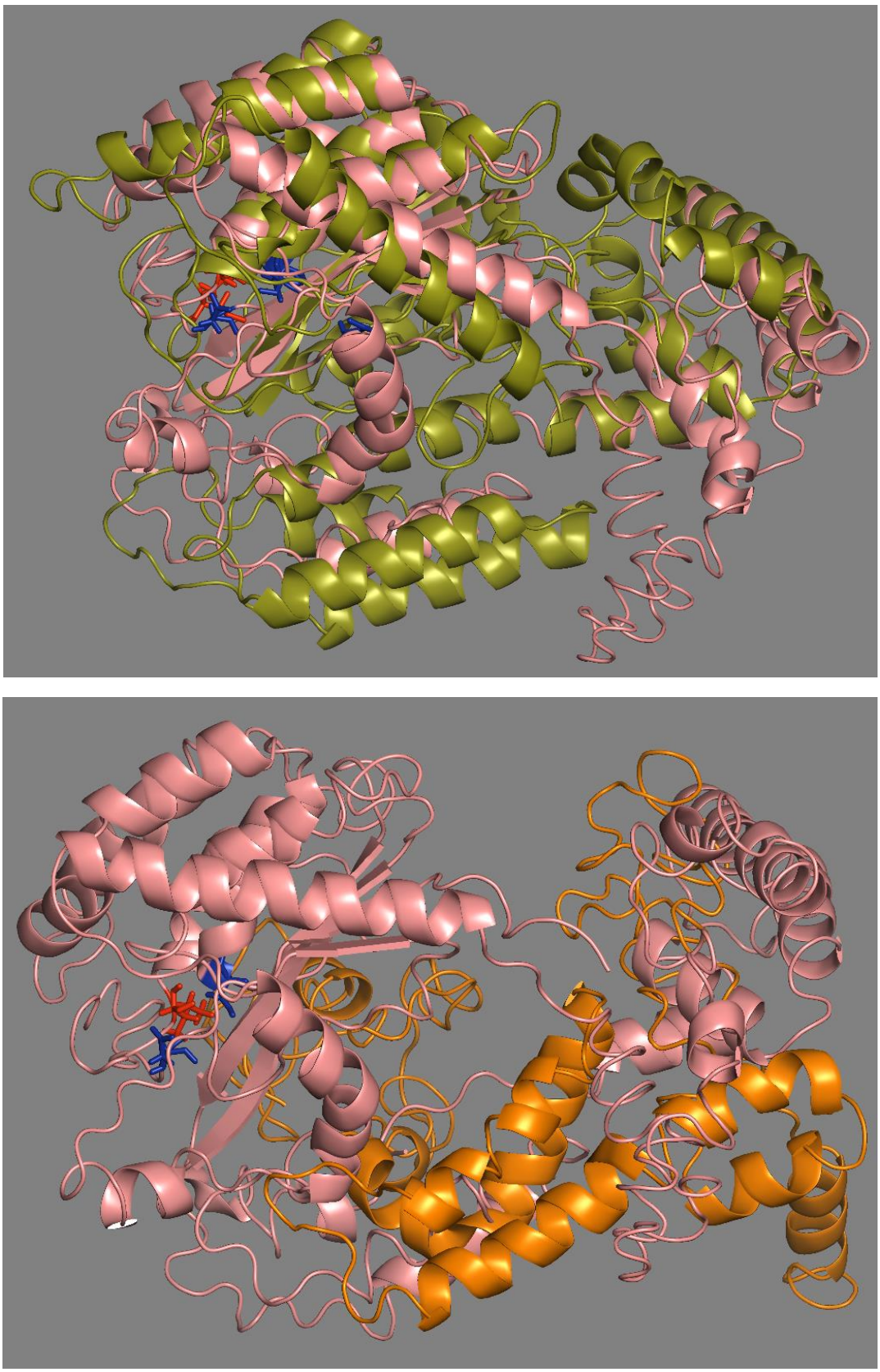


Figure 4.28 Aligned three-dimensional structures of Models 5, 8 and 9

β-sheets are represented as arrows, helices are represented as spiral turns. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.

Top panel; Model 5: salmon; Model 9: dark yellow.

Bottom panel; Model 5: salmon; Model 8: orange.

4.5.3 Active site conformations

Models 1,2,4,6 and 7 place all the key functional residues under investigation (S46, D166 and I148) in the same spatial location (Table 4.7; Figure 4.18). The distance between all three residues is smallest in these models; where S47 and D166 are 3.2 A apart, I148 and S47 are 4.2A apart and I148 and D166 are 3.2A apart (Figure 4.21).

Table 4.7 Distances between catalytic residues

Model	Distance (Å)		
	S47-D166	I148-D166	S47-I148
1	3.2	3.2	4.2
2	3.2	3.2	4.2
3	17.6	20.1	4.1
4	3.2	3.2	4.2
5	5.7	10.5	7.1
6	3.2	3.2	4.2
7	3.2	3.2	4.2
8	NA	NA	NA
9	7.4	12.8	7.4

Models indicating the same placement of the key functional residues are highlighted in blue

This positioning places S47 on a small elbow hinge region between strand 2 and Helix B. Both I148 and D166 are positioned on a random coil region between the end of helix E (in model 4 and 7, this is helix F) and the C-terminal residue. This coil folds on itself to give both residues a close spatial location.

On the coil, I148 is only two residues from the end of helix E, and D166 lies on a coil region which although not clear on visible inspection, is predicted by STRIDE[411] to be on a small β-strand. In model 6 this β-strand is clearly visible. This additional strand sits antiparallel to strand 2, forming a sandwich of three strands antiparallel to one another (Figure 4.28).

In models 5 and 9 the distance between these residues is increased, although to a different extent across the models. The position of S47 again is in the elbow region between strand 2 and Helix B, and both residues I148 and D166 are on a flexible loop region after Helix E (models 5 and 9 helix F). Helix E has turned slightly causing an increase in distance between S47 and I148. The loop after helix E is longer and stretches far beyond the active site, causing a significant increase in distances between D166 and the other residues.

Model 3 is the only model that significantly deviates from the active site architecture. While the catalytic serine S47 remain in the same location, the loop containing D166 has less structure, and leaves D166 positioned more distant from S47 and I148. The distance between S47 and D166 is the greatest of all models at 17.6 A, the distance between I148 and S47 is 4.1 A and the distance between I148 and D166 is 20.1A (Figure 4.29).

In model 5, the distance between S47 and D166 is 5.7A, the distance between I148 and S47 is 7.1A and the distance between I148 and D166 is 10.5A. In model 9, the distance between S47 and D166 is 7.4A, the distance between I148 and S47 is 7.4A and the distance between I148 and D166 is 12.8A (Figure 4.30).

4.5.4 Substitution of the I148M Variant

Models 1, 2, 4, 6 and 7 are the only models in which the catalytic residues S47 and D166 are positioned close enough together for the I148M variation to cause a clear significant impact on the active site (Figure 4.31).

In models 3, 5 and 9 the impact of this change cannot readily be observed on visual inspection alone, as there are greater distances between residue 148 and the catalytic residues which makes the change appear negligible. However, the orientation of residue 148 suggests that there is minimal impact on the spatial conformation of the active site.

In model 1, 2, 4, 6 and 7, the substitution of isoleucine for methionine causes the residue to protrude minimally into the opening to the catalytic site. In models 1, 2, 4, and 7 this hardly impacts the large opening into the active site (Figure 4.29). However, potentially in model 6 which has a smaller opening due to the presence of additional helices, this could have some mild impact. (Figure 4.32)

Within the structure residue 148 could potentially interact with the residues 147F, 149P 150F and 151Y. The presence of the I148M variant appears to decrease the distance to 147F by 0.7A, while all other distances increase. This has no clear impact on the structure.

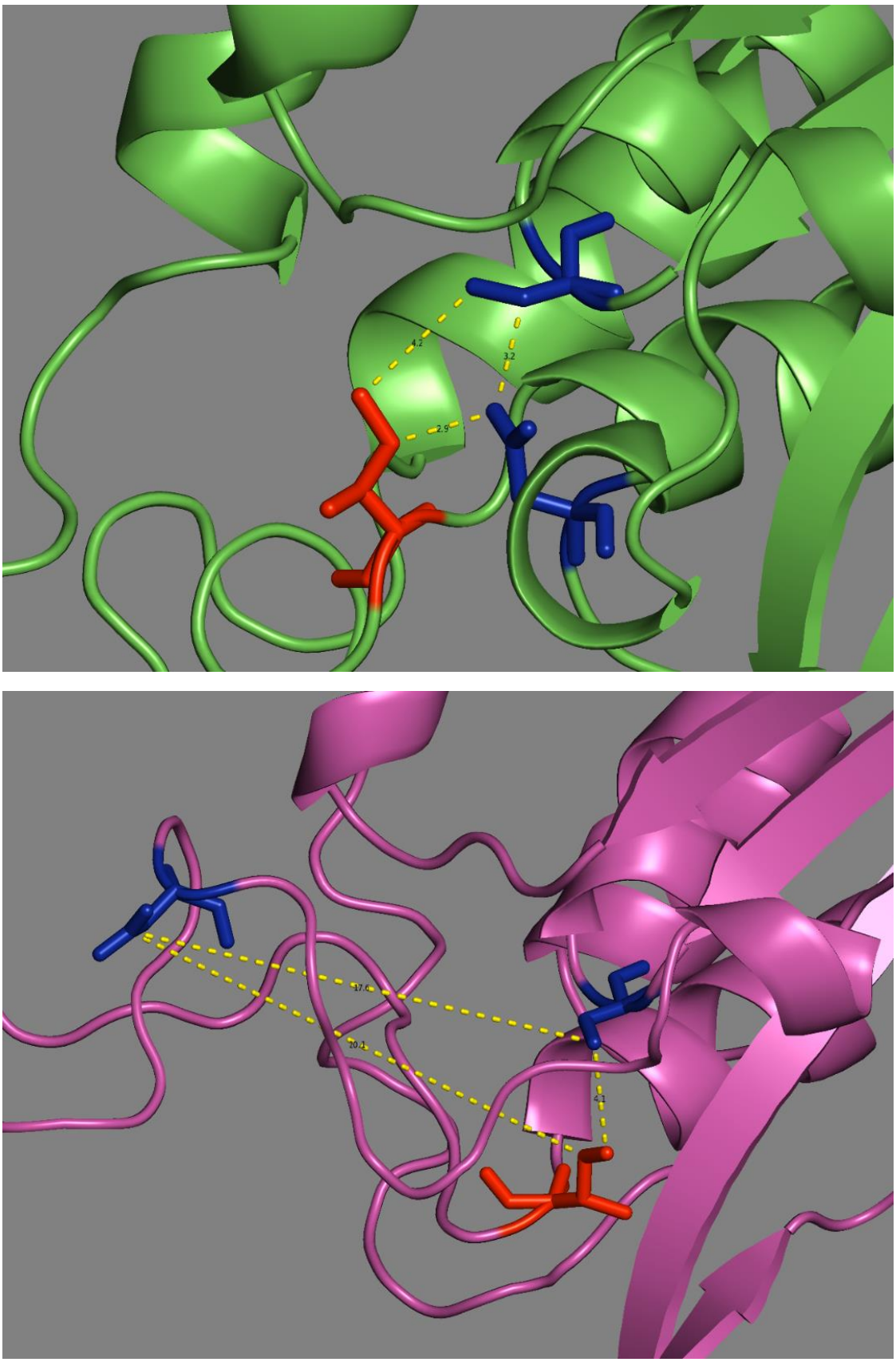


Figure 4.29 Active site residues of models 1 and 3.

The catalytic serine and aspartate highlighted in blue, isoleucine 148 shown in red. The distances between residues are shown as dashed yellow lines; distance between residues in Angstroms.

Top panel; Model 1.

Bottom panel; Model 3.

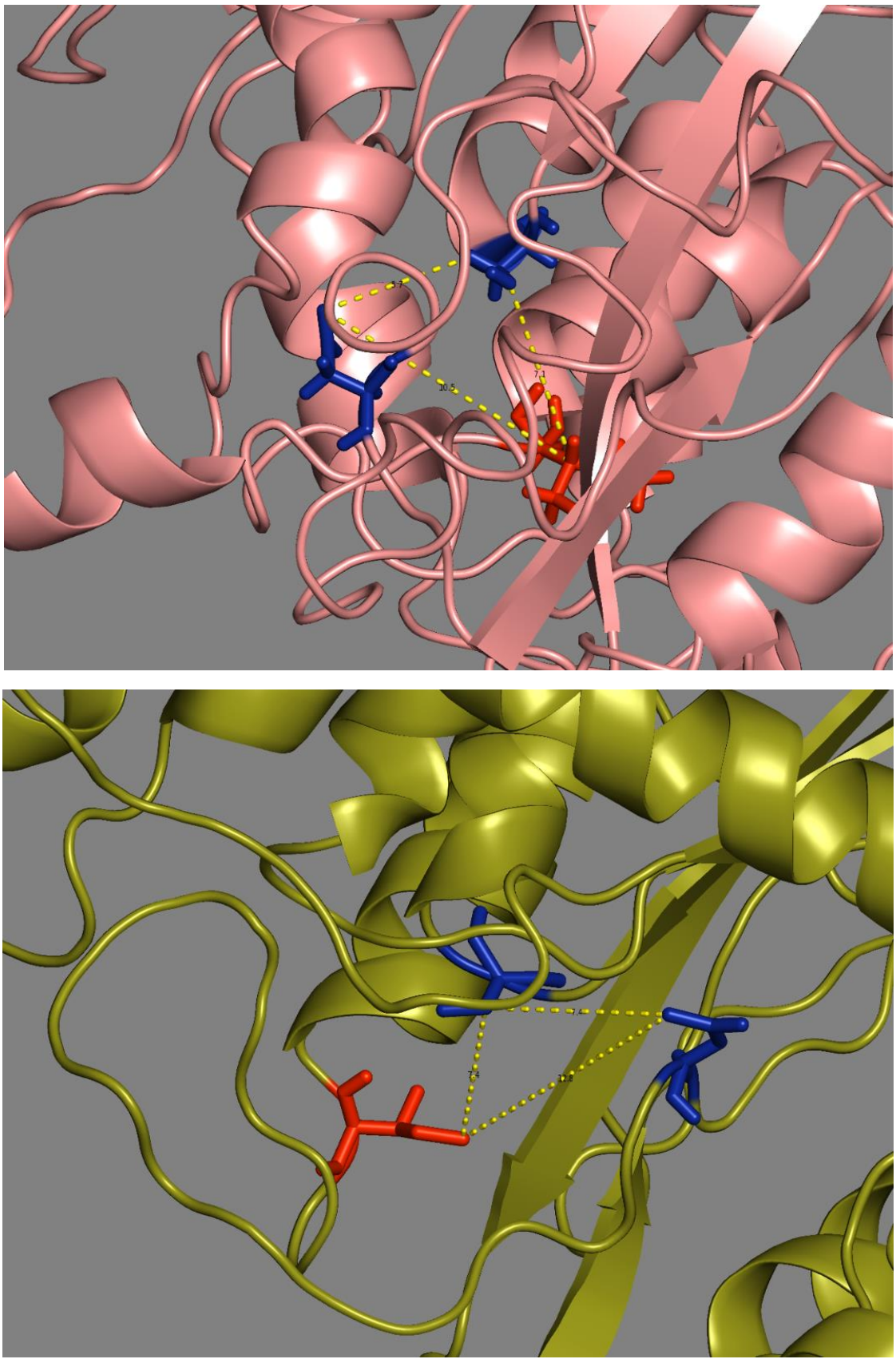


Figure 4.30 Active site residues of models 5 and 9

The catalytic serine and aspartate highlighted in blue, isoleucine 148 shown in red. The distances between residues are shown as dashed yellow lines; distance between residues in Angstroms.

Top panel; Model 5.

Bottom panel; Model 9.

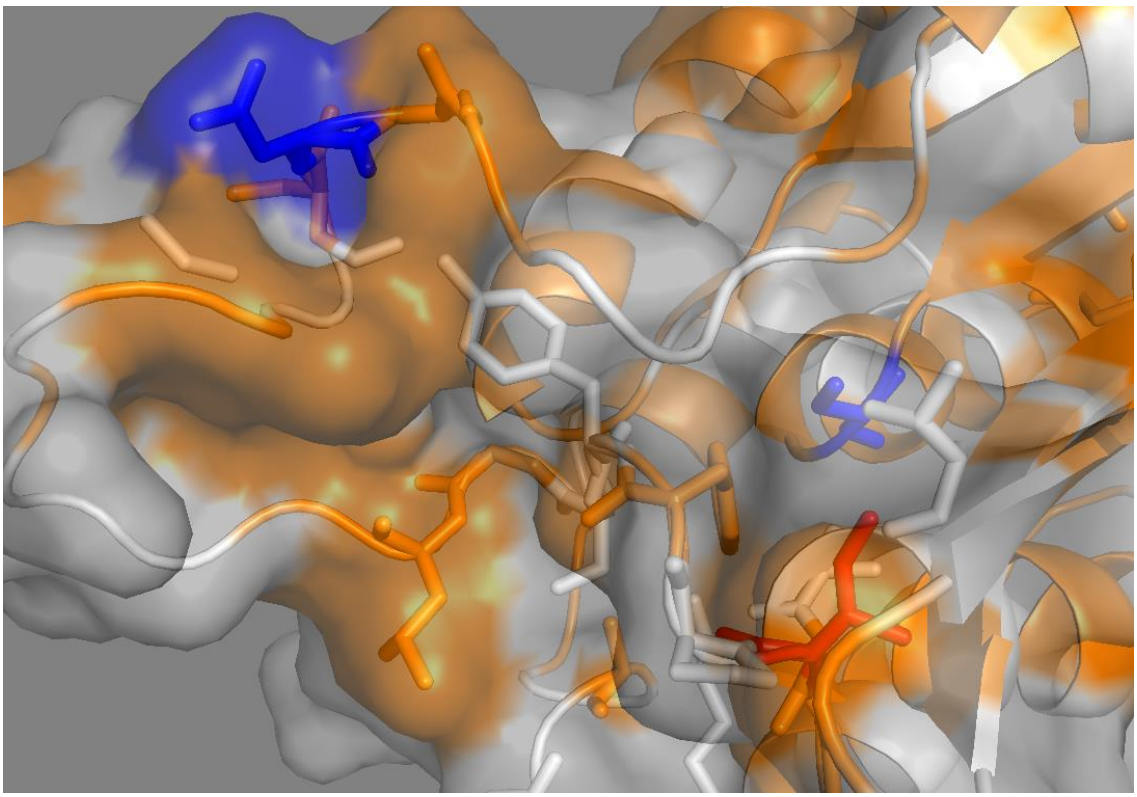


Figure 4.31 Model 3 potential interaction

Hydrophobic residues highlighted in orange, other residues in grey. Residue 148 is highlighted in red and the catalytic Aspartate and serine in green.

4.5.5 Surface hydrophobicity

In models 1, 2, 4 and 7, the face of the exposed pocket lies on region with a large number of hydrophobic residues which completely surround the active site pocket. This is similar in model 3, which also has a hydrophobic ring surrounding the active site; however, catalysis would require a significant movement of the active site in this model (Figure 4.33).

The smaller active site tunnel in model 6, disrupts the local hydrophobicity with an additional hydrophilic surface facing helix (Figure 4.34). This forms a potential lid region spanning residues 174 to 227 (Figure 4.35).

Similar to model 6, model 9 has more restrictive access to the active site and a small hydrophobic cluster around the catalytic residues; however, there are less total solvent exposed hydrophobic residues compared to model 6 (Figure 4.34).

The I148M Variant has no clear impact on the hydrophobicity of the active site (Figure 4.36).

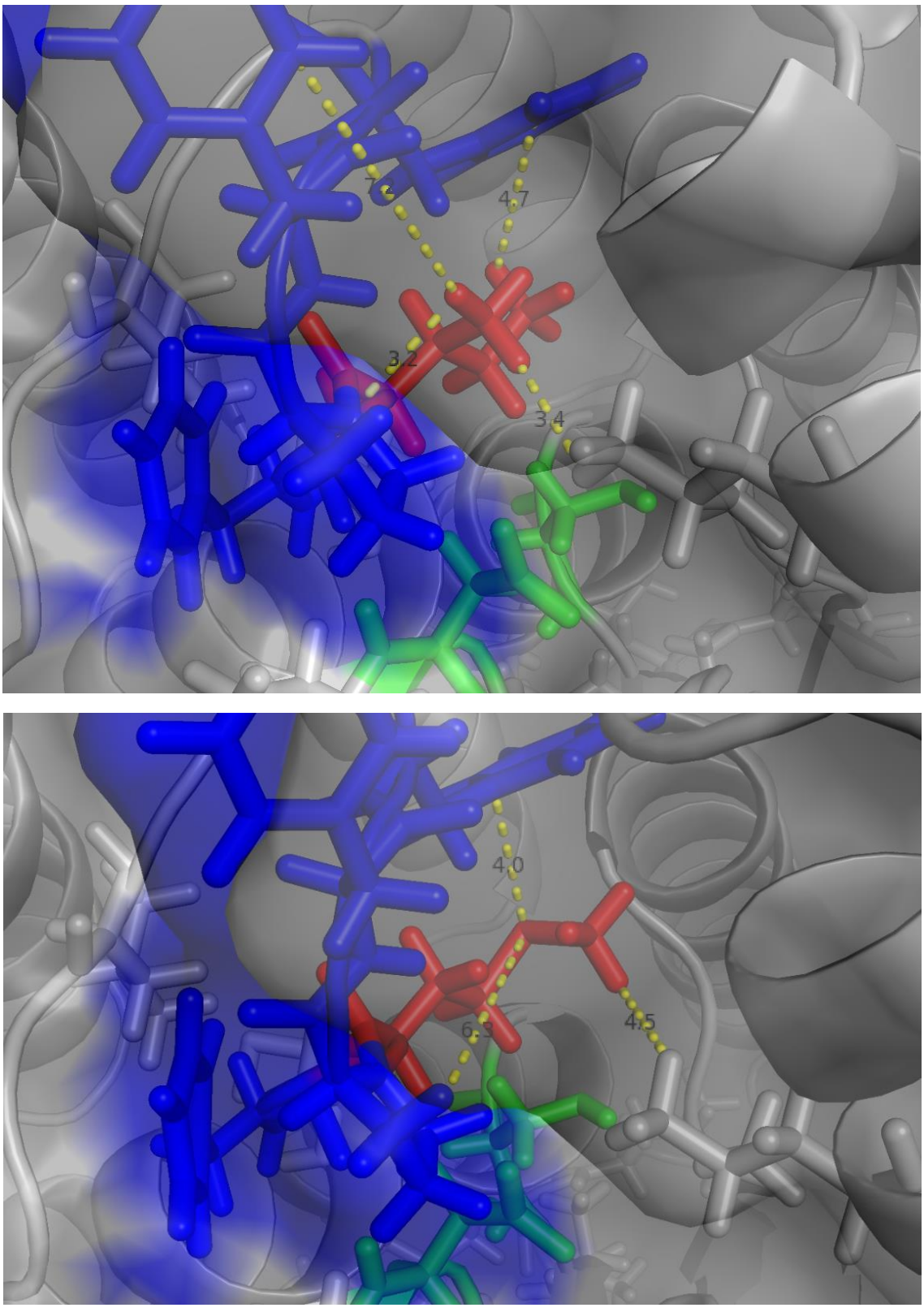


Figure 4.32 Model 6 residue 188 variant interactions

Catalytic serine and aspartate highlighted in green, methionine 148 shown in red. Distance between residues shown by dashed yellow line, distance between residues in Angstroms. Potential 148M loop interaction partners highlighted in blue.

Top panel; Wild type I148.

Bottom panel; variant M148.

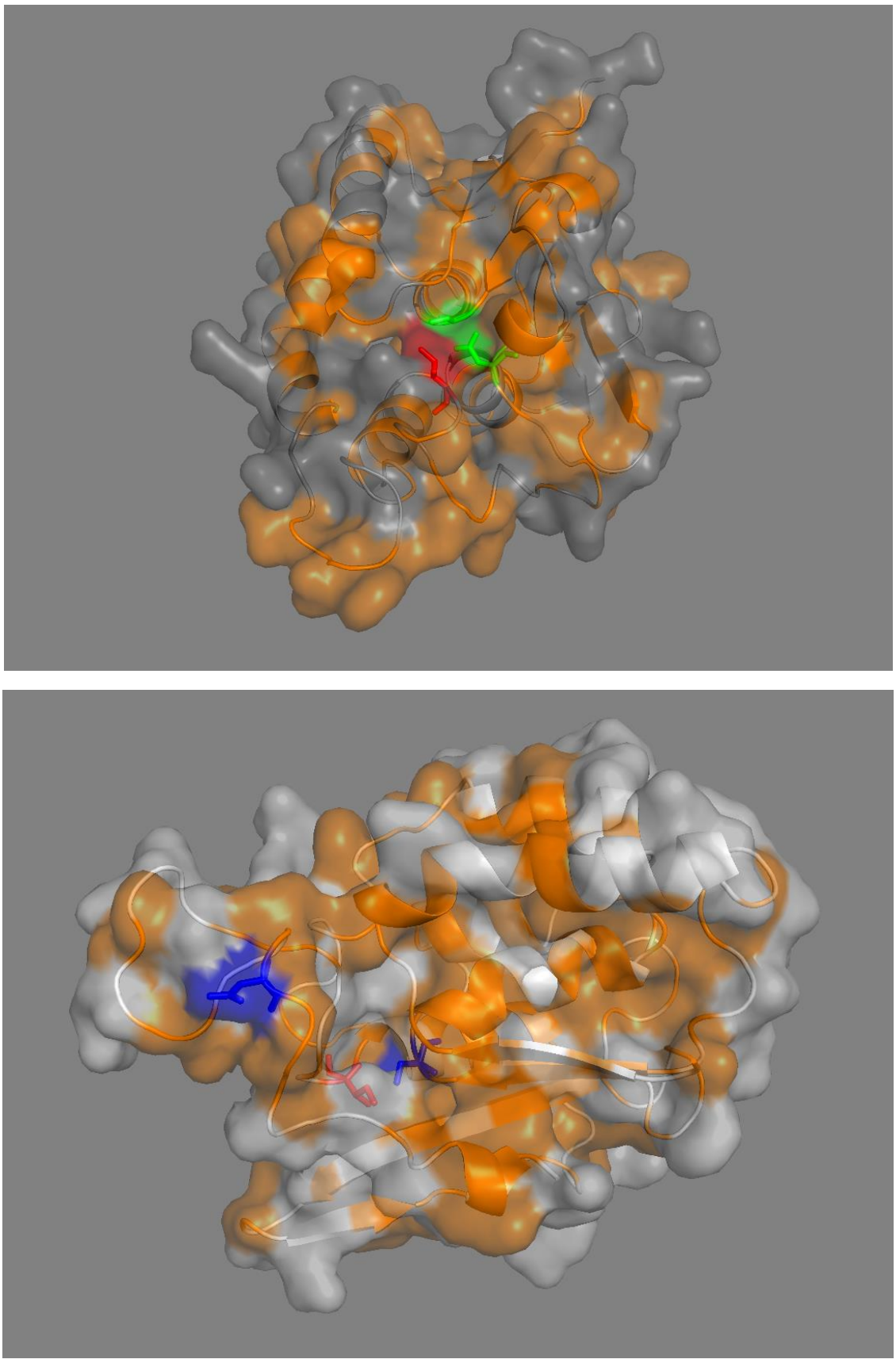


Figure 4.33 Hydrophobicity mapping to protein surface

The hydrophobic residues are highlighted in orange, other residues in grey. Residue 148 is highlighted in red and the catalytic Aspartate and serine in blue.

Top panel; Model 1.

Bottom panel; Model 3.

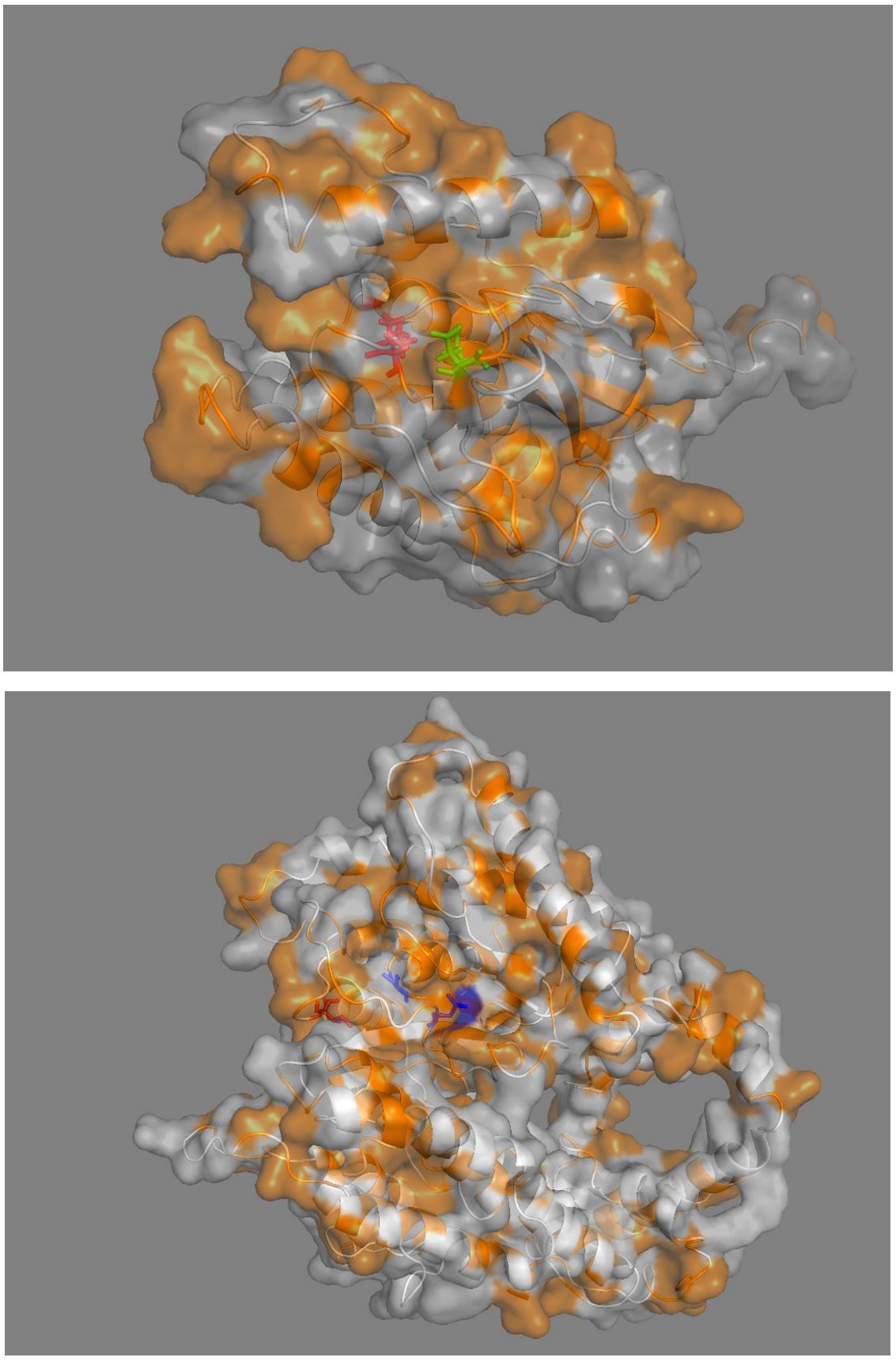


Figure 4.34 Hydrophobicity mapping to protein surface

The hydrophobic residues are highlighted in orange, other residues in grey. Residue 148 is highlighted in red and the catalytic Aspartate and serine in blue.

Top panel; Model 6.

Bottom panel; Model 9.

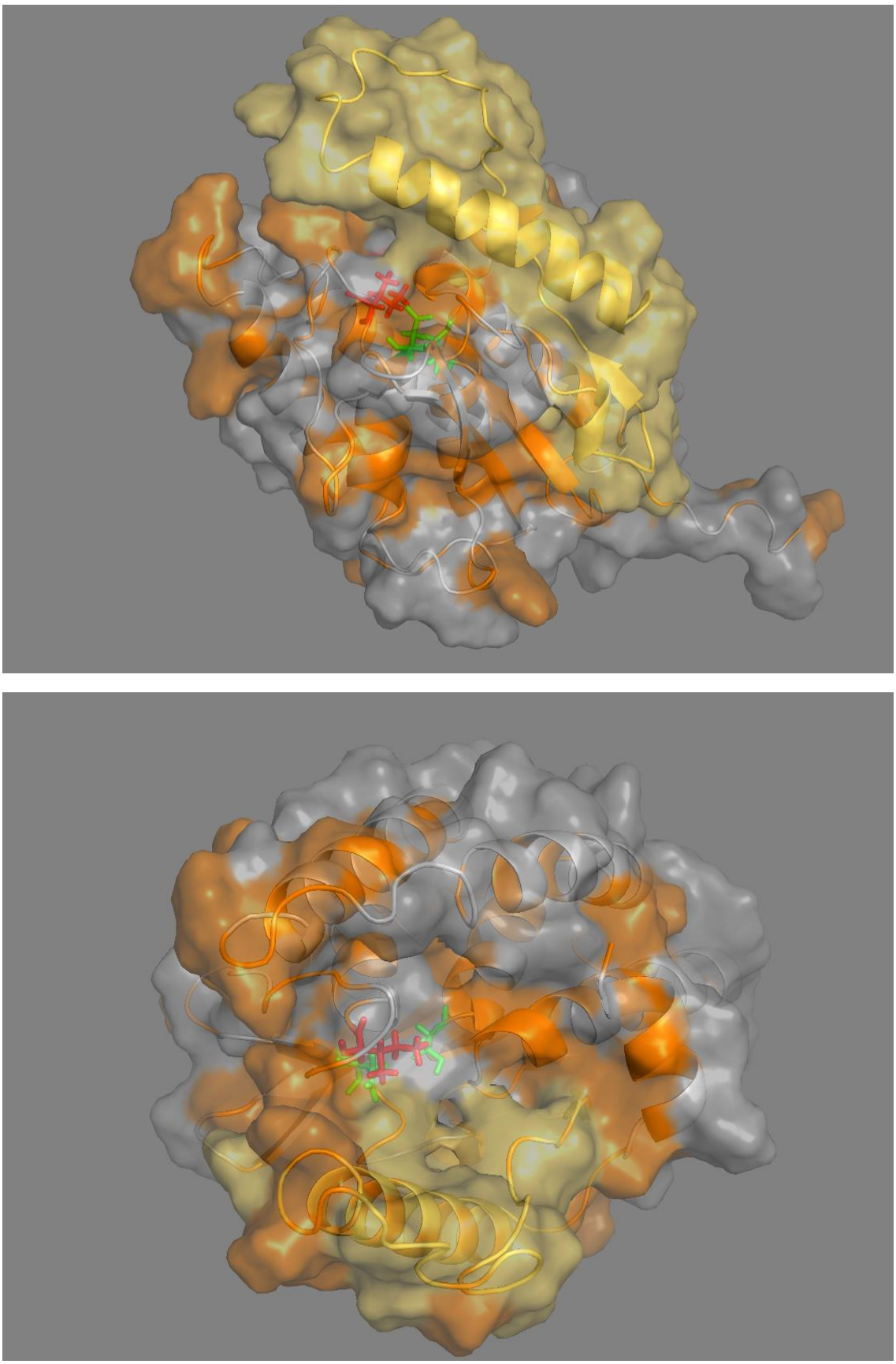


Figure 4.35 Model 6 hydrophobicity map with lid

The hydrophobic residues are highlighted in orange; the other residues are highlighted in grey. Residue 148 is highlighted in red and the catalytic aspartate and serine are highlighted in green. The potential lid helix is highlighted in yellow. Each panel presents a view from a different angle of the protein.

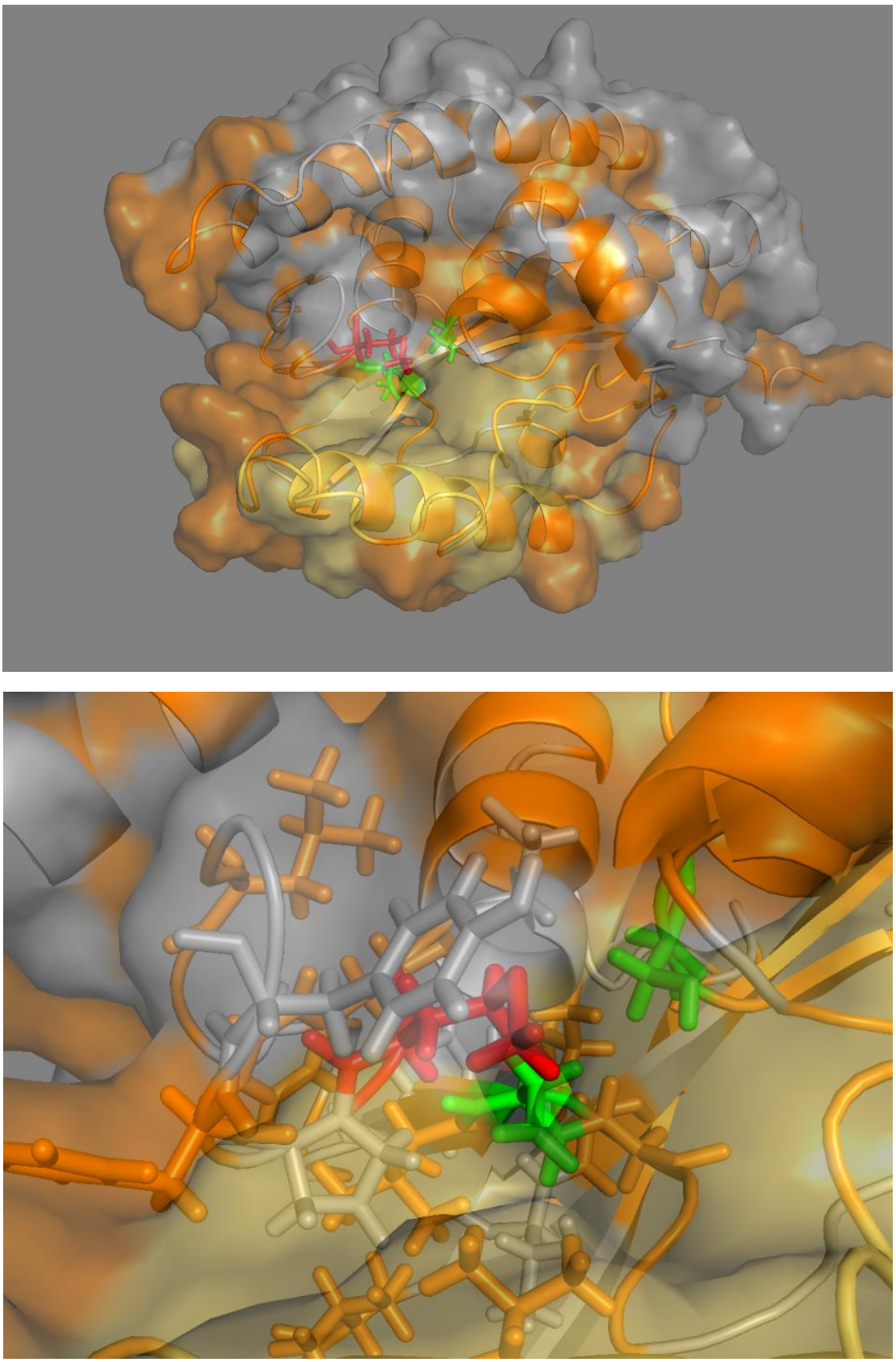


Figure 4.36 148 methionine substituted model 6 hydrophobicity map with lid

The hydrophobic residues are highlighted in orange; the other residues are highlighted in grey. Residue 148 is highlighted in red and the catalytic aspartate and serine are highlighted in green. The potential lid helix is highlighted in yellow. Each panel presents a view from a different angle of the protein. **Bottom panel;** shows additional residues as sticks.

4.5.6 SWISS-MODEL modelling quality

Homology models produced by SWISS-MODEL were automatically limited to regions of adequate homology. Each of the four models consist of the initial 180 amino acids of PNPLA3, representing the traditional patatin domain (Table 4.8).

Table 4.8 SWISS-MODEL quality assessment

Model	GMQE	QMEAN	Template	Sequence identity (%)	Method	Sequence similarity	Coverage
1	0.18	-4.91	1OXW	19.19	X-ray, 2.20 A	0.28	0.36
2	0.19	-3.78	4AKF	20.59	X-ray, 2.90 A	0.30	0.35
3	0.16	-4.58	3TU3	23.67	X-ray, 1.92A	0.30	0.35
4	0.18	-3.68	4KYI	20.71	X-ray, 3.08A	0.30	0.35

Sequence alignment of the 4 templates used by SWISS-MODEL to generate the models highlights the consistent regions of homology are within the initial 180 residues of the N-terminal domain. The coverage of all these alignments was almost identical between 35 and 36%. The best alignment was with the template of 3TU3, having the highest sequence identity with PNPLA3, at 23.67% over the shorter modelled patatin domain (Figure 4.37).

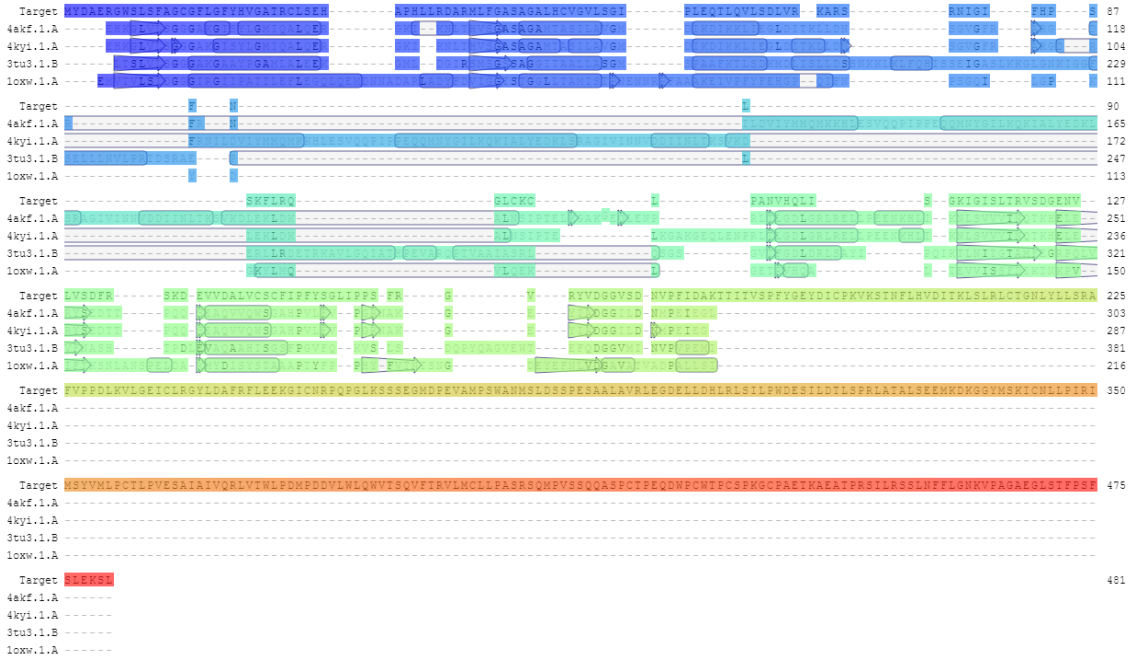


Figure 4.37 Alignment of templates used to generate SWISS-MODEL homology models against full length PNPLA3

Areas of highest similarity are highlighted in blue, those with the lowest in red.

The local quality of each model was generally between 0.4 and 0.8 along the length of the model. All models showed a dip in local quality around residue 80, which was particularly pronounced in models 2 and 4, where this region was extended slightly beyond residue 100. Model 3 had the highest local similarity over the first 140 residues and the smallest dip at residue 80; however, in this model there was a noticeable dip to 0.25 around residue 160 (Figure 4.38).

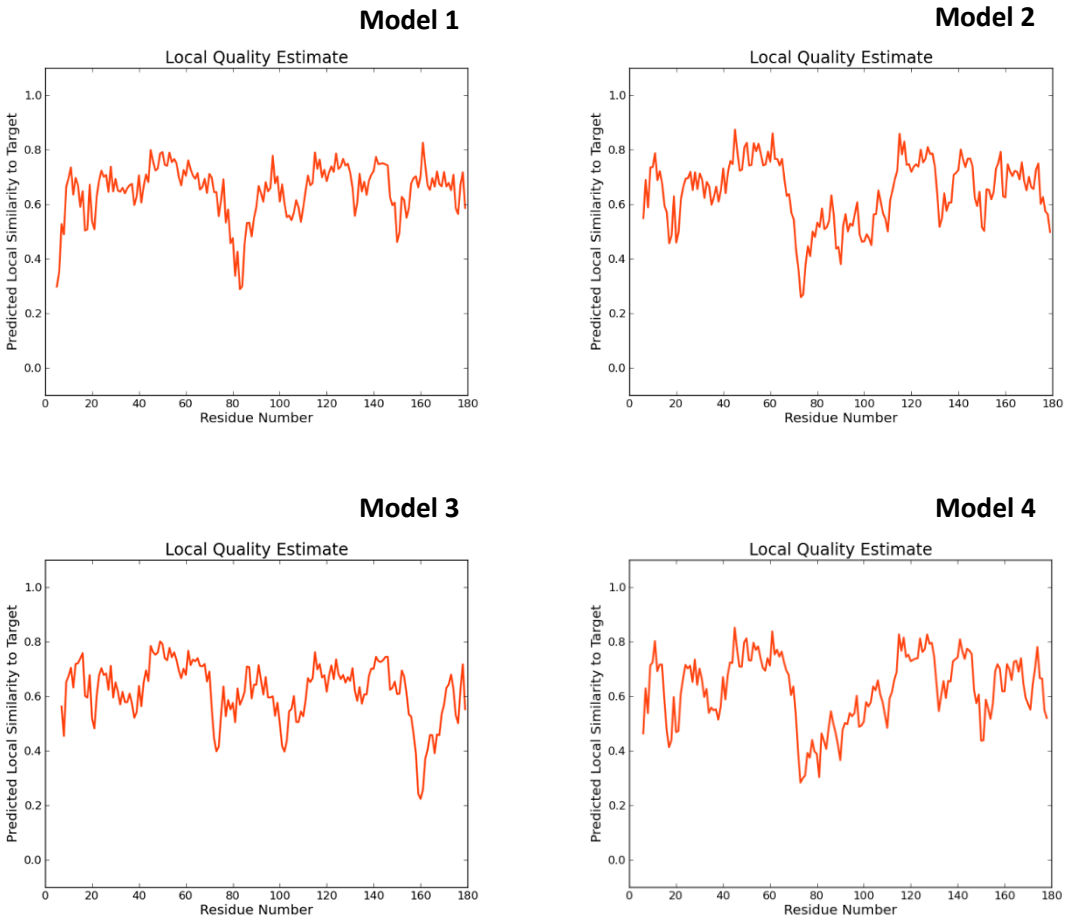


Figure 4.38 Comparison of SWISS-MODEL results local quality estimate

The GMQE scores for all four models were similar ranging from 0.16 and 0.19. Although Model 3 was based on the highest resolution structure (1.93Å) and shared the most sequence identity with PNPLA3, Models 2 and 4 had the highest confidence rating with QMEAN scores of -3.78 and -3.68 respectively.

Model 2 has the best All atom score at -2.68. Model 1 has the best solvation score, at -0.70 but the worst torsion score of -3.86. Model 4 has the best Cβ score at -1.06, and shared the joint best Torsion score with model 3 at -2.86. However, model 4 also has the worst solvation score at -1.95 and model 3 the worst Cβ score at -3.46 (Figure 4.39).

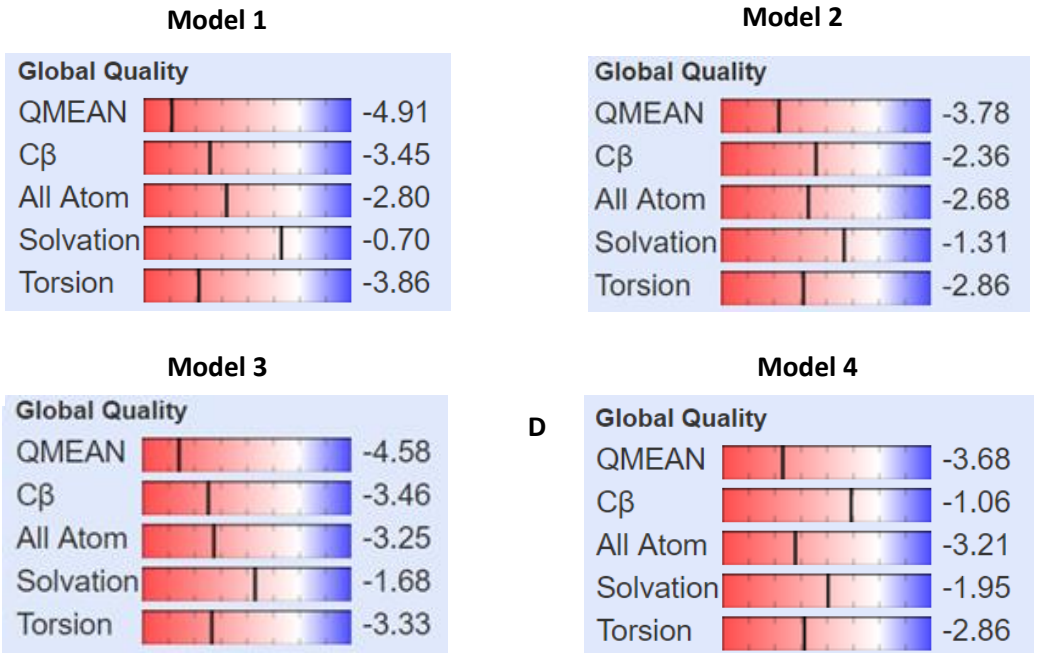


Figure 4.39 Global quality estimate of SWISS-MODEL generated homology models

The bar places the model on a sliding scale relevant to PDB deposited structures.

Comparing QMEAN score against a normalised non-redundant set of PDB structures showed that all four models were located below the average PDB quality as expected range for a model based on low homology (Figure 4.40).

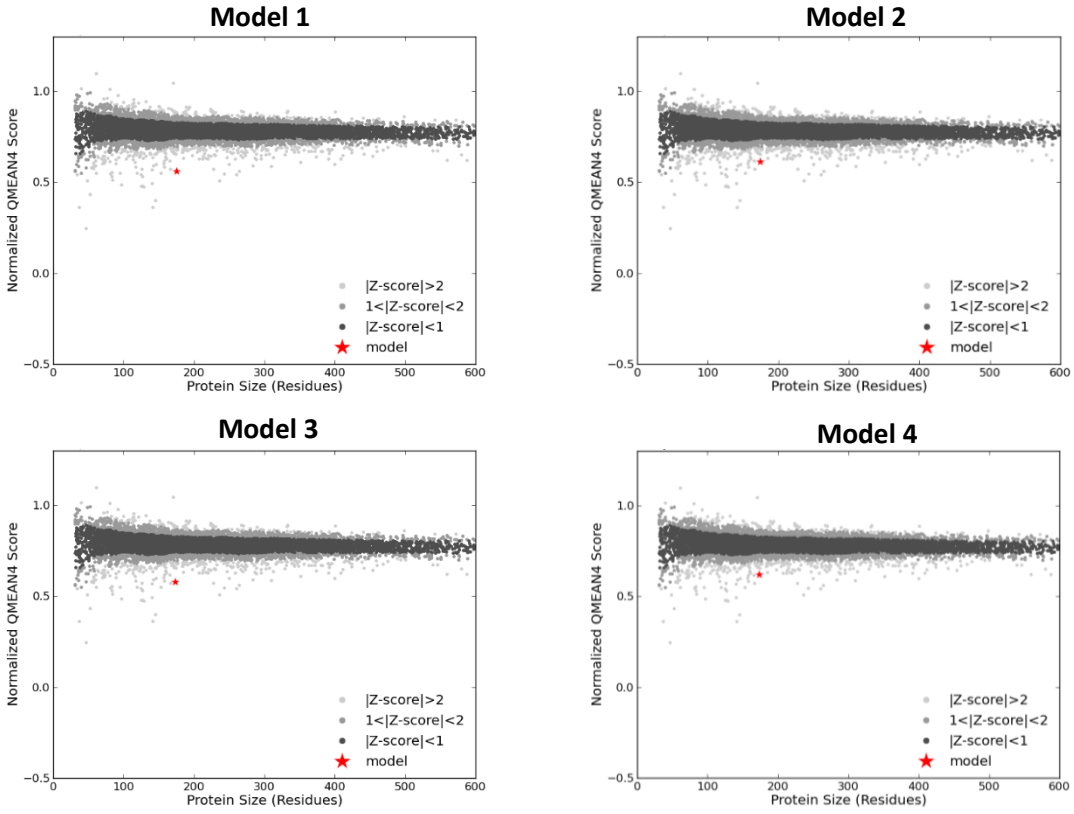


Figure 4.40 Comparison of the results obtained using SWISS-MODEL compared to the normalised QMEAN score for a non-redundant set of PDB structures

4.5.7 I-TASSER modelling quality

Homology models produced by I-TASSER do not have a minimum homology requirement with any templates and are therefore able to cover the entire length of the input sequence.

The C-terminal model 8 had the worst C-score of -2.33. The C-score of models of the patatin domain correlated with the length and the longer models had worse C-scores. Model 5 had the second lowest C-score at -2.07. Models 6 and 7 were significantly higher at -0.50 and 0.10 respectively. The C-score for model 9 was not generated within the manual run (Table 4.9).

The longer models had the worst average local quality estimates, and therefore model 7, the shortest model was the one with greatest average local quality predictions.

Table 4.9 I-TASSER quality assessment

Model	Residue range	C-score	TM-score	Average B-factor
5	1-481	-2.07	0.47	7.49
6	1-239	-0.50	0.65	3.19
7	1-179	0.10	0.73	2.24
8	239-481	-2.33	0.44	6.45
9	1-481	Not available	Not available	7.23

Looking more closely at the local quality estimated, based on simulated B-factors, we can see that in all of the models the highest confidence can be seen in the initial 150 residues with a small decrease around residue 80 as seen in SWISS-MODEL (Figure 4.41).

At 150 there is a short segment of decreased quality in all of the models; however, this is lowest in model 6 with the simulated B-factor only peaking at 6A. The largest peak in the models appears to be between residues 250 and 300, which corresponds to regions of predicted disorder, and simulated B-factor peaks at 15 in model 9 and 20 in model 5.

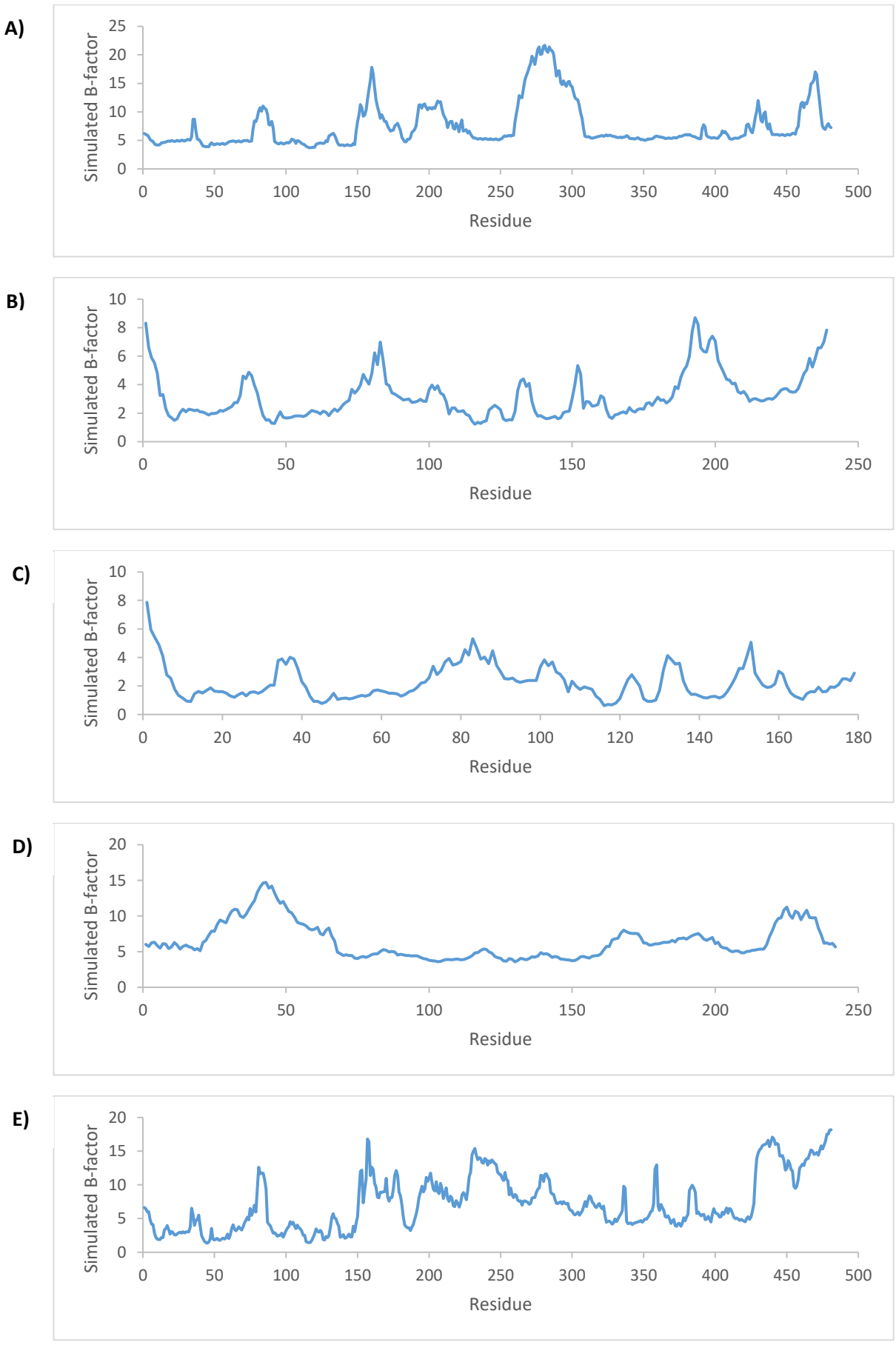


Figure 4.41 Simulated raw B-factors of I-TASSER generated models.

A) Model 5 **B)** Model 6 **C)** Model 7 **D)** Model 8 **E)** Model 9.

4.5.8 PROSESS model quality assessment

The overall quality score determined by PROSESS for the SWISS-MODEL models was 4.5, compared to an average overall quality of 2.1 for models generated by I-TASSER.

The model with the highest overall quality was model 3, with an overall quality score of 5.5. This model had significantly higher torsion angle quality when compared to the other models of 4.5 and shared the highest covalent bond quality of 7.5 (Table 4.10).

Table 4.10 Summary quality statistics produces by PROSESS

Model	Overall quality	Covalent bond quality	Non-covalent/ packing quality	Torsion angle quality
1	3.5	7.5	5.5	2.5
2	4.5	7.5	6.5	2.5
3*	5.5	7.5	5.5	4.5
4	4.5	7.5	5.5	2.5
5	1.5	3.5	3.5	0.5
6	2.5	4.5	4.5	1.5
7	2.5	6.5	5.5	1.5
8	1.5	3.5	3.5	0.5
9	2.5	6.5	3.5	1.5
1OXW	6.5	6.5	6.5	5.5
1CJY	4.5	6.5	5.5	4.5

The highest quality model highlighted in blue.

In all the models there were regions which had an increased number of outlier properties; these were located approximately between residues 70-80, 150-160, and 260-300 (Figure 4.42).

Models 1-4 had much better torsional angles, shown by less than 5 residues being strong outliers on the Ramachandran plots (Figures 4.43-4.46), Whereas models 5-9 had much worse torsional angles (Figures 4.47-4.51). Model 5 and model 8 had a particularly large number of outliers primarily corresponding to the C-terminal domain. Model 9 had the least number of outliers corresponding to this domain, potentially providing the most accurate C-terminal model.

There were certain regions of the protein which exhibited poor local quality in all of the models. Generally, these were in flexible loop regions, rather that within core helices and strands. A notable exception to this was the low quality anti-parallel sheets within the patatin domain of model 1 around reside 120-140 (strand 4).

Residues 70 - 80 corresponds on the shorter models to helix F, which is of poor quality. The exception is on model 3 in which this helix has less outliers, and the adjacent loop between sheet 5 is actually of poorer quality.

Residues 260 – 300 corresponds to a region within the C-terminal domain, Helix I based off model 9 architecture is of poor quality and almost the entire of helix J, extending along the loop all the way to helix K.

Residues 150-160 corresponds in all models, the loop between residues 148 and 166 and has a significantly larger proportion of outliers. This is a smaller region in the smaller models, and the best in model 3, although still of low quality. (helix E in models 1, 2, 3 and 6, or helix F in models 4, 5, 7 and 9)

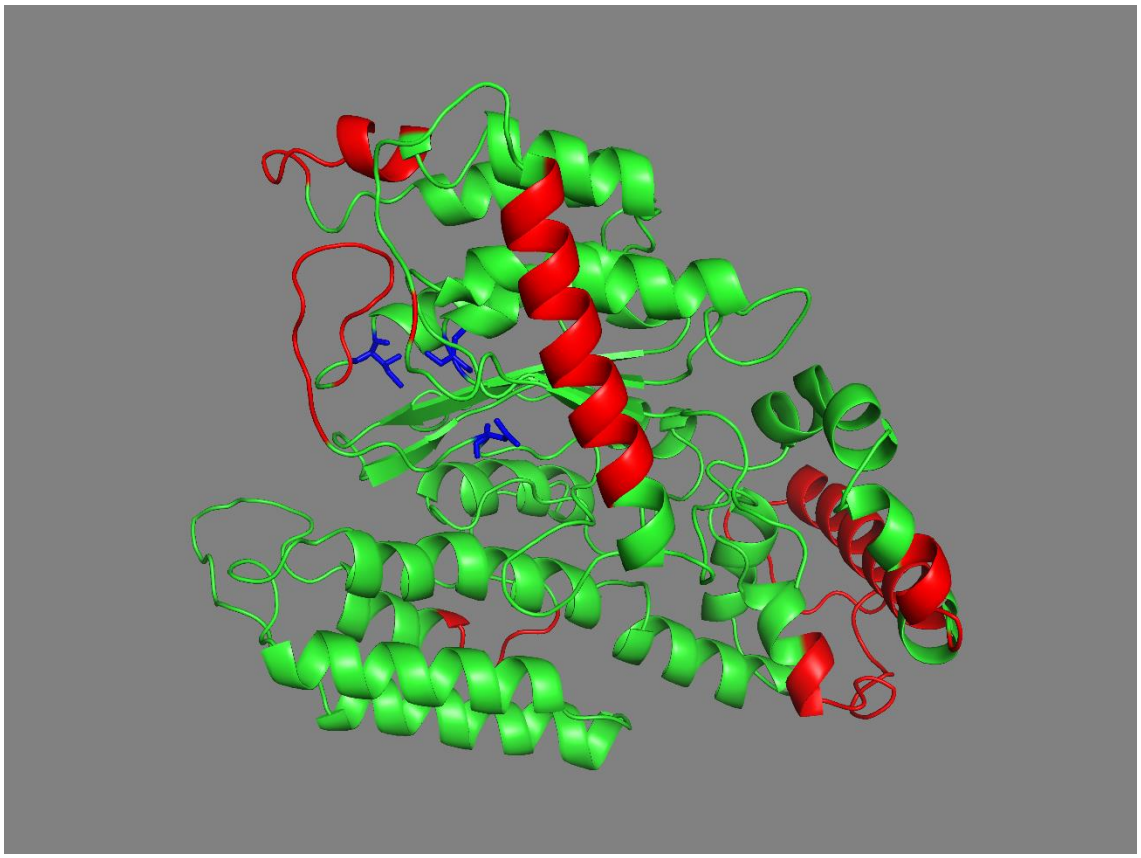


Figure 4.42 Simple plot of areas with high number of average residue outliers determined from models 1-9

Mapped onto model 9 structure, poorest quality highlighted in red. Key residues 47, 148 and 166 highlighted in blue.

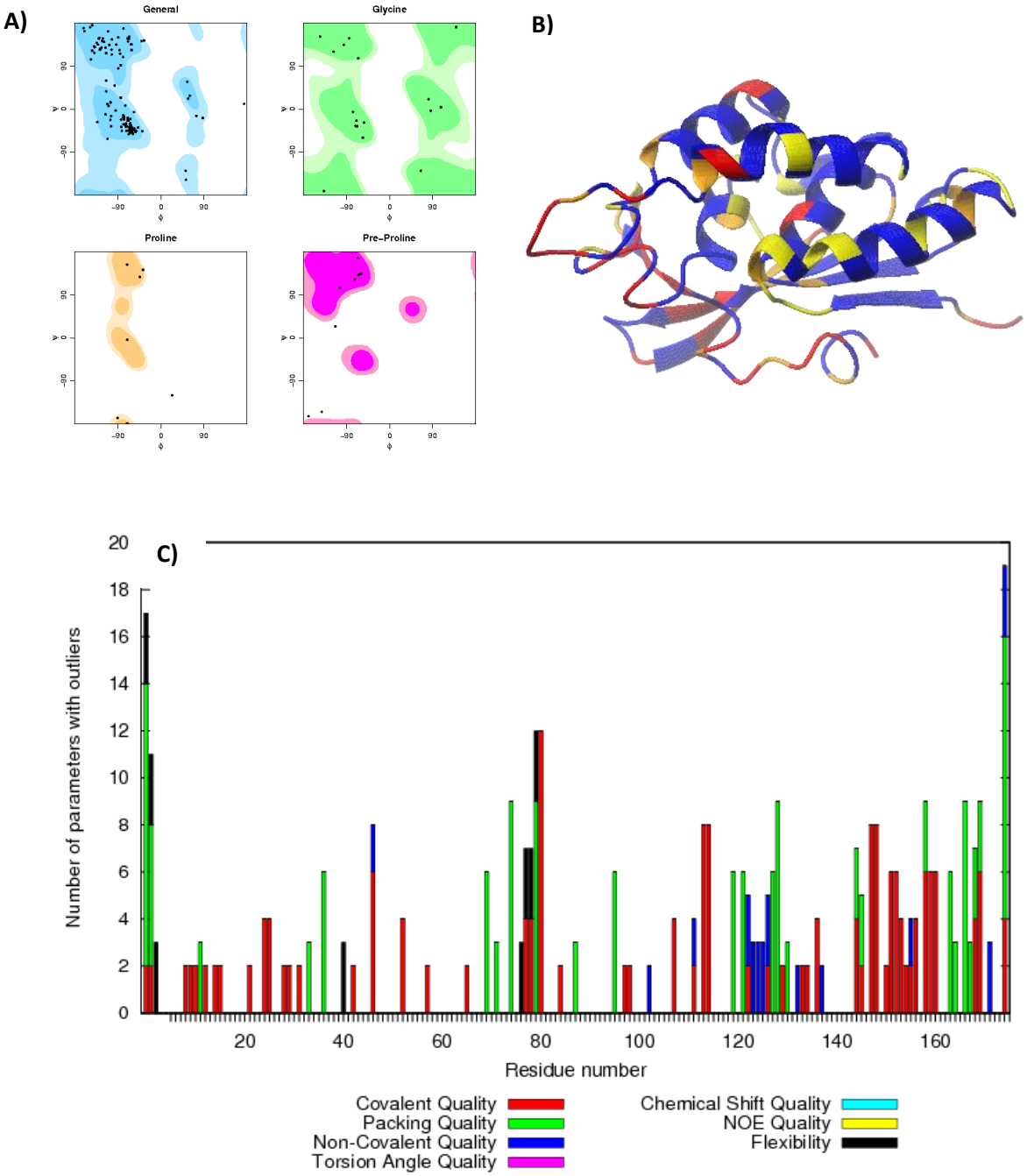


Figure 4.43 PROSESS determined model 1 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

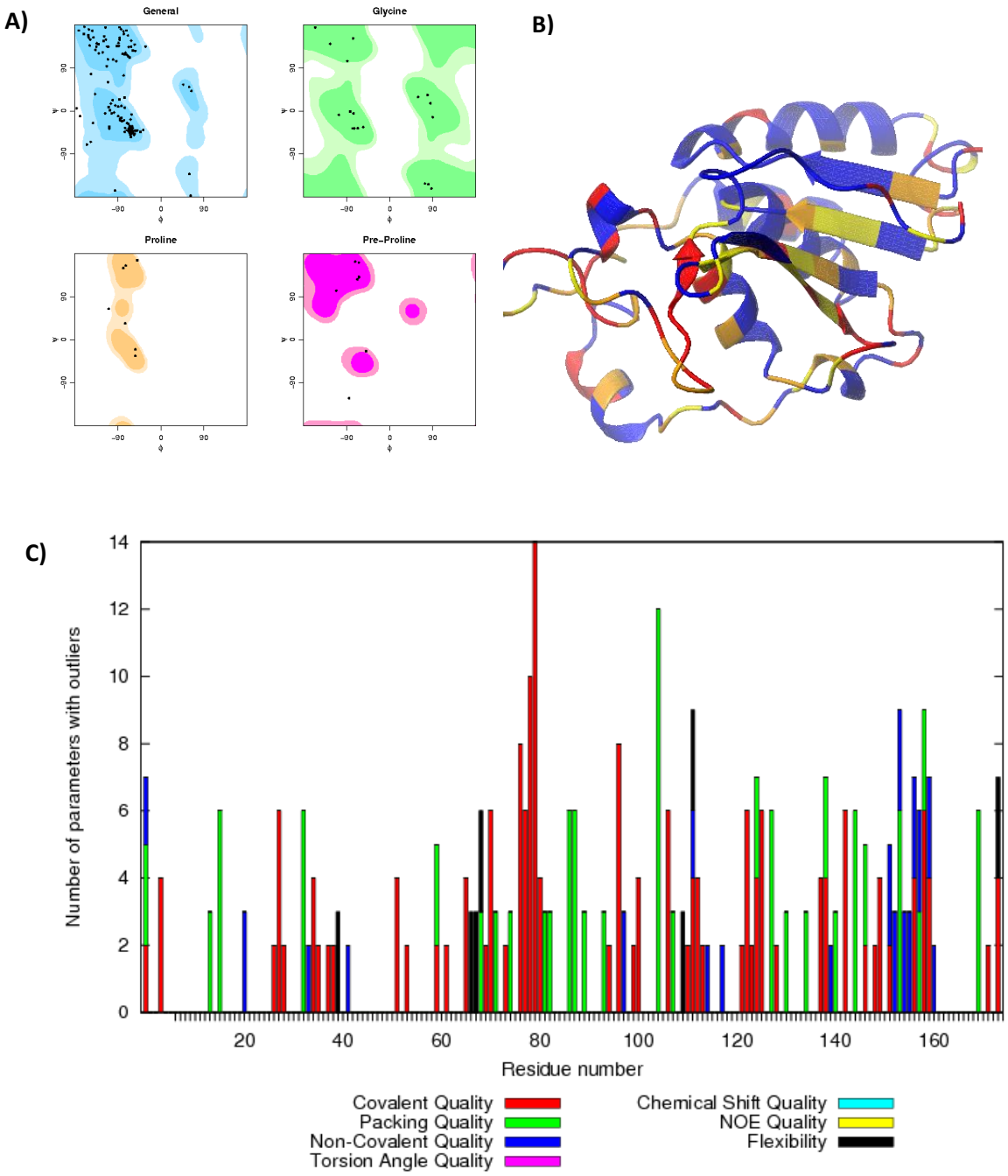


Figure 4.44 PROSESS determined model 2 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

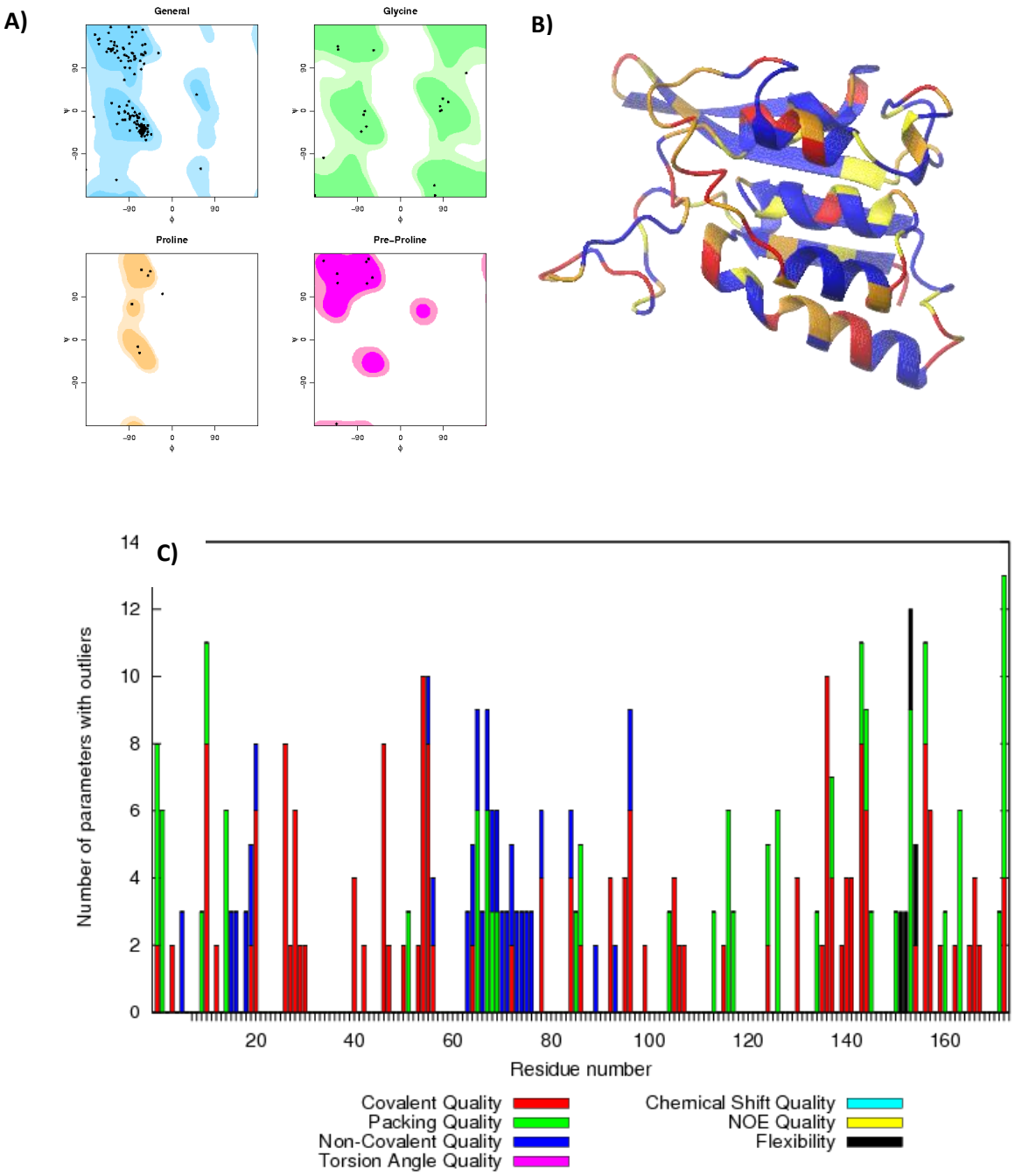


Figure 4.45 PROSESS determined model 3 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

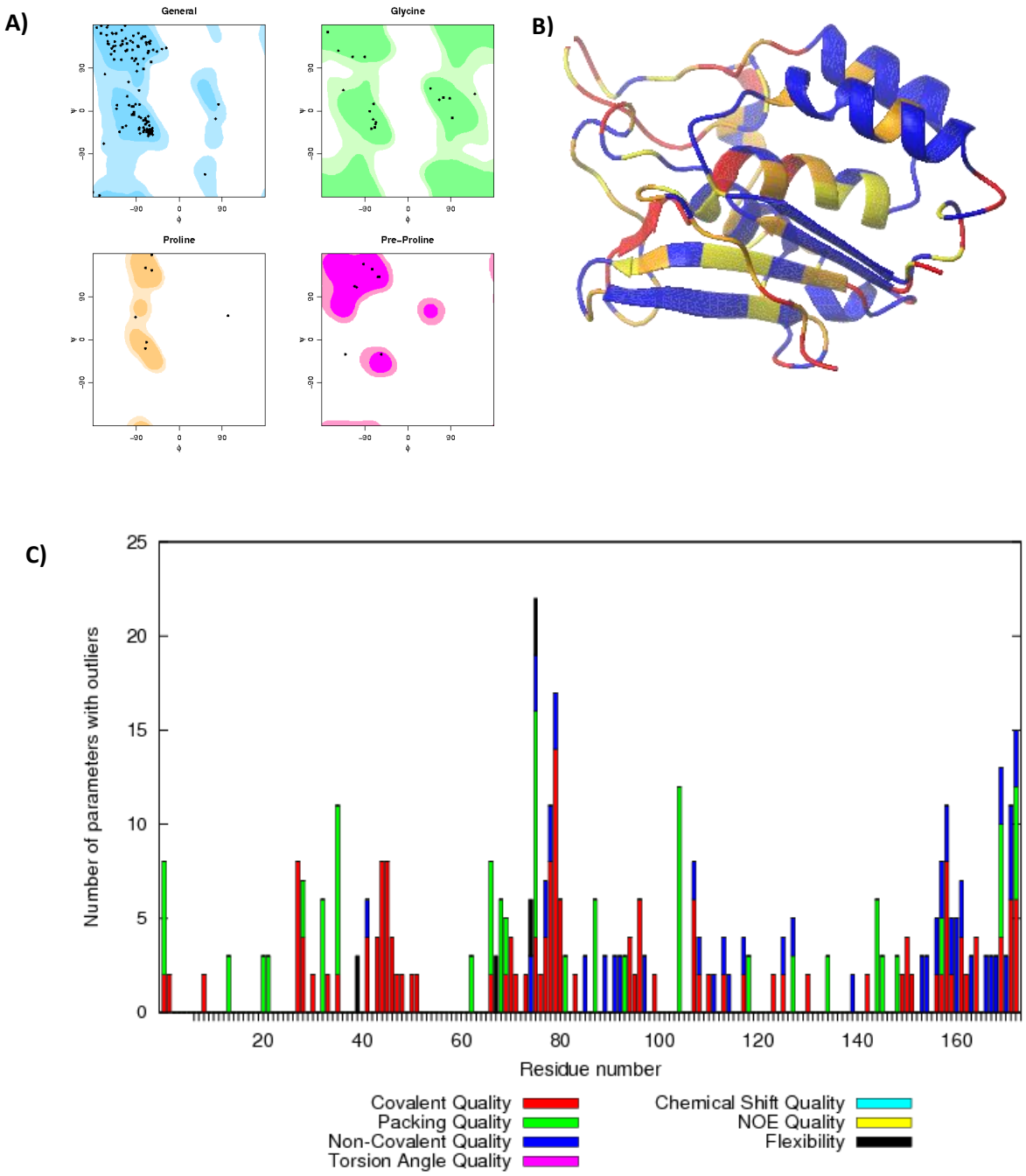


Figure 4.46 PROSESS determined model 4 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

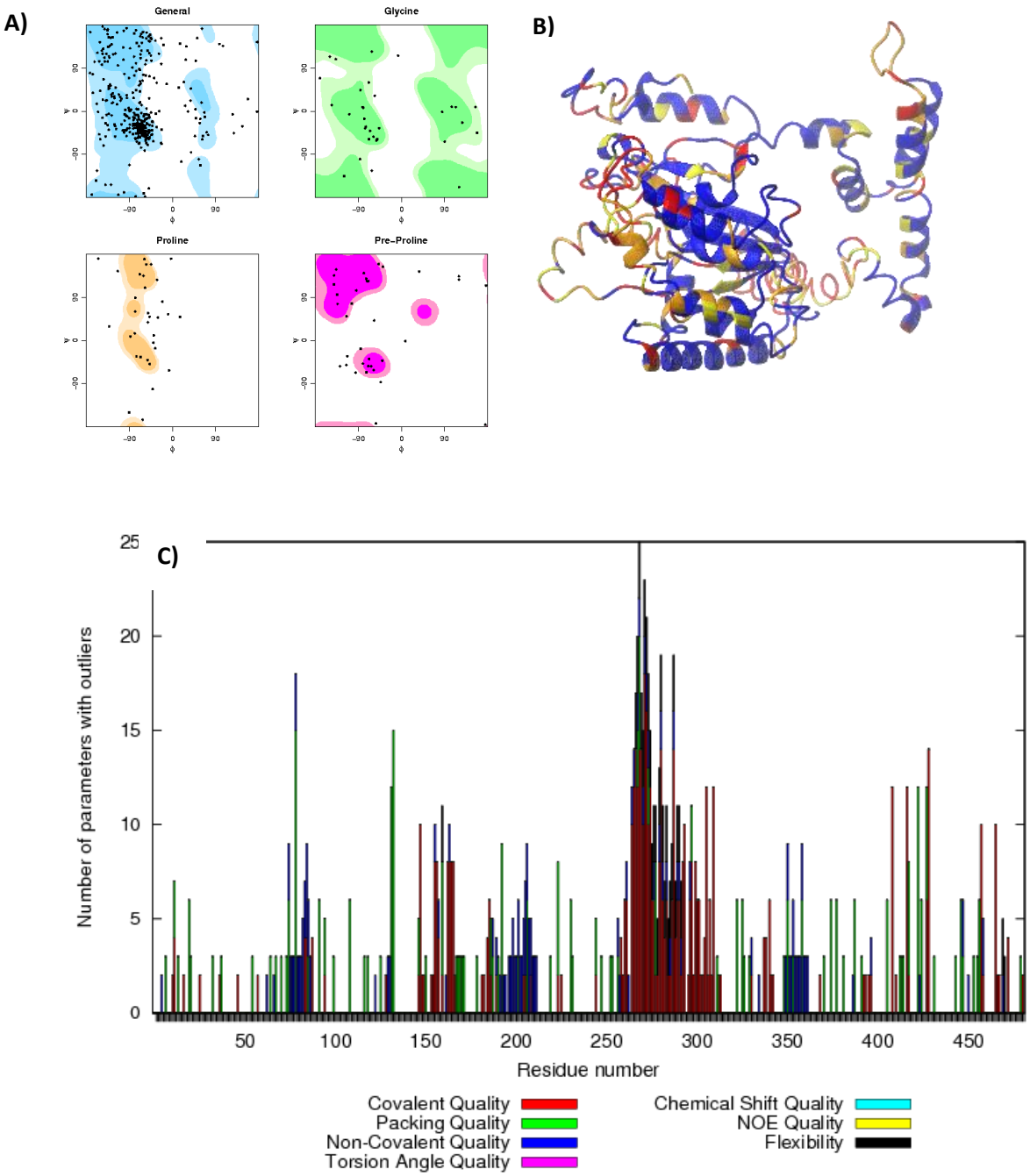


Figure 4.47 PROSESS determined model 5 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

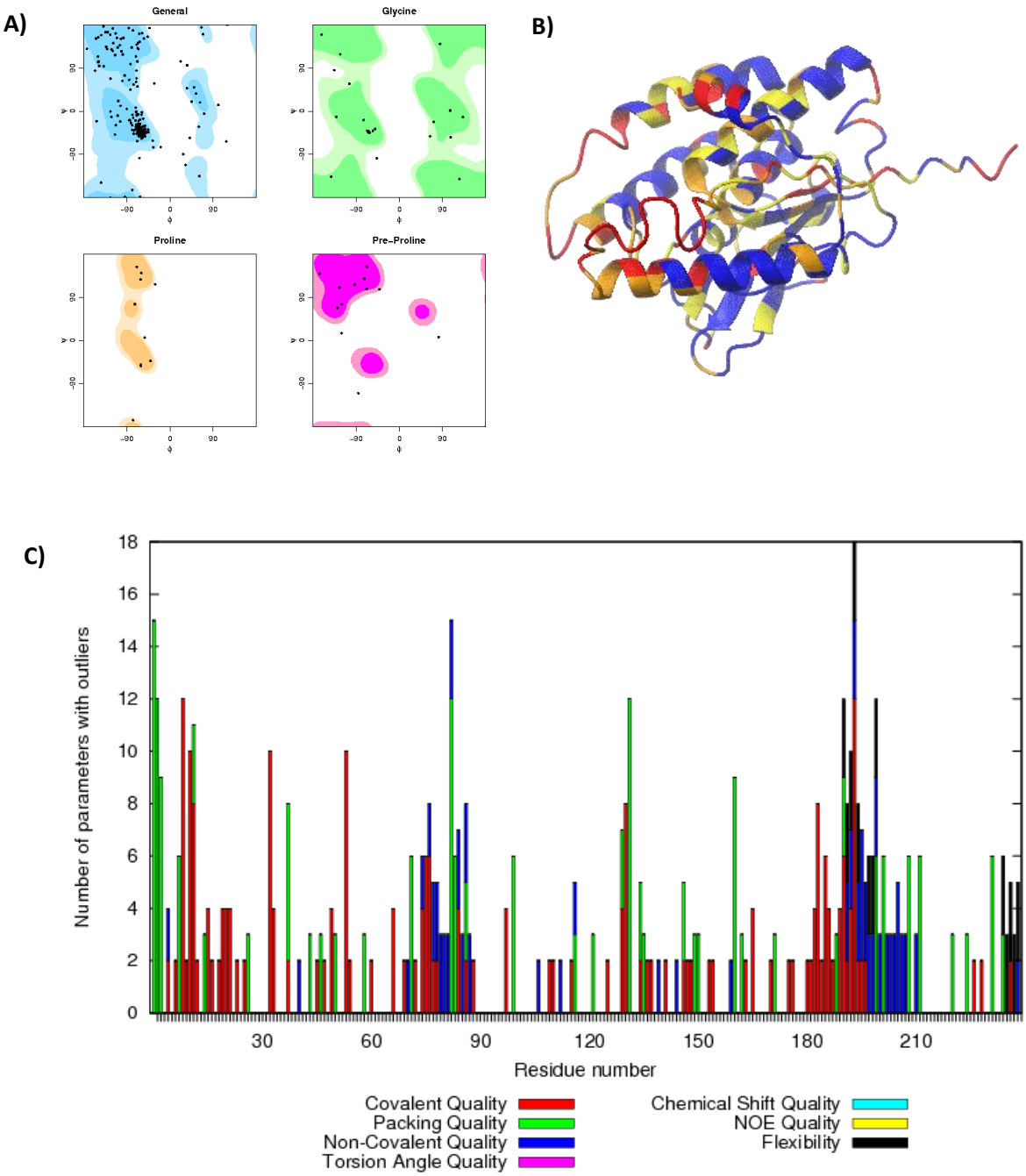


Figure 4.48 PROSESS determined model 6 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

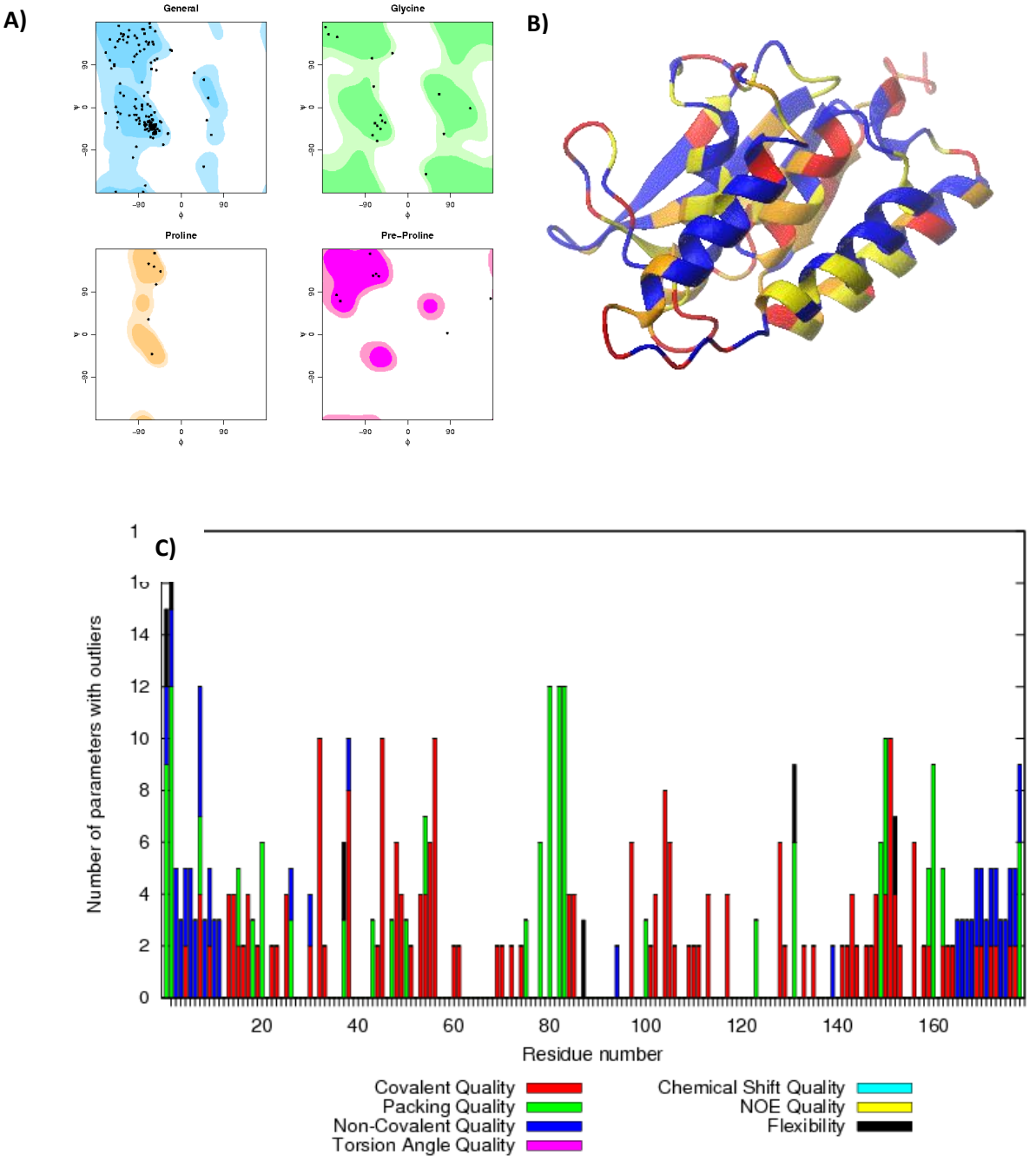


Figure 4.49 PROSESS determined model 7 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

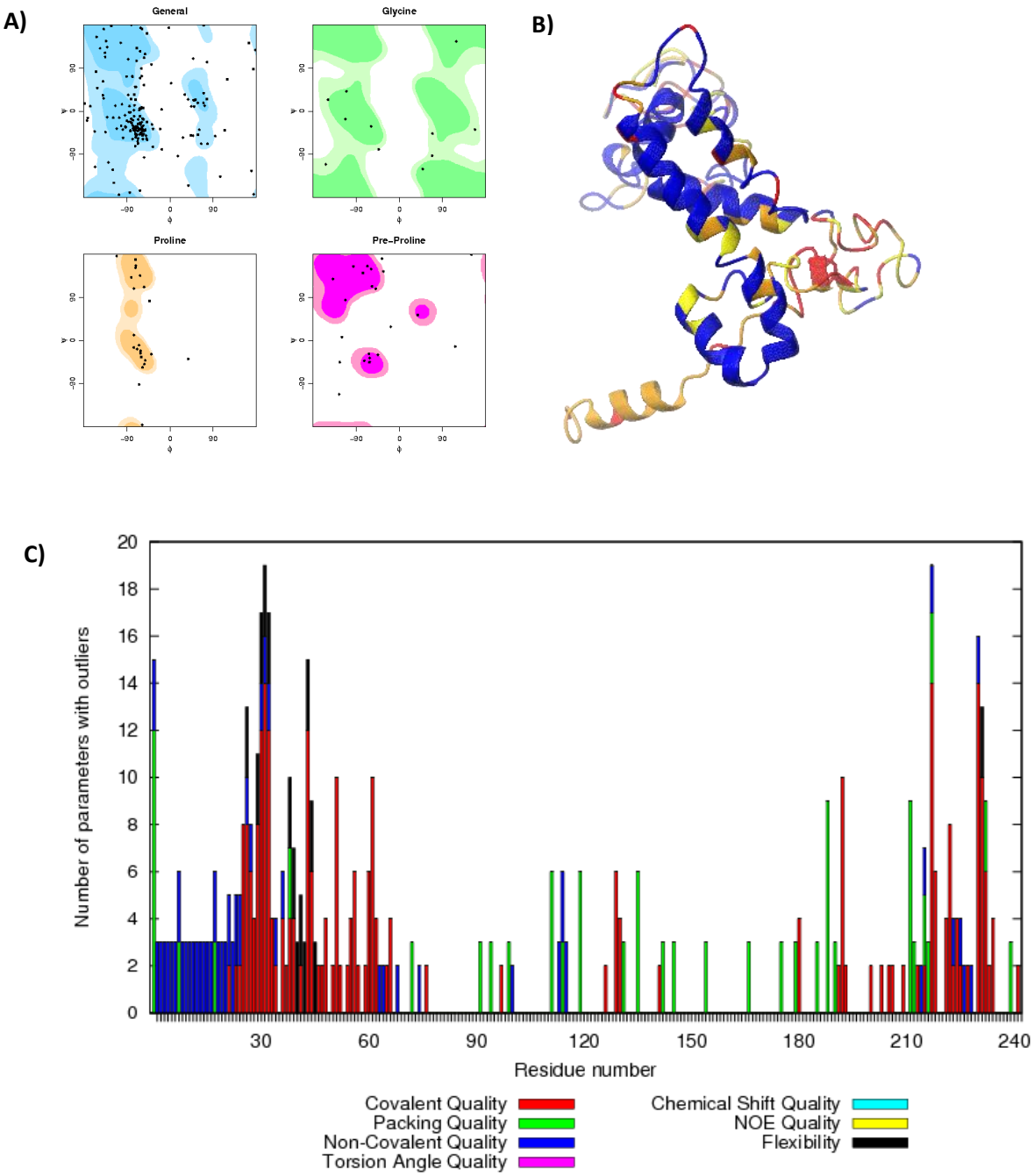


Figure 4.50 PROSESS determined model 8 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

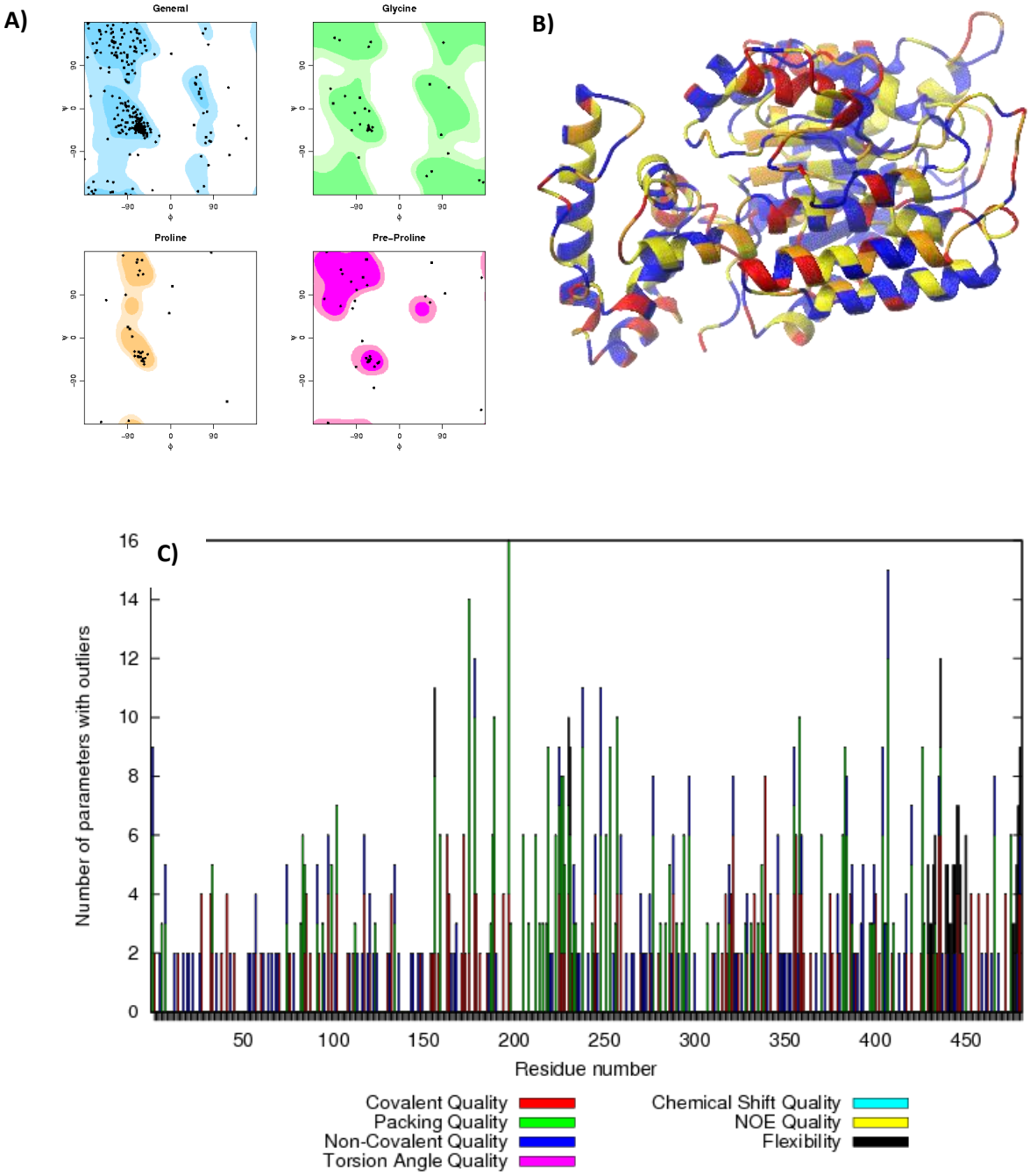


Figure 4.51 PROSESS determined model 9 residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

4.6 Discussion

4.6.1 Modelling quality

The most challenging part of any in silico investigation into protein structure based on homology modelling is determining the level of accuracy of each model. This is of particular concern when working with proteins that share low homology with known structures. Indeed, this was the limiting factor in the present attempt to develop homology models of PNPLA3.

While the N-terminal patatin domain does share some homology with other patatin-like proteins, the C-terminal residues share little to no homology with any known structures.

With this is mind, rather than relying on a single model, a range of models were generated in an attempt to attempt to find the best possible conformation for further investigation into the impact of the I148M variant on the three-dimensional structure. To mitigate the limited ability to assess the quality of the models, multiple assessment criteria were applied; Intrinsic homology was assessed as part of the modelling software internal quality assessment. The quality of the topology, which is more easily compared across models, was assessed by PROSESS.

All the initial models were predicted based upon the 148I variant. This variant was chosen as it is believed to be the ancestral variant, which likely carries higher enzymatic activity than the putatively inactive 148M variant. Although the similarity between isoleucine and methionine means that it is unlikely to significantly impact the modelling process in this instance, modelling a potentially inactive enzyme may cause the alignment to adjust to reflect the incompatible residue. It is unlikely that this impacted the quality of the models.

The range of quality scores across the generated models was wide. Thus, the overall quality score ranged from, 1.5 to 5.5 based on the topology. None of the models had a particularly high overall quality score, and all have a large number of outliers.

This is especially true in models which include the C-terminal domain, but to be expected given the low homology in this region, which highlights our need to treat these models with caution. However, these models should not be disregarded altogether, as the quality of these bond angles is not perfect even in the known crystal structures which were used as templates. For example, 10XW and 1CJY have quality scores of 6.5 and 4.5 respectively and have a number of outliers based on the PROSSES analysis (Figure 4.52-4.53).

Because I-TASSER is a more advanced modelling system which uses a combination of threading and *ab initio* model refinement, and is currently the top performing modelling software in a

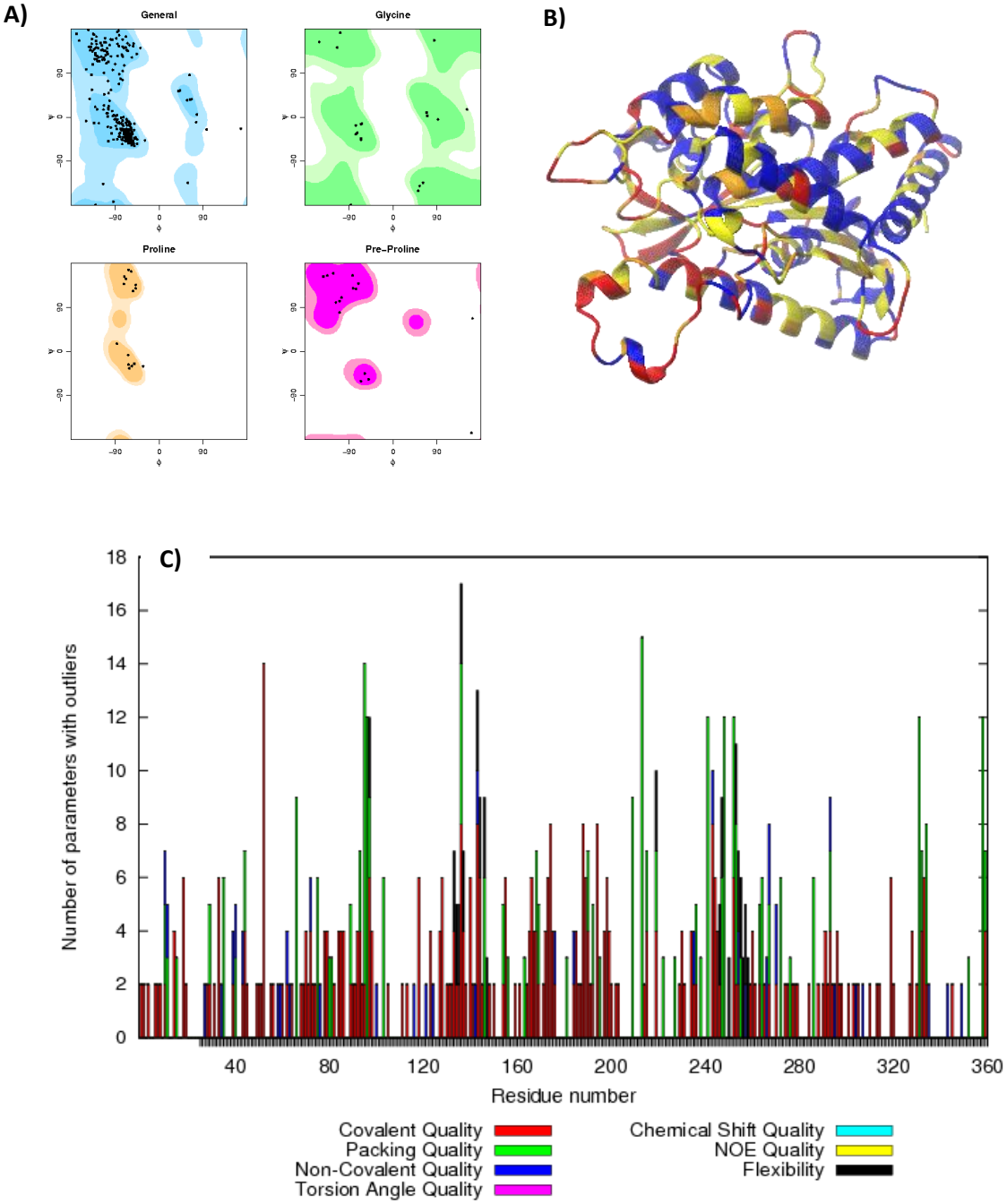


Figure 4.52 PROSESS determined 1OXW residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

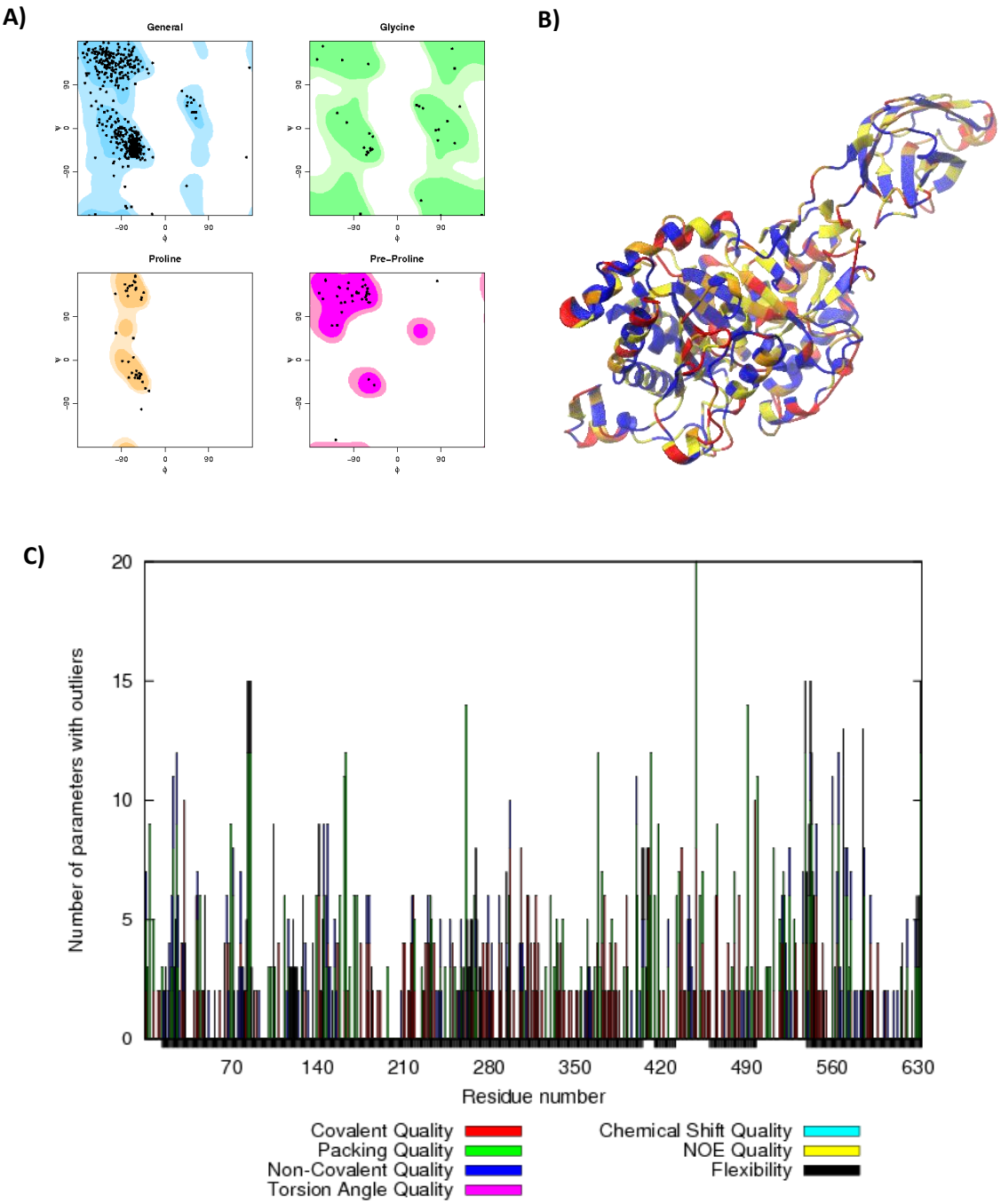


Figure 4.53 PROSESS determined 1CJY residue quality

A) Ramachandran plots showing torsional angles phi versus psi. In each plot, the accepted regions are represented as heat maps, light blue for the general case, light green for proline, orange for proline and pink for pre-proline. Each residue is plotted as a black dot.
B) Three-dimensional structure of model 1, coloured as a spectrum by overall local quality; blue being of high quality and red of low quality.
C) Number of parameters outlying the normal accepted region for each residue, colour coded by class of quality.

range of benchmark tests by CASP,[380] it was expected that I-TASSER would produce more accurate models. Interestingly the models which were generated by I-TASSER in this instance were predicted to have relatively poor quality when compared to the SWISS-MODEL results. Unfortunately, the internal quality prediction scores from each software cannot be easily compared. However, SWISS-MODEL resulted in better overall protein topology.

The fact that the more simply generated models showed better topology, was surprising. To rule out a bias caused by SWISS-MODEL only modelling the initial 170 residues, which are the most confident portion of the protein, I-TASSER was used to model these residues alone; however, still achieved poorer overall topology when compared to SWISS-MODEL.

One possible explanation for this difference in model quality it that the SWISS-MODEL approach includes an energy minimisation step for removing larger outliers, while I-TASSER only selects the lowest energy model in the cluster. Additional minimisation and dynamic simulation of the models may in fact cause these balances to shift.

Of course, I-TASSER was needed to make predictions regarding the C-terminal end of the protein, which is predicted to be of low quality simply due to a lack of homology with templates.

Despite the differences in modelling approaches and the templates employed there was a remarkable level of structural similarity between the models of the N-terminal. Thus, models 1,2,4,6 and 7 have close to identical structures derived from alignments with Pat17, VipD, VipD, PlpD and patatin respectively. This strongly suggests that the predicted topology of this domain is correct.

Reference to the conservation of residues mapped to this region shows that even disparate homologues share the highest levels of homology around the patatin domain, supporting these predictions (Figure 4.54). Furthermore, the conserved regions are clustered around the predicted active site in three -dimensional space in the present models of the patatin domain, lending further confidence to the predicted conformation.

Nevertheless, even with this high level of structural similarity between models their quality scores, range from 2.5-4.5. This implies that the quality is largely impacted by poor local topology of the side chains, rather than the overall architecture.

Further, all models have regions with increased numbers of outlier properties and a decrease in predicted local quality. These are generally located between residues 70 to 80, 150 to 160, and 250 to 300. The local quality assessments of these regions were poor with both SWISS-MODEL and I-TASSER.

MYDAERGWSL SFAGCGFLGF YHVGATRCLS EHAPHLLRDA RMLFGASAGA

LHCVGVLSGI PLEQTLQVLS DLVRKARSRN IGIFHPSFNL SKFLRQGLCK

CLPANVHQLI SGKIGISLTR VSDGENVLVS DFRSKDEVVD ALVCSCFIPF

YSGLIPPSFR GVRYVDGGVS DNVPFIDAKT TITVSPFYGE YDICPKVKST

NFLHVDITKL SLRLCTGNLY LLSRAFVPPD LKVLGEICLR GYLDAFRFLE

EKGICNRPQP GLKSSSEGMD PEVAMPSWAN MSLDSSPESA ALAVRLEGDE

LLDHLRLSIL PWDESILDTL SPRLATALSE EMKDKGGYMS KICNLLPIRI

MSYVMLPCTL PVESAIAIVQ RLVTWLPDMP DDVLWLQWVT SQVFTRVLMC

LLPASRSQMP VSSQQASPCT PEQDWPCWTP CSPKGCPAET KAEATPRSIL

RSSLNFFLGN KVPAGAEGLS TFPSFSLEKS L

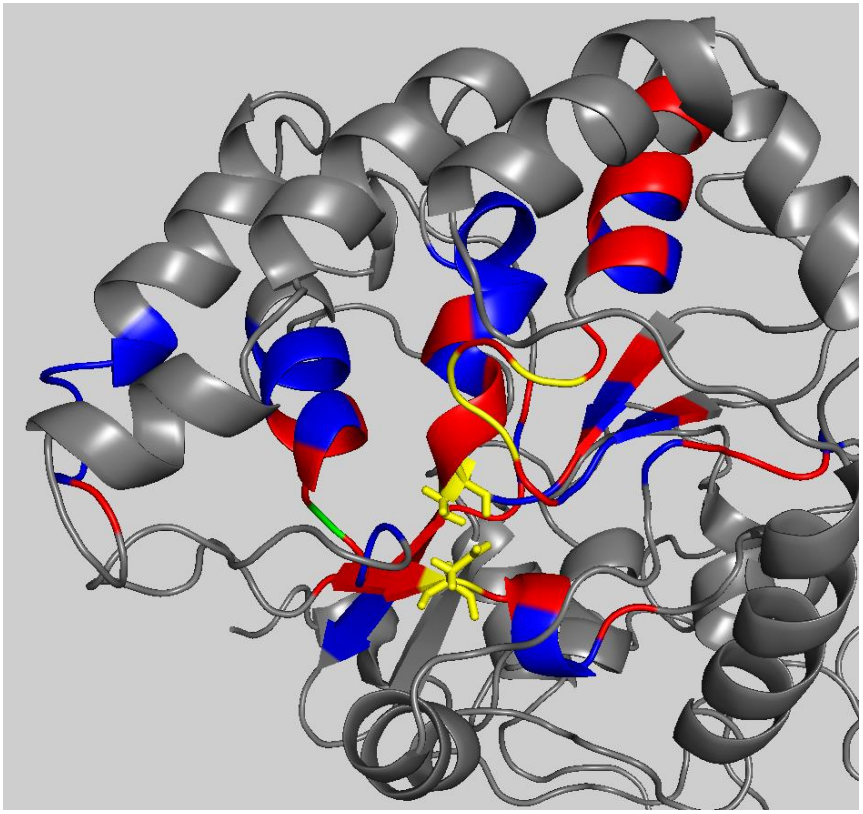


Figure 4.54 Sequence alignment and three-dimensional mapping of homologous protein sequences

Sequences were selected based on the most diverse members from the cluster of Patatin and CPLA2 superfamily sequences. Red shows highly conserved residues, blue shows low levels of conservation between sequences, grey shows no conservation between sequences and yellow represents predicted functional residues.

These regions all correspond to areas of predicted random coil or flexible loop which would be intrinsically difficult to model and may be open to large conformational changes. Residues 150 to 160 line up between two important residues, I148 and D166, which means that the location of these residues should be treated with some caution in the initial starting conformation. It is not unusual for key residues in the active site to be positioned on flexible loop regions, and the catalytic aspartate is not visible in the pat17 or ExoU structures because of this.

The longest of these regions of low quality is spans residues 250-300 and may represent a linker region between domains, which is more accommodating to variation and would therefore lose homology with distant homologues. This would align with a large region of random coil, which often forms these types of inter-domain connections.

By using an alternative I-TASSER run, an improvement in the simulated B-factors was achieved in model 9 in the N-terminal domain; however, the overall structural topology decreased in quality. In particular confidence in the C-terminal region decreased, which could be affected by skewing caused by the dominant alignments with the patatin domain, reducing confidence even further.

The model which gave the most confidence was model 3, which used ExoU (PBD ID: 3TU3) as a template. Compared to the previously published structure which was based on pat-17, this achieved an increase in overall quality from 3.5 to 5.5. It is worth noting this was based on the highest resolution template at 1.92A.

Taken together, these results suggest that overall structure of the models of the patatin domain are likely accurate, while the specific atomistic details may be incorrect. The C-terminal portion of the protein while informative should be treated with caution.

It is likely the long region of low confidence between 250 and 300 represents a linker region between domains, but the specific cut-off between the patatin domain and C-terminal is unclear. In particular the regions 70-80, 150-160, 250-300 should be regarded as uncertain and flexible, not as confident static entities in the models.

4.6.2 Structural features of PNPLA3 models

4.6.2.1 Structure of the patatin domain

The models which were generated are highly influenced by the homologous protein structures that provide a starting point for threading the PNPLA3 models. Since the models presented are often based around patatin and CPLA2 structures we would expect to see consistent structural trends between these proteins.

The structure of patatin (PDB ID: 1OXW) represents a classical α/β class protein fold. This consists of a core of 6 β-strands, 5 of which are parallel strands and one antiparallel strand toward the periphery. This β-sheet core is sandwiched between α-helices, on either side, creating an α/β/α structure. This shows clear conservation with the classic α/β hydrolase fold (Figure 4.55).

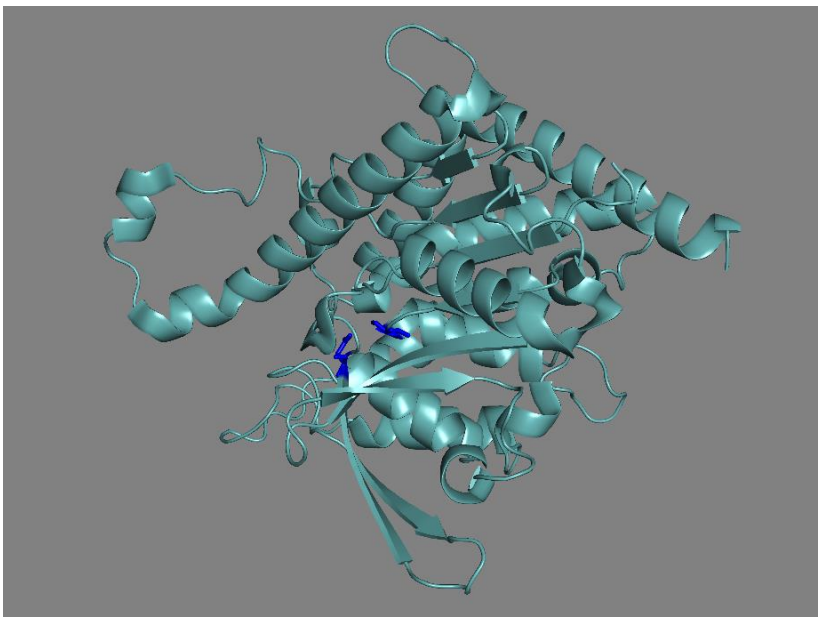
Cytosolic phospholipase A2 (CPLA2) another human phospholipase (PDB ID: 1CJY), also shows a similar α/β/α structure, with more structural complexity complicit with the need for interfacial activation, showing that this fold is conserved into the human proteome.

PNPLA3 models 1,2,4 and 7, were most structurally similar to this patatin fold, and all represent a very similar core. Models 2,4 and 7 have 3 parallel β-strands in the core and 2 antiparallel strands. While model one also had an additional parallel strand. Thus, Model 1 shows greatest similarity to the classic patatin domain and is consistent with the previously published model which has been used for earlier computational observation.[158,412]

One thing that is notable about these structures, is that in patatin and CPLA2 the β-sheets are sandwiched between helices on either side. However, in each of these PNPLA3 models, the helices only have structural support from helices on one side, while the other side is solvent exposed.

Model 6 also maintained a highly similar core to patatin; however, while models 1,2,4, and 7 were only constructed to included residues up to residue 179, model 6 was constructed to residue 239. These additional residues formed an additional 2 helices, helix F and helix H, which sits on the other side of these β-sheets offering the expected additional structural support.

Helix F appears to be a long helix which fits directly into the structure, while helix H is at the periphery of the protein. It could be that having these additional helices on this side of the protein might be important for stabilising the core patatin structure and could act to partially shield the highly hydrophobic active site pocket from hydrophilic solvent.



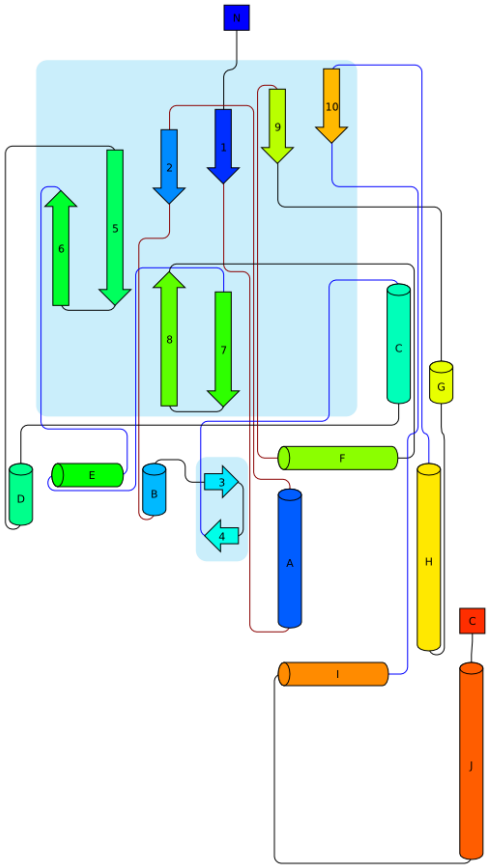


Figure 4.55 Three-dimensional structure of 1OXW

Top panel: The β-sheets are represented as arrows; the α-helices are represented as cylinders. Residue 148 is highlighted in red and the catalytic dyad (serine and aspartate) in blue.
Bottom panel: Richardson diagram of the structure coloured from the N to C terminus. β-sheets are represented as arrows and α-helices as cylinders. The blue panel represents spatially conserved β-sheets.

This stands out as a key structural feature which was missing from previous predictions of the patatin domain and could have significant impact on the interpretation of results because of the local impact on the active site.

4.6.2.2 Structure of C-terminal domain

Only two models were generated that spanned the full length of the PNPLA3 protein, Models 5 and 9. Interestingly, these models clearly reinforce the prediction that PNPLA3 is a multidomain protein, with the C-terminal forming a distinct unit.

Both models have an identical β-sheet core in the patatin domain, consisting of four parallel and one antiparallel β-sheets, which again aligns well with the patatin fold. However, in both models the C-terminal domain contained no β-strands. This contrasts notably with the C-terminal domains of aligned structures and implies that the alignment is probably poor. These models also do not align well with one another although their overall structure is similar. The most noticeable difference is that the C-terminal region is less tightly packed in Model 5 than in Model 9.

When modelling multi-domain proteins, the quality of the model can often be improved by modelling each domain individually. This is particularly important for the domain with less homology which is more easily skewed by the alignment with the other domain. Although I-TASSER uses a threading process with some automatic domain separation, this could still influence confident clustering biased toward this domain.

Model 8 was the only model which was constructed solely of the C-terminal domain and interestingly, also has highest similarity with ExoU (PDB Id: 3TU3). In contrast to Models 5 and 9, which encompassed the C-domain this model predicted the presence of two short antiparallel β-strands toward the end of the domain. However, this still contrasts with the structure of the typical α/β hydrolase or CPLA2, which both have a significant number of β-sheets throughout the entire length of the protein.

The overall lack of homology with the C-terminal domain, and poor topology of the models, show that we should be highly cautious before relying on the model of this domain. However, because of the vast difference we see with other known proteins within this class, it could be postulated that this region has a unique function, and potentially a novel protein fold, which is functionally independent of the C-terminal regions of patatin, CPLA2 and other PNPLA enzymes.

4.6.2.3 Putative lid region

CPLA2, unlike patatin has a flexible lid region, which is not part of the protein core. This region consists of four β-strands and five α-helices which form a cap around the active site (residues 370 to 548), of which a smaller section is deemed to be the actual lid (residues 413 to 457)[413] (Figure 4.56). This lid region is imperative to the interfacial activation of the protein and is thought to not exist in patatin because of no obvious parallel in the structure and a lack of this mechanism.

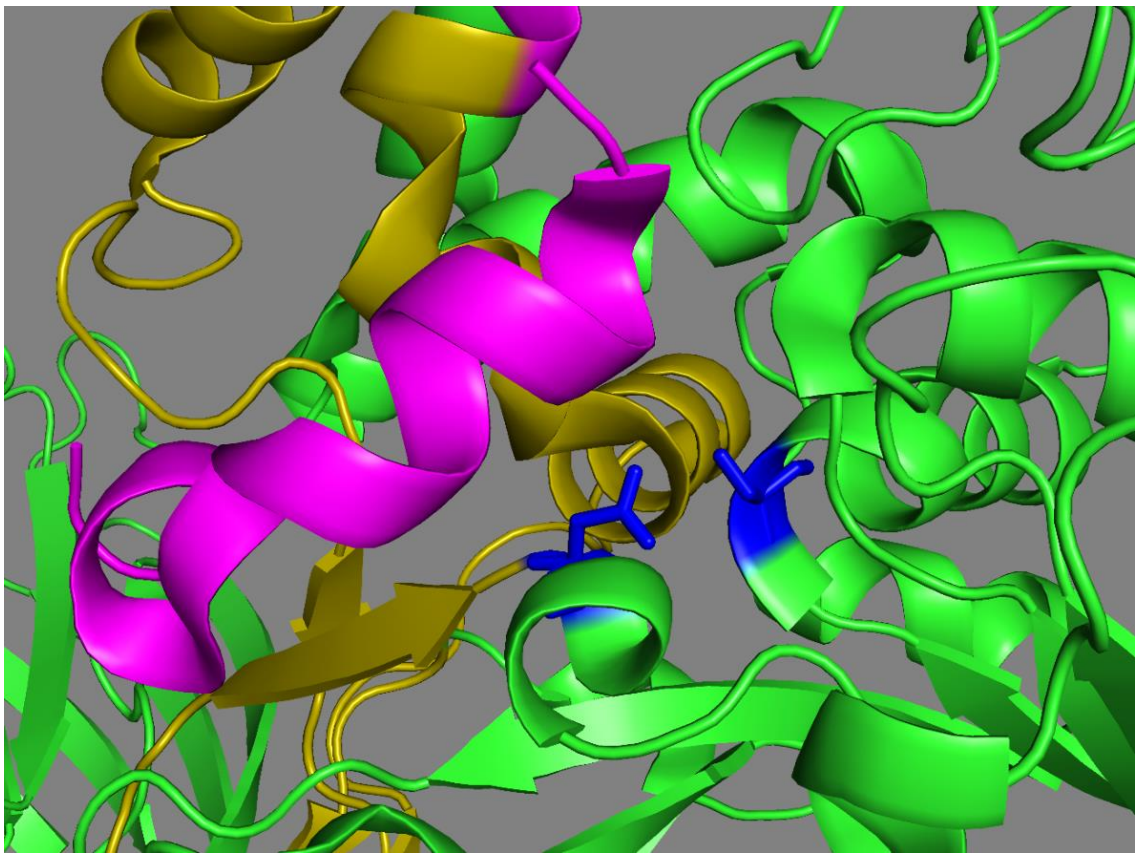


Figure 4.56 Active site residues of CPLA2 (PDB ID: 1CJY)

The catalytic serine and aspartate are highlighted in blue (Serine 228 and aspartate 549). The main backbone is highlighted in green. The cap region (residues 370-548) is highlighted in olive and the lid (413-457) in magenta.

There is no evidence of a similarly large lid region in any of the models of PNPLA3, as would be expected since the protein is considerably shorter. Where CPLA2 has a range of additional helices after the catalytic aspartate, PNPLA3 goes straight into the next core β-strand, in a fashion similar to patatin. However, there is a high concentration of hydrophobic residues

around the region of the active site which lend credence to the possibility of interfacial activation; supported by the fact PNPLA3 is known to associate to lipid droplets.

There are two regions of the modelled structure which cover the active site and could contribute to a lid region; however, they would be significantly smaller than in CPLA2.

In the models with a conserved active site (models 1, 2, 4, 6 and 7), the loop which contains residue 148 and 166, acts to cover a portion of the active site. While in most of these models this is insignificant, in model 6 where the active site is more shielded, this could provide a significant barrier to the active site. Flexibility in this loop could allow for a lid function and putative interfacial activation.

Furthermore, in model 6, the additional helices at the C-terminal end of the patatin domain bridge over the active site pocket, and because of the lack of an extended helix network, could also perform a lid like function. This would still form a much smaller cap than observed in CPLA2; however, the position of this lid and similarity to CPLA2, makes this a convincing possibility.

While in no way conclusive, this further supports the importance of residues 179-239 to be included in any further research.

4.6.2.4 The active site

Models 1, 2, 4, 6 and 7, predict that both the catalytic residues (S47 and D166) and residue 148 lie within a catalytic pocket, in the same position. This places the catalytic residues within 4.2 angstroms from one another, a distance suitable for reaction. This provides confidence in both the previous and present models of the active site. Models 3, 5 and 9, in contrast, predict that the catalytic residues, while located in the same region of the protein are further apart; in Model 3 this distance amounts to as much as 17 Å.

In all models, the catalytic serine was positioned in a turn between a small elbow hinge region between strand 2 and Helix B. This is highly conserved spatially, between all the canonical α/β hydrolase fold containing enzymes, including patatin and CPLA2 (Figure 4.57) and so we can treat the location of this residue with high confidence.

The catalytic aspartate (D166) is positioned on a coil region which is not modelled with high quality and so the true position of the coil is not clear. It is reasonable to assume that the correct position maintains a conserved active site with the catalytic residues in close proximity; however, this cannot be decided based on these models alone.

In Model 3 which has the highest quality prediction based on the topology of the protein, the distance between S47 and D 166 is 17Å clearly too far for a reaction to occur. This raises the question as to how biologically relevant this model can be.

Notably this model does still place Asp166 directly on a flexible loop. It is not unreasonable to believe that PNPLA3 may have an APO and HOLO conformation, and model 3 may represent an APO conformation, while the other models HOLO. This loop could potentially reposition to form a stable active site under the right conditions, for example upon binding of a ligand or binding partner.

Looking in more detail, D166 is positioned on a coil region which although not clear by visible inspection, is predicted by STRIDE[411] to be on a small β-strand and in model 6 this β-strand is clearly visible.

It is notable that in model 3, there is a complete loss of this strand (strand 5) along with the loss of active site structure. Similarly, where this strand is less formed, and shorter in models 5 and 9, the distance between catalytic residues increases. This suggests the formation of this strand is important to the overall structure.

In both patatin and CPLA2, the active Aspartate residue lies at the terminus of the last β-strand which lies outside of the core β-sheet segment, while in ExoU for example the flexible loop in this region is not visible in the structure. This supports the possibility of conformational shifts, occurring primarily in this region of the coil.

If models 3, 5 and 9 are accurate then that the position of this key residue is more flexible than in patatin. This could result in an overall increase in active site flexibility, and in turn cause a lower overall rate of reaction. This may explain, in part, why the activity of PNPLA3 is significantly lower than the activities of other phospholipases such as CPLA2. However, it might also confer the protein with greater ability to support a broader range of complex substrates.

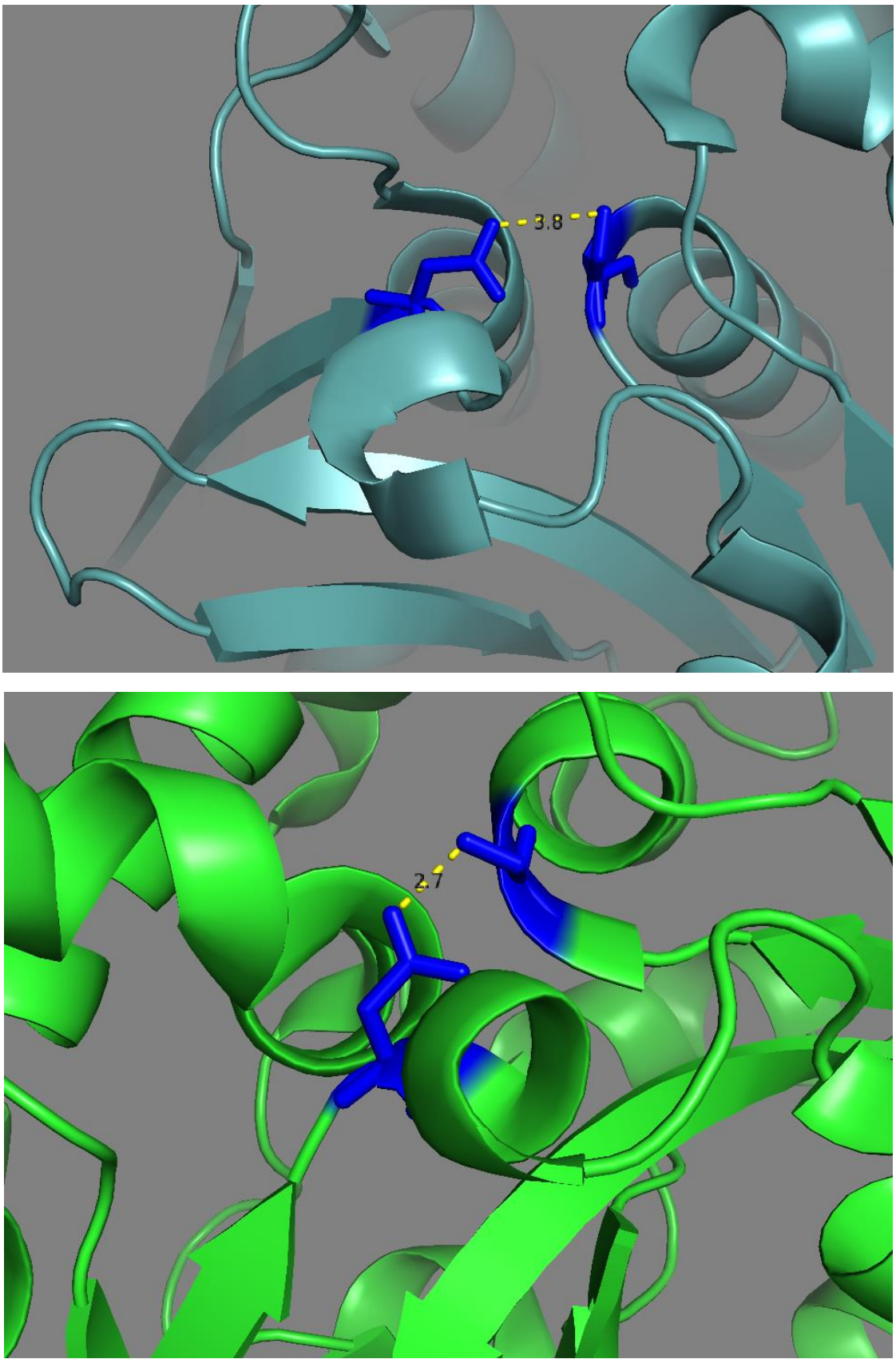


Figure 4.57 Active site residues of 1OXW and CPLA2

The catalytic serine and aspartate are highlighted in blue. The distance between residues, measured in Angstrom, is shown by dashed yellow line.

Top panel; 1OXW.

Bottom panel; CPLA2.

4.6.2.5 Oligomerisation

One of the issues with homology modelling is that the product generated is viewed as a monomer, while many of the templates used to generate the models are in a range of oligomerisation states. For the model, itself, this is not a major issue, as this should not affect the static structure. However, for further investigation using molecular dynamics the oligomerisation state could be important to the protein stability and structure.

There is a large body of evidence that patatin-like proteins often require binding partners for activity (VipD with rab5[414], ExoU with ubiquitin[415], PNPLA2 with cgi-58[416]), and although no specific evidence has highlighted a binding partner for PNPLA3, there is a strong possibility this will be identified in the near future.

All of the templates used were observed in multimeric forms within the crystal structure. The patatins were observed as homo-mers and the bacterial proteins hetero-mers.

There were concerns that the multimeric state of these templates could lead to poorly performing dynamic simulation of monomeric protein subunits. However, although the crystal structure of patatin (1OXW) displayed a stable trimer confirmation, this is likely not biologically significant, as dynamic light scattering experiments showed the protein in solution to be monodisperse, and the conformation was simply a result of crystal packing.

VipD has phospholipase activity, which only becomes active upon binding with a mammalian host protein Rab5 [404]. Binding to Rab5 causes a flexible loop to move, opening the active site pocket. There is the potential that PNPLA3 would need to undergo similar domain movement upon binding to a protein partner to become active.

ExoU similarly relies heavily on binding partners, needing SpcU which was observed in the template for excretion and ubiquitin binding for activity.[417] It is notable that in all of the above cases, binding partners interact with the C-terminal domains.

When these partners occur in the template, this will influence the structure in PNPLA3, but could also suggest the need for a binding partner to maintain stability in the C-terminal domain. It is difficult to determine whether the similarity in this domain is caused by alignments themselves biasing the PNPLA3 predicted structure, or an effect of true similarity.

Although the patatin domain of PNPLA3 is known to associate with lipid droplets,[158] it is also possible that the C-terminal domain plays a secondary role in this interaction, whether directly, or through interaction with a binding partner. It is possible that the previous observation of the patatin domain in lipid droplet fractions were caused by other factors for example incomplete

lysis of cells, or the highly exposed hydrophobic residues on the surface causing aggregation; this hydrophobic region is not only a good potential interaction surface for lipid droplets, but for one another, and inclusion bodies and larger aggregates could be expected which may also fractionate to lipid droplet containing fractions.

As no information was available on possible binding partners for PNPLA3, no attempts to model in partners were made to avoid confounding error at this stage. It is highly likely that this will impact the C-terminal region in particular, and it is expected binding partners will be identified in the future, as large flexible loops in the region would act as very good sites for binding.

4.6.3 impact of the I148M variation

The PNPLA3 variant is associated with the development of liver disease as well as its progression and even ultimate survival. This makes delineating the structure and function of this variant of prime importance.

While this change is believed to result in a loss of function, isoleucine is not believed to have any catalytic role in the protein because the isoleucine side chain is very non-reactive and is thus rarely directly involved in catalysis. However, it can play a role in substrate recognition and in particular can be involved in binding/recognition of hydrophobic ligands such as lipids. Methionine is also a hydrophobic residue and although slightly less hydrophobic than isoleucine, the small degree of change in hydrophobicity is unlikely to be significant.

The current hypothesis on the impact of this variant, based on previously generated models, is that the change from isoleucine to methionine creates steric hindrance at the active site by protruding further into the active site pocket. This would serve to explain the loss of catalytic activity observed in this variant.

In the present models, this change appears to have no clear impact on access to the active site. The residue does protrude into the active site, but only marginally, but rather more importantly it also forms closer contact with nearby residues, for example Y151, which will inevitably change the active site structure potentially opening the active site even further. Overall this type of steric hindrance seems unlikely to play a role in the loss of activity and cannot be determined by observing the structure alone without dynamic information.

It would seem much more likely that the effect on catalytic activity is caused by the change in the biochemical environment affecting the structure of the active site, and its ligand binding

ability. There are several changes in biochemical properties which may cause changes within this conserved region.

First, isoleucine, like valine, and leucine is a branched chain amino acid. Whereas methionine contains only one non-hydrogen substituent attached to its C-β carbon. This creates more bulkiness near to the protein backbone of isoleucine; thus the surrounding amino acids are more restricted in the conformations they can adopt. Clearly the loss of this stability could enhance the flexibility of the loop which is structurally important for the position of D166.

Second, Model 6, like the 1OXW, has additional helices looping over the active site. It is possible that, in this model, the methionine variant has a different interaction with these helices. In particular, it may form a stronger hydrophobic centre with L212 with which, because of side-chain positioning, it is in closer proximity. This could either stabilise or transition the helix closer to residue 148, reducing access to the active site, or else it could repel it opening the site further and disrupting the conformation. This could lead to a change in catalytic activity depending on the nature and extent of the interaction

Finally, the change to methionine in this region will result in close contact between two aromatic residues in the local chain, P149 and Y151. These residues form part of a loop which curve around residue 148, and in effect shield the active site pocket. This loop is clearly a key element in determining access to the active site and the position of the catalytic residues.

The change to methionine in this region, will cause a greatly enhanced interaction between these aromatic residues as well as removing additional steric hindrance which forces them apart in the model of the wild type protein. This action will inevitably change the properties of this chain, especially because compared with a purely hydrophobic interaction, the Met-aromatic motif yields an additional stabilization of 1–1.5 kcal/mol.[418] This arises as a result of a unique single sulphur-aromatic interaction, which is rarely considered and leads to stronger bonds at longer distances. The preferred intermolecular distance for this interaction is 5.5A between the sulphur and the ring centre, and there is an orientational preference of 30 to 60° between the sulphur and the normal vector defined by the plane of the aromatic ring.[418]

The energy associated with a single sulphur-aromatic interaction is comparable with the interaction energy of a single salt bridge. However, the sulphur-aromatic interaction also occurs at a larger distance (5-6A) than a salt bridge (typically 4A) and may be less sensitive to local environmental changes, including the pH values of the solvent and the solvent itself. Additionally, salt bridges have a significant energy penalty associated with side chain desolvation; however, the nonpolar nature of the methionine and aromatic residues minimizes

the desolvation energy penalty. The combination of these attributes makes this a significant interaction to consider within the context of the PNPLA3 I148M variation.[418]

Approximately one-third of all known protein structures contain an energetically stabilizing Met-aromatic motif and mutations involving methionine are associated with a number of pathological conditions, including Alzheimer, Creutzfield-Jacob, and von Willebrand disease, putatively because of changes in these Met-aromatic interactions.[418]

In all of the models generated in this study, the impact of these stabilising or destabilising interactions has the potential to effect catalytic activity, although the specific mechanism would differ, depending on which model structure is considered.

In Models 1, 2, 4, 6 and 7, where residue 148 is positioned close to the active site pocket, destabilising the loop could lead to a loss of the conserved catalytic active site, and therefore a loss of catalytic activity. However, if we hypothesise that the models with more distant positioning of the catalytic residues represent an APO conformation, stabilising interactions caused by the methionine in this position, could lead to increased activation energy and similarly reduced activity. In Model 3, there is a close proximity between F159 and M148, which could facilitate a stabilising interaction as described above. In this case, the additional stability of this region could increase the activation energy needed to pass through transition state, for example based on interfacial activation, in which the loop would be displaced for the aspartate residue.

While it is clear the I148M variant will have an impact on the structure of the active site this cannot be determined based on visual observation alone. It is likely, given the many possible effects of the I148M variation, that the impact is not caused by any one single factor, but a complex effect which strongly impacts the stability of the active site.

4.6.4 limitations to structural modelling

There are a number of limitations to the modelling procedure used in this study. First to create confident protein structural models at least 25% homology with another known structure is needed. This was not available for PNPLA3 making it difficult to create models with acceptable confidence. In particular, the C-terminal domain had almost no homology with any known structures, relying almost entirely on a threading with small fragments and ab initio structural prediction. Thus, the quality of the models was difficult to assess.

An automated server-based modelling approach was taken rather than an expert driven approach. The automated system produces less biased results, but could potentially reduce the accuracy of the result in such a difficult target.[387] It is debatable whether this is the most appropriate approach. However, based on the lack of confident experimental data, it was predicted that the potential to incorrectly bias the model was greater than the ability to improve the accuracy.

It has become increasingly possible to improve models through the use of a range of energy minimisations and dynamic simulations. The models were used as produced and no additional refinements were performed. This was because it is likely certain regions of the protein are not confidently modelled and these steps may distort the protein further from the native state and allowed a fair comparison of all the models generated; however, that the models may not be optimal and could be further refined.

As discussed throughout the chapter, there is difficulty assessing the quality of the models. While we attempted to use as many measures as available to help elucidate regions of the model with higher and lower confidence scores, the true accuracy of the structure cannot be assessed without experimental validation.

Finally, while the primary interest in this model was to assess the impact of the I148M variant, the interpretation of this change is extremely difficult to predict by visual inspection alone. While a range of possibilities have been discussed, it is clear that a static model does not provide enough insight into the active site and I148M to make confident claims regarding this structure.

4.7 Conclusion

Key findings:

Support for previous findings:

- Similarities between the 9 models support the structure of patatin domain.
- Ser47, Ile148 and Asp166 are predicted to be spatially conserved.

Novel findings:

- The overall structure of the models of the patatin domain are likely accurate, while the specific atomistic details may be incorrect. The information on C-terminal portion of the protein, while informative, should be treated with caution.
- Different local conformations around active site suggest APO and HOLO conformations.
- A full-length model of PNPLA3 was predicted, showing PNPLA3 is likely to be a multidomain protein.
- Two additional helices have been identified as a key part of the patatin domain.
- The I148M variant seems unlikely to inhibit the enzyme via steric hindrance.

In summary, despite difficulties generating high quality models because of low homology, models were deemed to have sufficient quality to allow structural insight into the overall architecture of the PNPLA3 protein.

Nine molecular models of PNPLA3 were generated, including the first full length model of PNPLA3. The consistent architectural similarities between models lend support for the overall protein structure of the patatin domain and show that both I148 and D166 lie on a flexible loop region near the active site pocket. Multiple conformations adopted by different models suggest a potential APO and HOLO conformational shift, by which the predominant shift is performed by this flexible loop.

The I148M variant, which is a significant risk factor for the development and progression of chronic liver disease, does not appear to inhibit the enzyme through steric hindrance to the active site pocket as previously predicted, and likely functions through more complex chemical changes in the local environment, most likely effecting the energy state and in turn conformation of the D166 containing flexible loop.

This is the first report of two additional crucial helices which appear to form an integral part of the patatin domain. This leads us to redefine the patatin domain from residues 1-239.

The prediction that PNPLA3 is a multidomain protein was supported by the models generated, with a linker region approximately spanning residues 250-300. Through observations of

disordered regions and similarities with homologues, it is likely that PNPLA3 has important protein binding partners, which may be needed for proper protein function.

In order to further investigate the structure of PNPLA3 using these models, dynamic simulation will be needed. This will help to better understand the impact of the I148M variant.

Chapter 5

Dynamic simulations of PNPLA3

"I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me."

Isaac Newton

5.1 Overview

The impact of amino acid substitutions on a proteins function relies not only the change in the local chemical environment, but also on the intrinsic flexibility and in turn dynamic movement of the protein within the cell.

The dynamic motion of PNPLA3 has undergone little investigation to date, despite its importance in understanding the function of the protein and impact of the I148M variation, in part because of difficulty generating quality models of the protein.

In this chapter the implication of the I148M variation on the dynamic motion of PNPLA3 are investigated using molecular mechanic simulations of the PNPLA3 models generated in the previous chapter.

The full-length model of PNPLA3 (model 5) remains stable over long-time scale simulations of 100ns, adding confidence to the initial structural model.

A significant difference in the active site architecture is observed between isoleucine and methionine variants of this model. Of note, during the simulation of the methionine variant, the catalytic residues are separated, disrupting the otherwise stable active site. This suggests a novel putative mechanism by which the I148M may cause loss of lipase activity.

5.2 Introduction

Proteins are highly mobile structures which undergo frequent and large conformational changes. There are a variety of motions that this can entail, ranging from large scale domain movements, local rearrangement of residues around a ligand, and small organic shifts caused by side chain rotations which occur on a much shorter timescale. These movements can occur because of ligand binding, changes in local chemical environment, molecular interactions, or due to the basic thermal dynamic motion of the protein.

It is important to note that the dynamic fluctuations in the protein structure are not simply a superficial by-product of the structure itself. It is becoming increasingly clear that these fluctuations are vital to the underlying protein function, and it is likely that these motions are at least as important as the overall three dimensional structures in facilitating these functions.[419] In fact, it is the rate of conformational change in many proteins that often presents the bottleneck in catalytic conversion from substrate to product.[420–423]

Because of the key role that protein flexibility and dynamics are postulated to play in catalysis, it is reasonable to assume that proteins not only have optimum 3-dimensional structures, but also an optimal dynamic state for catalysis. This would energetically balance the structural flexibility needed for catalysis, and the stability to maintain conformations available for binding.[424]

Allosteric effectors, including the traditional binding of ligands in allosteric binding sites, but also mutation in sites which have allosteric effects, can result in the redistribution of the proteins conformational ensembles, and alter their rate of interconversion thereby modulating active site or binding site geometries either toward or away from this optimal dynamic state.

In this way, allosteric events alter protein conformational equilibria based on the amino acid networks that enable communication between distant sites of the protein.[424]

The remarkable sensitivity to single amino acid mutations within these allosteric networks, highlighted by single amino acid modifications of the ligand binding domain of the glucocorticoid receptor, emphasises the interdependence between disparate residues within the structure of proteins, and the large potential impact of changes to amino acid networks on catalysis.[425]

Thus, it stands to reason that small variations in protein structure must be assessed not only on their impact directly on the local chemical environment as has been traditionally performed, but also the impact on the flexibility and dynamic behaviour of the entire protein.

In particular, potential allosteric effects on the flexibility and conformational ensemble of the protein caused by disease related amino acid variations, such as the I148M variant of PNPLA3, cannot be overlooked.

Investigation of protein dynamics is extremely challenging to achieve experimentally. NMR spectroscopy is the most suitable technique, facilitating some investigation into protein dynamics through the estimation of conformer populations and the rates of kinetic processes.

While there are several studies using NMR relaxation and dispersion experiments investigating stable proteins such as RNase A,[426] these investigations are costly, time consuming and only amenable to certain proteins which behave well *in vitro* conditions, as discussed in the previous chapter.

The need for an approach to investigate the dynamics of other non-ideal proteins, has led to the development of the broad field of molecular dynamic simulations, which has become a routine approach used to investigate the dynamic movements of proteins on shorter timescales *in silico*.

5.2.1 Molecular dynamic simulations

Molecular dynamic simulations are computational simulations which explore the movement of atoms and molecules within a defined system. These can be used to explore a range of different system types ranging from solid state materials, to more complex biomechanical systems.

During molecular dynamic simulation, the classical Newtonian equations of motion are solved for each atom in the system with numerical methods over the picosecond and nanosecond timescale. By inputting known physical laws and simplifications of atomic properties, this results in a simulation of the motion of the physical system over time, and provides a way to access the previously elusive nature of protein structural fluctuations.[179]

5.2.2 Advantages of molecular dynamic simulations

The majority of known structures to date, have been solved by X-ray crystallography which has the highest resolution of structural data. However, this technique offers little insight into the dynamics of the protein system. Some information on the dynamics can be inferred by the B-factors and missing flexible loops, but this information is limited, and it can be difficult to differentiate between regions of high flexibility and lower quality data.

When investigating amino acid variations, these static protein structures give clues as to potential functional states and interactions that each variant may impact; however, the sheer complexity of these systems means that accurate interpretation based on visual inspection and simple calculations alone is not possible. While visualising the protein structure can be useful to an expert in the biochemistry of that protein, it may or may not be useful for hypothesizing the effects the SNP may have on the dynamic structure, and activity of the protein. This is confounded by the fact that allosteric interactions can be subtle and visually non-obvious.

Molecular dynamic simulations can make use of the already large database of known structures, to infer crucial dynamic information and provide the ultimate detail regarding each atom in the system. They can therefore be used to answer specific questions about the properties of a model system, through quantitative prediction. This has the advantage of relying less on the advanced knowledge of the researcher and is without the intrinsic bias which comes from qualitative interpretation.[427]

Furthermore, altering the simulation conditions can facilitate investigation into different properties of the protein's behaviour. By performing simulations with fixed protonation state and pH, the flexibility and stability of the protein under physiological conditions can be explored. This allows the prediction of subtle changes on stability and flexibility that may be caused by point mutations in the protein structure, which is of particular relevance when dealing with disease associated SNPs.[428,429]

By performing simulations at different temperatures, it is possible to reveal the thermo-stability of a protein and rare dynamic events, which is valuable to the investigation and design of enzymes for more efficient industrial processes.[430]

In the post genomic era, we have access to the sequence information of a huge number of proteins, and rare SNPs which may play important roles in disease. When dealing with such large volumes of SNPs, investigating them all experimentally is not feasible, because of the high cost both financially and temporally. Molecular dynamic simulations offers a solution for investigations into larger volumes of SNPs of interest, without the need for experimental optimisation.[431]

The above applications are based around the intrinsic value of investigating protein dynamics and are amenable to being used with both experimentally determined structures as well as structural models. However, dynamic simulations can also be used specifically with protein models, to both improve the quality of the model by allowing the relaxation of abnormal high energy bond confirmations, and additionally to observe the intrinsic quality of the model based

on stability under physiological conditions. For this reason, it has become more prevalent to use dynamic simulations, not only to investigate dynamic space of proteins but also as a final stage in model refinement even in complex proteins.[432]

5.2.3 Varieties of Molecular Dynamic Simulation

Not all molecular dynamic simulations are run with the same level of atomistic detail. In general, molecular dynamic simulations can be classified under three primary categories, molecular mechanics (MM) simulations, quantum mechanics (QM) simulations and quantum mechanics/ molecular mechanics hybrid (QM/MM) simulations.

5.2.3.1 Molecular mechanics simulations

Molecular mechanics simulations offer the lowest level of atomistic detail, which allows larger system sizes to be simulated for longer and biologically relevant timescales. In a molecular mechanics simulation, the so-called molecular mechanics forcefield is used to compute the potential energy of the system using simple chemical concepts.

The force on each atom within the system is computed using an energy function including terms representing bond lengths, bond angles, torsional angles, van der Waals interactions and electrostatic contributions. The energy contributions of each term are then defined within the molecular dynamic forcefield, based around the potential energy function.

The acceleration of each atom can be calculated for each atom via Newton's law of motion. This can be integrated to define motion from a given position at a given time over the simulation timestep.

Electrons are ignored in the molecular mechanics forcefield, being approximated within the simulation using polarisation of charges within the atom type descriptions. This means that processes involving electronic rearrangement such as chemical reactions cannot be observed using this method of simulation.[433]

5.2.3.2 Quantum mechanics simulations

Quantum mechanics simulations are performed using approximations of the Schrödinger equation to fully evaluate realistic and complex energy surfaces based on a first principle understanding of the target system.

This allows the simulation to achieve the highest level of atomistic accuracy, which can also model charge transfer interactions and environmental dependent polarisation of atoms.[434]

Although QM simulations are able to achieve remarkable accuracy, even with computational advances they remain extremely resource intensive, limiting simulations to low atom number and short time scales.[435]

5.2.3.3 Quantum mechanics / molecular mechanics hybrid simulations

More recently a combinatory approach has become computationally feasible which allows quantum mechanics/ molecular mechanics (QM/MM) hybrid simulations to be performed. These allow a smaller subset of atoms within the system to be treated at an appropriate level of quantum chemistry, while the remainder of the system is described by a molecular mechanics force field.

This has a distinct advantage in protein biochemistry, by allowing chemical reactivity to be investigated in specific regions of biological interest such as active site residues and polymorphisms, while reducing computational costs and still including important information from the larger substructure of the protein. While this has become much more popular, it remains highly computationally expensive which limits the widespread application of the technique.[436,437]

5.2.4 Limitations of molecular dynamic simulations

The biggest limitation of molecular dynamic simulations remains the high computational cost of simulating high atom number systems over significant timescales to observe large scale protein movements.

Hence, to allow simulations of biological systems to take place over a useful timeframe requires a number of simplifications to be applied to the system which is being simulated. Determining which simplifications and assumptions can be made while maintaining the biophysiological

nature of the protein during simulation is a unique challenge, which still has to be dealt with on a case by case basis for each and every system simulated.[438]

Additionally, the method of velocity calculation within the system will mathematically generate cumulative errors based on rounding errors during integration steps. While this is generally deemed to not cause significant errors over short time scale simulations can become problematic when working in the µs timescale and above.[439]

Finally, much like structural modelling of the protein, assessing the validity of the simulations is an area that remains challenging without experimental data. This means that quality assessment often relies on the intrinsic stability of the system as a precursor for biologically accurate, whereas this may not always be a valid assumption.[440]

5.2.5 Steps of molecular mechanics simulation

Progression from a static structure to production quality molecular dynamic simulations is achieved via several distinct steps, which ensure the system is adequately prepared and attempt to avoid the natural dynamics of the system from being perturbed.

A molecular mechanics simulation can be said to go through 5 key stages: system preparation, energy minimisation, heating, equilibration and final production (Figure 5.1).

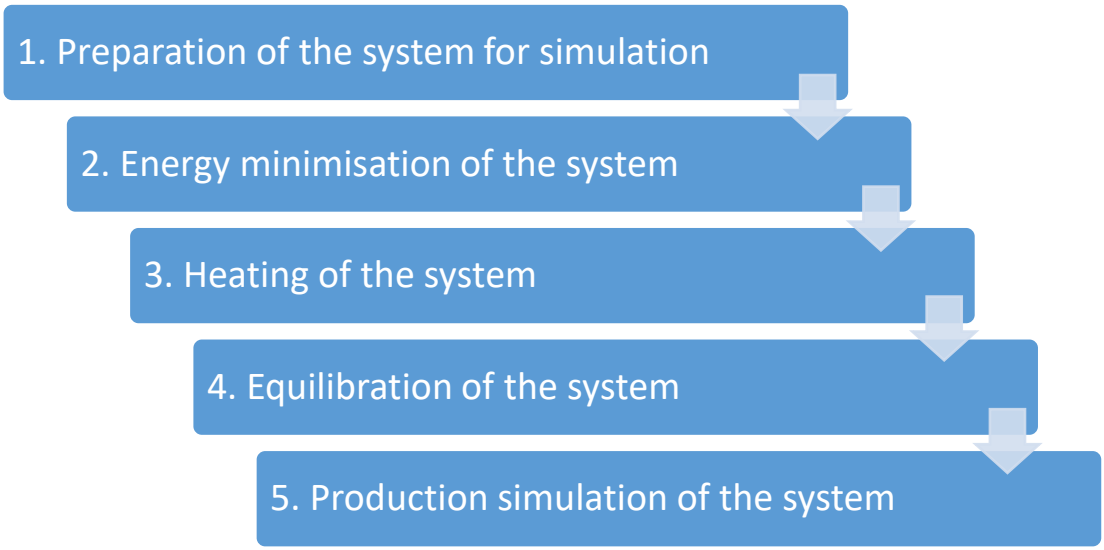


Figure 5.1 The stages of producing a molecular dynamic simulation

5.2.5.1 Preparation of the system for simulation

The initial preparation of the system involves ensuring that you have a full description of the system being investigated. The forcefield, which contains parameters to fully describe the forces within the system must be selected. It is recommended a well validated biomolecular forcefield is used to investigate proteins.[441]

You must ensure each atom is adequately described in your forcefield. Different atom types are assigned to not only different elements, but different hybridisation states and chemical environments. This is generally performed through the use of an atom typing software; however, most traditional biochemical protein atom types are included in the common forcefeilds.[442–445]

Other parameters for each atom type to fulfil the potential energy equations are then read from pre-existing tables, which have been developed for the specific forcefield.[441]

Finally, you will complete the set-up of your system to be simulated, by ensuring all other molecules are added. For example, many simulations are run within a periodic boundary water box with additional ions to replicate aqueous solution *in vitro*.

5.2.5.2 Energy minimisation of the system

Before beginning simulation, the system must be refined using an iterative minimisation algorithm. This relieves the local stresses introduced into the system by poorly described initial geometry, for example overlapping atoms, bond length and angle distortions and newly added solvent molecules.

A gradient descent algorithm is generally applied, as this allows a computationally efficient approach to minimising energy. Notably this means any minimised structure will sit in a local energy minima rather than a global minimum, and so will be heavily dependent on the starting structural confirmation.[438]

5.2.5.3 Heating and equilibration of the system

Velocities are added into the minimised system by gradually increasing the kinetic energy. Atoms are randomly assigned kinetic energy based on a Maxwellian distribution of energy at the given

temperature. A simulation is performed, and additional kinetic energy is added in a stepwise fashion, until the desired simulation temperature is reached.

At this stage, the system must undergo prolonged simulation steps to ensure that the kinetic energy has been evenly distributed throughout the system and an energy equilibrium is obtained, ensuring no skewing of the production data occurs.

5.2.5.4 Production simulation of the system

The final simulation is run as a continuation of the equilibration over a far longer timescale. Simulations will each investigate a different portion of phase space, meaning the data of interest is the average information obtained over the course of the simulation. This means the confidence is directly proportional to the timescale of the simulation. Computational advances have allowed simulations to take place over nanosecond timescales and can range from 10 ns to 20 μs depending on the size of the system under investigation and atomistic detail used.[446]

5.2.5.5 Analysis of trajectory

Analysis of the completed trajectory must be performed to collect biologically relevant information from the simulation. There are many questions that can be asked of this data by analysing the overall flexibility of the protein, specific key contacts throughout the trajectory and ligand interactions within the structure.

The simplest and most common analyses performed are based around root mean square deviations (RMSD) and root mean square fluctuations (RMSF) of the protein structure.

While these can be calculated based on different subsets of the data, in general the RMSD is be based on the root mean square deviation between normalised atomic positions within the protein structure. This can be used to track changes in the global topology over the time course of the simulation.

On the other hand, the RMSF, is generally based on the root mean square fluctuation of a given atom averaged over the time course of the simulation. This gives an overview of more local atomic shifts which have occurred in the simulation and the degree of flexibility along the length of the protein chain.

Further analysis will then be performed based on the investigation into unique portions of the simulation which are relevant to the system in question, for example observing the unique local area around a disease related SNP.[446,447]

5.2.5.6 Quality assessment

The gold standard for quality assessment of a molecular dynamic simulation is through experimental validation. This can take place either via direct study of the dynamic state of the system using NMR, or indirectly through predicted biological responses to mutagenesis of key residues.

In practice, experimental data cannot be obtained for many simulations, due to practical difficulties working with the target protein. To counteract this, intrinsic properties of a simulation can be used as a proxy for simulation quality.

The factors used to assess quality are based around evaluating the thermodynamic stability and convergence within the system, via the RMSD, RMSF and energy profiles. The principle challenge of this approach is determining to what extent instability is due to an issue with the simulation input, or because of native biological instability.[440]

5.2.6 Software variations

There is a range of software available that offer the ability to perform molecular dynamic simulations, for example AMBER (Assisted Model Building and Energy Refinement),[446] CHARMM (Chemistry at HARvrd Marcromolecular Mechanics),[448] NAMD (Nanoscale Molecular Dynamics)[449] and GROMACS (Groningen Machine for Chemical Simulations).[450]

Each software is designed for specific simulation related tasks and has a variety of different tools available as well as access to additional forcefields; however, the underlying principles behind the simulations remain the same. In fact, it is the forcefields and settings applied which will have the largest impact on the simulation rather than the software used to execute it.

For this project AMBER was chosen to run simulations, because of the versatility and full spectrum of tools available within AMBER and AMBER-TOOLS packages, as well as the advanced optimisation for graphics processing unit (GPU) implemented code to accelerate simulations.

5.2.7 Simulations of PNPLA3

Because of the difficulty in obtaining an experimentally derived structure, or high confidence homology models, only one investigation into the dynamic behaviour of PNPLA3 has been attempted by Xin *et al.*, 2013.[412]

In this investigation, four simulations were performed of PNPLA3 WT and I148M variants, both substrate free and bound with palmitic acid, over a short time scale of 10ns.

The group showed the binding energy between palmitic acid and PNPLA3 was reduced with the I148M variant. In addition, it was observed that the active site and substrate binding channel was reduced in size.

These factors were predicted to be the cause of reduced catalytic activity in this variant, supporting experimental evidence consistent with the I148M variant causing a loss of lipase activity; although they suggest residual activity may remain in variant protein.

While these simulations mostly showed adequate stability in the short term, the substrate bound simulation exhibits a 2A increase in RMSD over the 10ns simulation which does suggest the simulation has not yet converged and there is potential unfolding of the protein. It is not reported which atoms were included in this evaluation making assessment of quality of the simulations problematic.

Additionally, all simulations were performed using the truncated model of PNPLA3 based on Pat17, which has low levels of homology compared with novel structures in the PDB such as ExoU.

In conclusion, while this was provided novel insights into the behaviour of each PNPLA3 variant, limitations in both the starting structures and the length of simulation leaves unanswered questions as to whether this is a true representation of the biological system and mechanism of disease.

Today, a huge increase in computational power and the ability to utilise the power of GPU processing has opened up the ability to perform simulations with greater level of detail over longer timescales. This once again makes PNPLA3 an excellent target to explore further with molecular dynamic simulations.[412]

5.3 Aims

The aim of this chapter is to further investigate PNPLA3 using a set of molecular dynamic simulations to 1) refine the previously developed structural models; and 2) investigate the dynamic profile of the protein.

It is hoped that by further exploring the dynamic behaviour of both the wild type and I148M variant of PNPLA3 potential implications of the I148M variant on the dynamic equilibria of the protein may be revealed.

5.4 Methods

5.4.1 Investigatory 20ns simulations

The correct protonation state at pH7.5 was calculated for each residue using the H++ server (http://biophysics.cs.vt.edu/H++).[451–453] The extracted files were then prepared for simulation with the tLeap module of AMBER 16 and AMBER Tools 17.[446]

The standard residues in the enzyme complex were simulated using the ff14SB AMBER forcefield.[454] The enzyme complex was solvated in a periodic TIP3P water box with a 12 Å boundary. The net charge was neutralised with appropriate chloride or sodium ions and topology (prmtop) and coordinate (inpcrd) files generated for simulation.

Dynamic simulations were run using pmemd.cuda a GPU accelerated implementation of the pmemd amber simulation software. Runs were submitted via BASH script (Appendix III). [455,456]

All simulations were performed using particle Mesh Ewald method to characterise long range electrostatic interactions with a cut-off distance of 8 Å, and pairwise summation involved in calculating the effective Born radii also limited at 8 Å.

5.4.1.1 Minimisation

Minimisation steps were performed with initial restraints on the system which were periodically decreased over time to inhibit gross movements caused by poor contacts in the initial structural model and random solvent positioning.

step 1 - 50,000 cycles of steepest gradient descent with a restraint of 1000 kcal/mol/A^2 over the entire protein backbone.

step 2 - 50,000 cycles of steepest gradient descent with a restraint of 500 kcal/mol/A^2 over the entire protein backbone.

step 3 - 50,000 cycles of steepest gradient descent with a restraint of 200 kcal/mol/A^2 over the entire protein backbone.

step 4 - 50,000 steps of steepest gradient decent, followed by 25,000 cycles of conjugate gradient decent without restraint.

5.4.1.2 Heating

The system was heated in several stages, to allow for recalculation of periodic box boundary, during volume expansion. The temperature was controlled using the Langevin thermostat with a collision frequency of 1.0 ps-1. At each stage,

Step 1 - The temperature was raised from 0K to 100K, over 40,000 steps at constant volume, NMR restraints were applied to allow heating gradually over 20 ps with a 0.5 fs timestep. Shake remained off, so all bonds were unrestrained.

Step 2 - The temperature was raised from 100K to 250K at constant volume, NMR restraints were applied to allow heating gradually over 20 ps with a 0.5 fs timestep.

Step 3 - The temperature was raised from 250K to 280K at constant pressure, NMR restraints were applied to allow heating gradually over 15 ps with a 0.5 fs timestep. In this step pressure scaling was introduced to correct for density changes in the system on heating.

Step 4 - The temperature was raised from 280K to 310K at constant pressure, NMR restraints were applied to allow heating gradually over 15 ps with a 0.5 fs timestep.

5.4.1.3 Equilibration

Equilibrations were performed with constant pressure at 310K. Simulations ran for 1.5 ns with a 2 fs timestep. The shake function was enabled to constrain bonds involving hydrogen atoms. Temperature was controlled via the Langevin thermostat with a collision frequency of 1.0 ps^{-1}.

5.4.1.4 Production

Production simulations were performed with constant volume at 310 K. Simulations ran for 20 ns with a 2 fs timestep. The shake function was enabled to constrain bonds involving hydrogen atoms. Temperature was controlled with the Berendsen thermostat which applies the weak coupling algorithm.

For each initial simulation, the starting model number corresponds to the simulation number. (simulation 1 – model 1; simulation 2 – model 2; simulation 3 – model 3; simulation 4 – model 4; simulation 5 – model 5; simulation 6 – model 6; simulation 7 – model 7; simulation 8 – model 8; simulation 9 – model 9).

Two additional runs were tested to check for the influence of simplified conditions. One run was run with shake turned off throughout the simulation, and one run with NPT maintained throughout.

5.4.2 Final 100ns production simulations

Model 5 was chosen for long time scale molecular dynamics to investigate the I148M variant, based on the simulation stability and integrity of active site in the initial runs.

Isoleucine was replaced in the initial model with methionine using PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC. The model was prepared for simulation as described previously, and a short 20ns simulation run to assess stability.

The final frames of each simulation were extracted and again minimised and prepared for simulation using the above protocol as a heating and equilibration step. This provided two equally stable isoforms for full length simulation (Figure 5.2).

The simulations were run under the same conditions as the production standards used in the 20ns simulations for a 100ns period.

5.4.3 Molecular dynamic simulation processing and analysis

Perl scripts included in AmberTools 17, were modified for separate extraction of the energetic data from the minimisation, equilibration and production runs (Appendix III). Each dynamic simulation was analysed and processed using CPPTRAJ.[457]

For each model simulated, the RMSD and RMSF of the protein were calculated over the course of the full production simulation using the first frame as a reference. Heating and equilibration steps were also investigated for quality control (results not shown). Additional information was extracted using CPPTRAJ and included the distance between the key catalytic residues and variant (S47, D166, I/M148), the RMSD of the flexible loop region containing residues 166 and 148, the RMSD of the patatin domain (residues 5-179) and where applicable an extended patatin domain (residues 5-239).

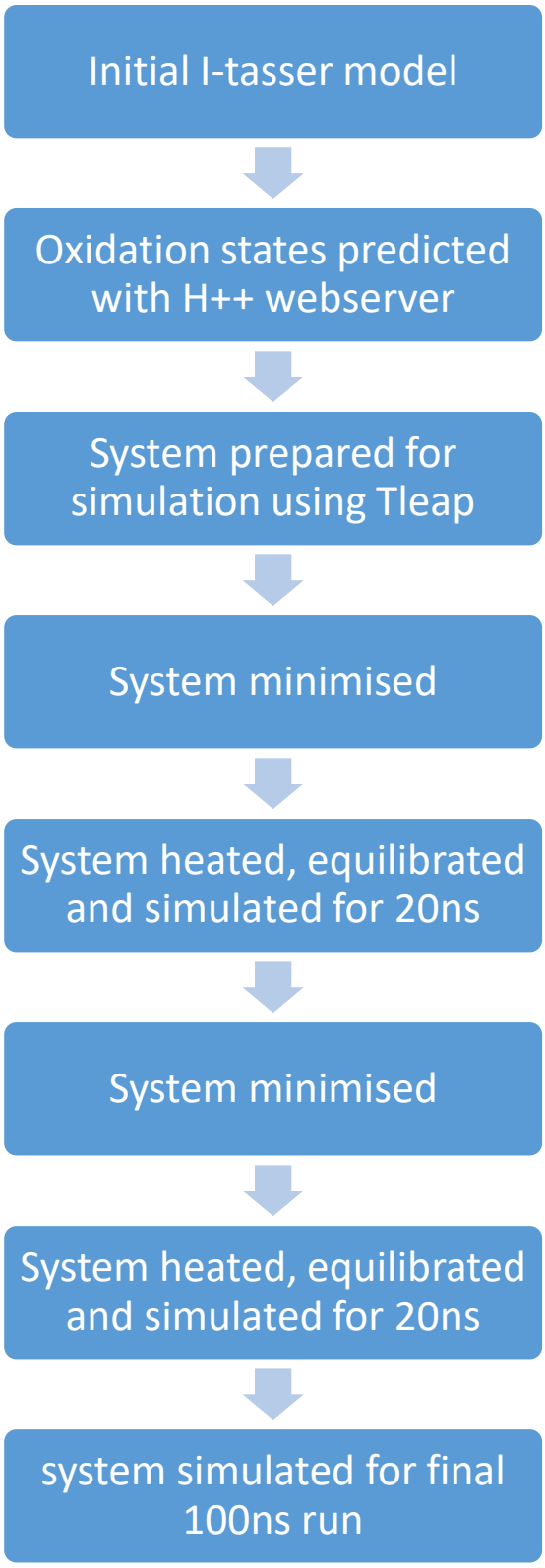


Figure 5.2 Full process diagram for final full length PNPLA3 molecular dynamic simulations

The results were visualised using a custom script in R studio, using base R and the GGPLOT2 and KNITR packages.[458–461]

Two-dimensional Richardson diagrams were generated using Pro-origami.[410] Secondary structure selected using STRIDE, distance matrix used rather than heuristics for placing helices, helices prevented from being placed between neighbouring sheets and separate numbering for helices and strands.[410]

5.4.4 Detection of tunnels to active site

Protein tunnels providing access to the catalytic residues were predicted using MOLE 2.5, to predict and characterise substrate access to the binding pocket in each variant.[462]

The search was initiated from cavities around the selection of SER 47. With a probe radius of 3.00, interior threshold 1.25, minimum depth of 4.07, bottleneck radius 1.25, origin radius 10.0, surface cover radius 10.14, bottleneck length 0.00, cut-off ratio 0.90, and no minimum tunnel or pore lengths. The tunnel detection was based on a Voronoi weight function, used to detect channels based on shortest pathways to the protein surface.

Detected pore data was exported and visualised using a custom R Shiny application, using base R and the GGPLOT2 packages.[458–461]

5.4.5 Ligand docking

Flexible docking, of fourteen PNPLA3 ligands (Table 5.1) reported in the literature, was performed using AutoDock Vina against both wild type and variant protein, based on the final frame in each 100ns simulation.[463]

All ligands were docked into 40A cubic box, generously surrounding the active site and nearest surfaces, which was positioned manually. 20 modes were recorded for each ligand, and the search exhaustiveness set to 40 to prioritise accuracy. The best mode was for each ligand was used for further analysis and imaging.

Three-dimensional images of the protein structures and docked ligands were generated using PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.

Table 5.1 Ligands which were docked to PNPLA3

Ligand	Fatty acid type	Fatty acid chain length
1,2-diolein	DAG	18 (1 cis double bond)
1,2-dipalmitin	DAG	16
1,3-dilinolein	DAG	18 (2 cis double bonds)
1,3-diolein	DAG	18 (1 cis double bond)
linoleic_acid	FA	18 (2 cis double bonds)
oleic_acid	FA	18 (1 cis double bond)
palmitic_acid	FA	16
retinoic_acid	FA	10 (6 carbon ring, 5 double bonds)
retinol	Alcohol	10 (6 carbon ring, 5 double bonds)
triarachidonin	TAG	20 (4 cis double bonds)
trilinolein	TAG	18 (2 cis double bonds)
trimyristin	TAG	14
triolein	TAG	18 (1 cis double bond)
tripalmitin	TAG	16

5.4.6 Multiple one nanosecond repeats

Isoleucine was replaced in the initial model with Alanine using PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC. All three models including each variant (Isoleucine, methionine and alanine) were simulated for 10 individual 1ns simulations using the same method as above, to explore the reliability of the simulations.

The collected data were compared using pairwise Wilcoxon rank sum tests performed using R version 3.3.2 and visualised using Microsoft Excel.

5.5 Results

5.5.1 Initial simulation results

All simulations were deemed to be adequately minimised and an energy minima successfully reached. The heating of each system showed no abnormal system fluctuations, with energy and temperature increasing at a smooth rate as determined by the input files. The temperature was maintained within 0.5^{o}C throughout the simulations, and only a slight decrease in system energy was observed (results not shown).

5.5.1.1 Model 1:

Simulation of model 1 produced a partially stable protein structure as demonstrated by small RMSD fluctuations. There was a gradual increase of the RMSD from 1.5 to 3A throughout the simulation, which suggests the stability is lower than would be desired from a high confidence simulation and the protein may be about to unfold or undergo a conformational shift. The RMSD fluctuations are reduced slightly, but not significantly when reducing the calculation to only the backbone atoms or α carbons, showing the instability is not due to side chain movements alone (Figure 5.3).

The RMSD of the flexible loop (residues 147 to 151) has an initial conformational fluctuation of roughly 1A over the first 2,500 frames of the simulation, before stabilising with RMSD fluctuations of only 0.2A.

The model has a relatively flat RMSF profile along the protein chain. No regions of the domain were fluctuating abnormally, suggesting a consistent level of stability across the structure.

The distance between the catalytic residues (S47 and D166) show 2A fluctuation in the first half of the simulation but maintain an average distance based on their centre of mass of around 6A. The distances between each catalytic residue and residue 148 also remain consistent throughout the simulation, and places them within close proximity of one another, with centre of masses under 10A apart; showing good conservation of the active site throughout simulation.

5.5.1.2 Model 2:

Model 2 underwent a shift in conformation between frames 1,000 and 2,000 represented by an increase in RMSD to 4A. After this shift, the RMSD continued with fluctuations of only 0.5A, showing a stable conformation has been reached (Figure 5.4).

The RMSD of the loop fluctuated between 0.5 and 1.5A throughout the simulation and was less stable than most of the alternate model simulations.

Model 2 also has a relatively flat RMSF profile along the protein chain, with all regions showing RMSF below 5A. No regions of the domain were fluctuating abnormally, suggesting a consistent level of stability across the structure.

The distances between the residues of interest all remained low throughout simulation and decreased slightly after frame 4,000. The average distance between the centre of mass of the catalytic residues was 5.5A. The distance between I148 and S46 were approximately 10A with roughly 2A fluctuations. The distance between I148 and D166 fluctuated between 14A and 22A showing very high levels of deviation from the standard.

5.5.1.3 Model 3:

Simulation of model 3 produced an instable protein structure. There was an overall increase of RMSD to 5A, which while slightly faster early in the simulation, generally occurred with a consistent upward trend throughout the whole simulation. In addition, the RMSD fluctuations were around 1.5A throughout the simulation (Figure 5.5).

The RMSD decreased to fluctuations between 2 and 4A when reducing the calculation to only the α carbons, showing the fluctuation is in part due to the movement of side chain atoms.

The RMSD of the flexible loop is initially stable, but exhibits large fluctuations for a short chain of between 0.5 and 1A. The RMSD increases steadily to 1.5A over the final 2,500 frames of the simulation, mirroring the upward trend of the overall protein RMSD.

The model has a large RMSF profile, which is far greater than the previous models. Along the protein chain multiple regions fluctuating over 20A. This high level of fluctuation appears abnormal, suggesting poor stability across the structure.

The distances between the catalytic residues vary wildly between 10 and 18A throughout the simulation, with no discernible pattern to the movements. The distances between catalytic

residues and I148 only fluctuate by around 4A; however, the total distance remains over 8A throughout simulation, suggesting little direct interaction between them.

5.5.1.4 Model 4:

Simulation of model 4, showed an initial instable structure, which lead to a conformational adjustment with RMSD increasing to 6A over the first 3,000 frames of the simulation. At this point there appears to be a stable conformation reached within the protein structure, and fluctuations are observed to be below 0.6A. The RMSD only decreased by roughly 1A when reducing the calculation to only the backbone atoms or α carbons, showing the instability is not due to side chain movements alone (Figure 5.6).

The RMSD of the flexible loop is stable throughout, with RMSD fluctuations of only 0.5A for the majority of the simulation.

The model has a large RMSF profile along the protein chain, with certain regions fluctuating over 15A, although it is lower than model 3. This high level of fluctuation appears abnormal, suggesting poor stability across the structure.

The distances between the catalytic residues remained at approximately 11A apart; however, there were large fluctuations in distance of up to 4A. Throughout the simulation, the distance between reside I148 and S47 decreased and stabilised from 13.5A, to 12.5A. The distance between I148 and D166 increased constantly throughout the simulation from 7A to 10A. This gradual increase in distance throughout the simulation, suggesting there may be an unfolding of the loop region.

5.5.1.5 Model 5:

Simulation of model 5, showed a generally instable structure with a large RMSD increase between 2 and 6A. The RMSD fluctuation decreases by approximately 2A when only considering the α carbons, showing that while they make a large contribution, the instability is not only due to side chain movements (Figure 5.17).

When considering the patatin domain alone (residues 5 - 179), or the extended patatin domain, the RMSD quickly reached a stable level with fluctuations of only 0.5A, between 1.5A and 2A across the initial 8,500 frames of the simulation. In the last 1,500 frames, the RMSD shows signs of increasing rapidly from 2A to 2.5A RMSD, suggesting a second structural shift.

The loop has a stable RMSD throughout the simulation, with fluctuations of less than 0.7A, with a very slight upward trend.

The model has a relatively flat RMSF profile along the entire protein chain. Notably the patatin domain of the protein was clearly lower than the C-terminal half of the protein. No regions of the domain were fluctuating abnormally, suggesting a consistent level of stability across the structure. There was a peak of flexibility in the region of residues 250 to 380, which had an RMSF up to 12A. All the residues of interest were located on regions of median flexibility with an RMSF of around 3A.

The distances between the catalytic residues were very stable fluctuating between 5 and 5.5A. The distance between I148 and S47 increased across the simulation from 7A to 8.5A, which appears to be a trend which could continue. The distance between I148 and D166 also remains constant with an approximate distance of 9.8A between the centre of masses.

5.5.1.6 Model 6:

Simulation of model 6 produced a partially stable protein structure, with RMSD increase from to 4A throughout the simulation. After the initial 1,500 frames the RMSD does appear to stabilise, but an overall upward trend remains. The RMSD does decrease slightly when reducing the calculation to only the backbone atoms or α carbons; however, again the RMSD increase remains (Figure 5.8).

When considering the patatin domain alone, the RMSD increase doesn't change in intensity, but rather appears more gradual throughout the simulation. This further implies the patatin domain may not be stable in this simulation or may be about to undergo a conformational change.

The RMSD of the flexible loop is very stable at 0.7A throughout the simulation.

The model has a relatively flat RMSF profile along the protein chain, all below 5A. No regions of the domain were fluctuating abnormally, suggesting a consistent level of stability across the structure.

The distance between the catalytic residues show 1.5A decrease from 7.0A apart to only 5.5A within the first 2,000 frames of the simulation. The distance between I148 and S47 decreased across the simulation from 10A to 8.5A, which appears to be a trend which could continue. The distance between I148 and D166 also remains constant with an approximate distance of 7A

between the centre of masses. This shows good conservation of the active site throughout simulation.

5.5.1.7 Model 7:

Simulation of model 7 produced unstable results with RMSD increasing from 1 to 6A throughout the simulation. A partially stable confirmation was achieved between frames 2,000 and 8,000, but this conformation was subsequently lost. The RMSD decreased by only 1A when reducing the calculation to only the backbone atoms or α carbons, showing the instability is not due to side chain movements alone (Figure 5.9).

When considering the patatin domain alone, the RMSD profile remained unchanged, showing the entire protein is equally contributing to the deviations.

The RMSD of the flexible loop had increasing fluctuations throughout, gradually increasing to 0.8A.

The model has a large RMSF profile at the beginning of the protein chain within the core patatin domain, at 15A. This could imply poor stability within this domain.

The distance between the catalytic residues gradually increased from 12 to 20A throughout, with a stable distance of 16A mirroring the overall RMSD. The distance between S47 and D166 remained roughly constant at 10A after the initial 2,000 frames. The distance between I148 and D166 also increased throughout simulation from 6 to 14A. The increasing distances were greater toward the end of the simulation, suggesting a continued unfolding of the domain.

5.5.1.8 Model 8:

Simulation of model 8 showed an unstable structure with RMSD climbing 12A from the starting structure over the first 3,000 frames. The remainder of the simulation is still unstable, with RMSD fluctuations between 8 and 12A, but does not appear to be degenerating further (Figure 5.10).

This did not decrease significantly when reducing the calculation to only backbone atoms and α carbons.

The RMSF along the protein chain are not abnormally pronounced. In particular, the first hundred residues seem to have greater fluctuation of over 14A, than the rest of the protein.

5.5.1.9 Model 9:

Simulation of model 9, showed a generally instable structure with a gradual increase in RMSD from 2 to 7A. The RMSD decreases by approximately 1A when only considering the α carbons, showing that while they make a contribution, the instability is not only due to side chain movements (Figure 5.11).

When considering the patatin domain alone (residues 5 - 179), or the extended patatin domain, the RMSD reached a stable level after the initial 4,000 frames, with fluctuations of only 1A, between 2.5A and 3.5A.

The flexible loop has a stable RMSD throughout the simulation, with fluctuations of less than 0.5A.

The model has a relatively flat RMSF profile along the entire protein chain. Notably the patatin domain of the protein was clearly lower than the C-terminal half of the protein. No regions of the domain were fluctuating abnormally, suggesting a consistent level of stability across the structure. There was a peak of flexibility in the region of residues 250 to 380, which had an RMSF up to 12A. All the residues of interest were located on regions of median flexibility with an RMSF of around 3A.

The distances between the catalytic residues varied between 9 and 14A throughout the simulation. This occurred gradually, suggesting movement in this domain as opposed to a specific conformational shift. The distance between I148 and S47 remained stable at approximately 9.5A. The distance between I148 and D166 also remains constant with a large distance of approximately 18A.

5.5.1.10 Additional quality control simulations

To confirm no important performance loss was occurring by constraining hydrogen bonds and not directly controlling for pressure, additional runs of model 1 were performed under each condition. In each instance the system was either similarly or less stable than the primary runs (Appendix III).

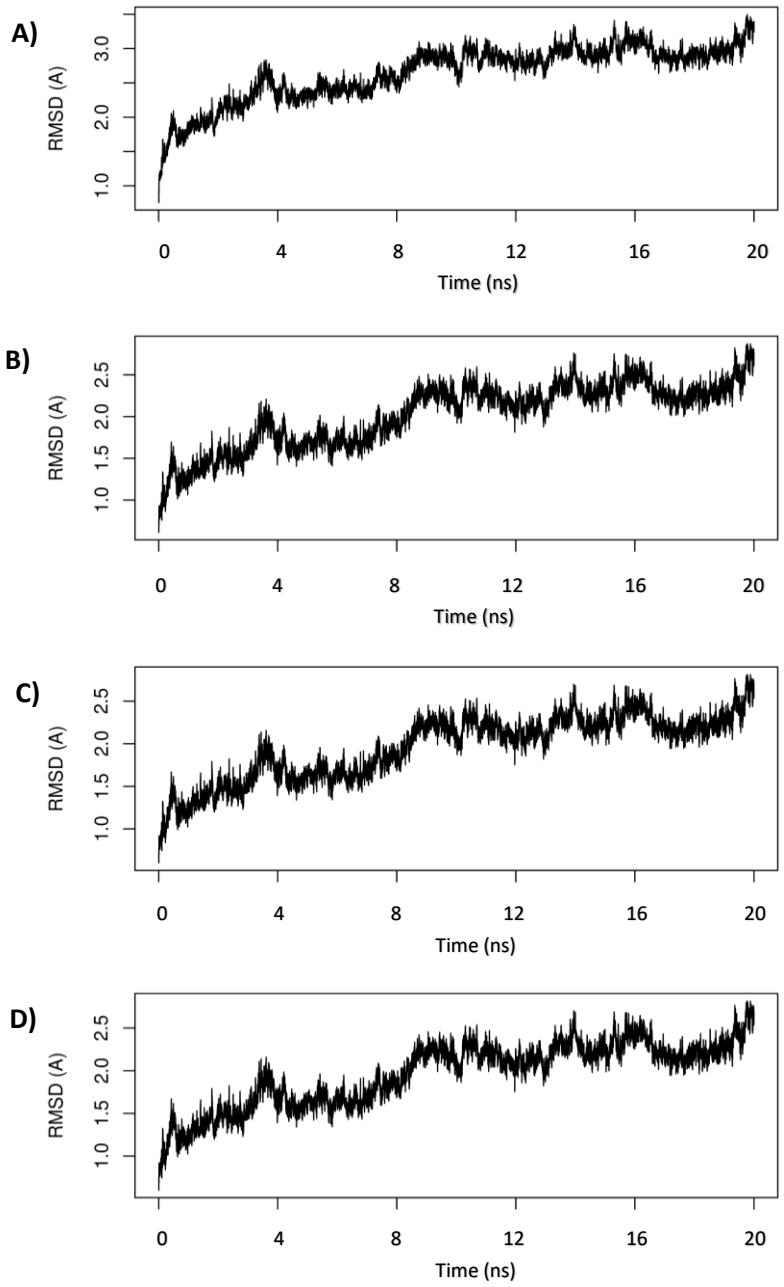


Figure 5.3 Model 1 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.
A) RMSD of entire protein chain
B) RMSD based on backbone chain atoms C, CA and N
C) RMSD based upon α carbons
D) RMSD including only the patatin domain (residues 5-179)

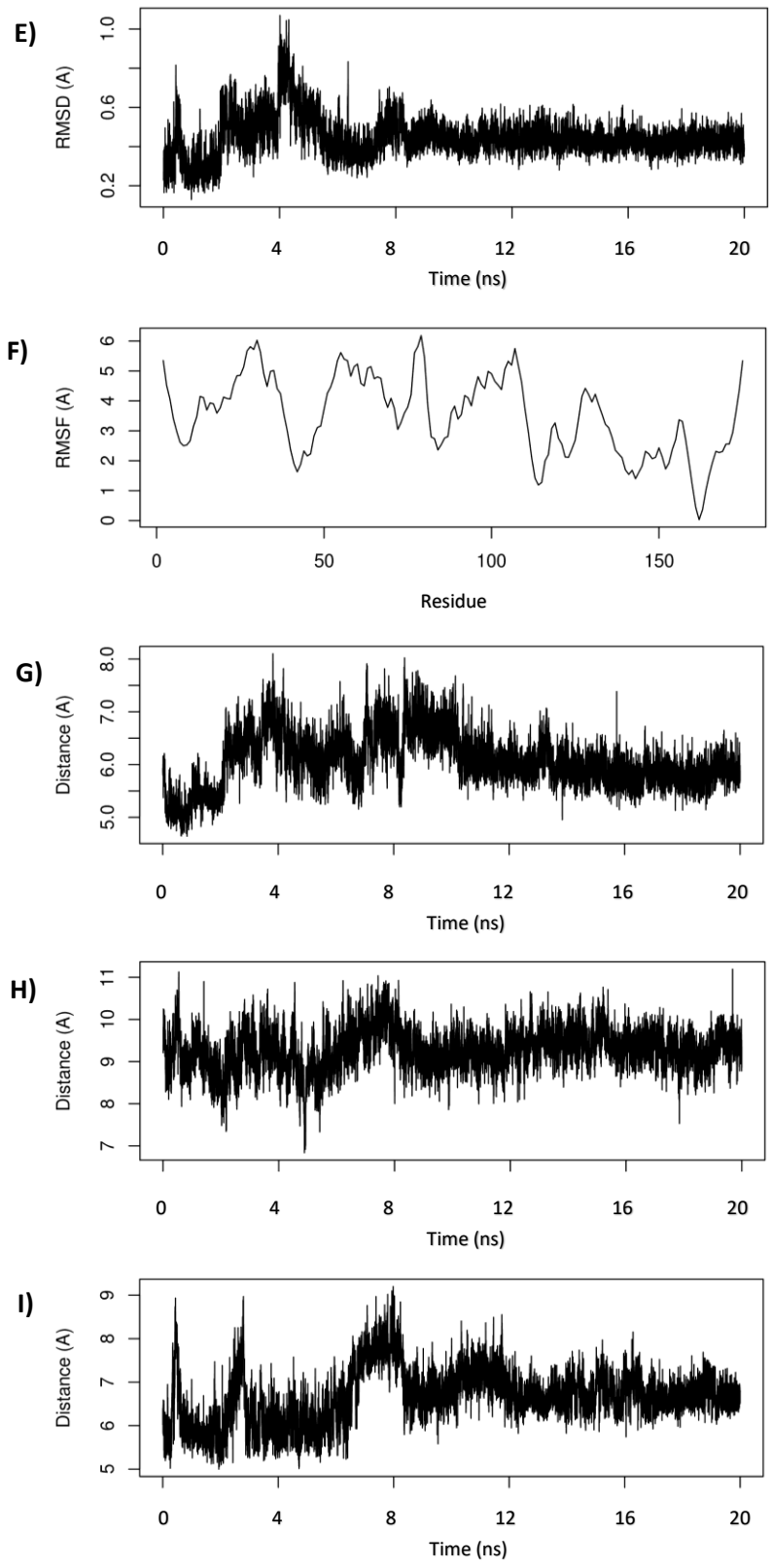


Figure 5.3 (Continued)

E) RMSD of I148M containing loop (residues 147-151)

F) Root mean square fluctuations along the protein chain averaged for full simulation

G) Distance between S46 and D166

H) Distance between S46 and I148

I) Distance between D166 and I148.

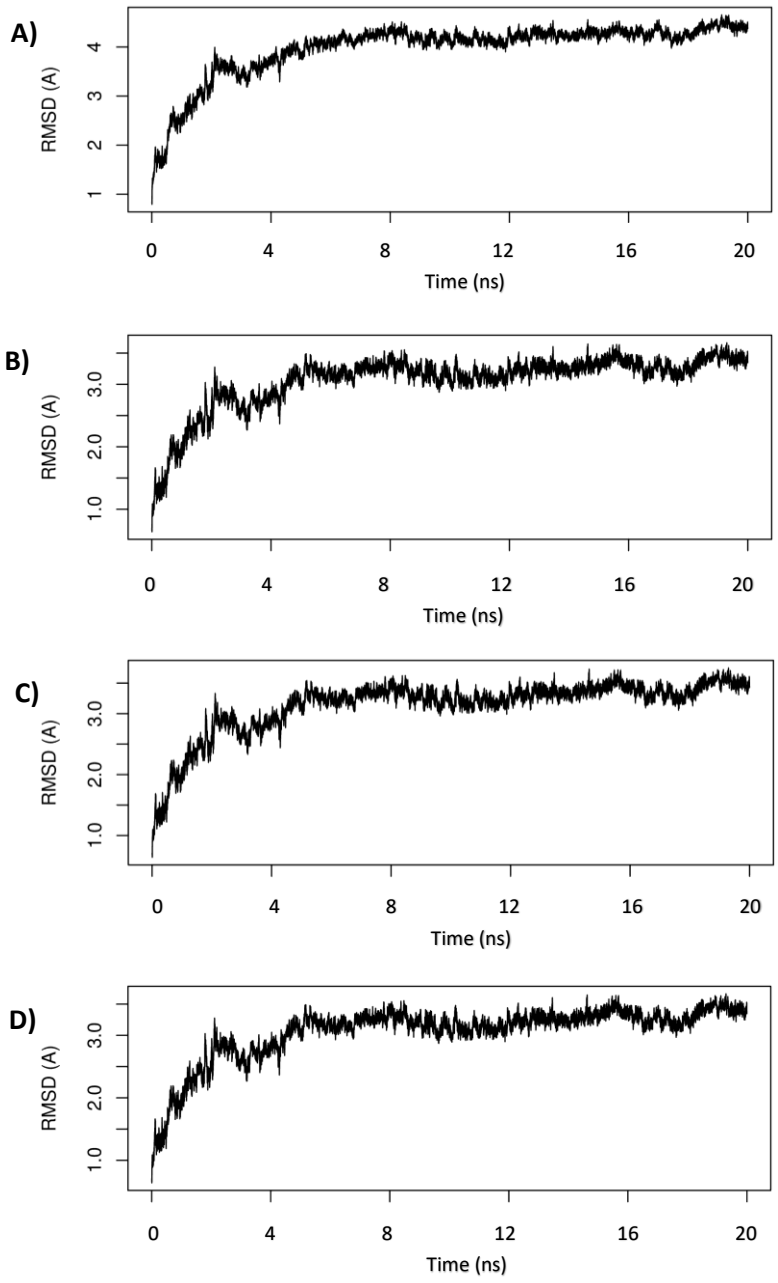


Figure 5.4 Model 2 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.
A) RMSD of entire protein chain
B) RMSD based on backbone chain atoms C, CA and N
C) RMSD based upon α carbons
D) RMSD including only the patatin domain (residues 5-179)

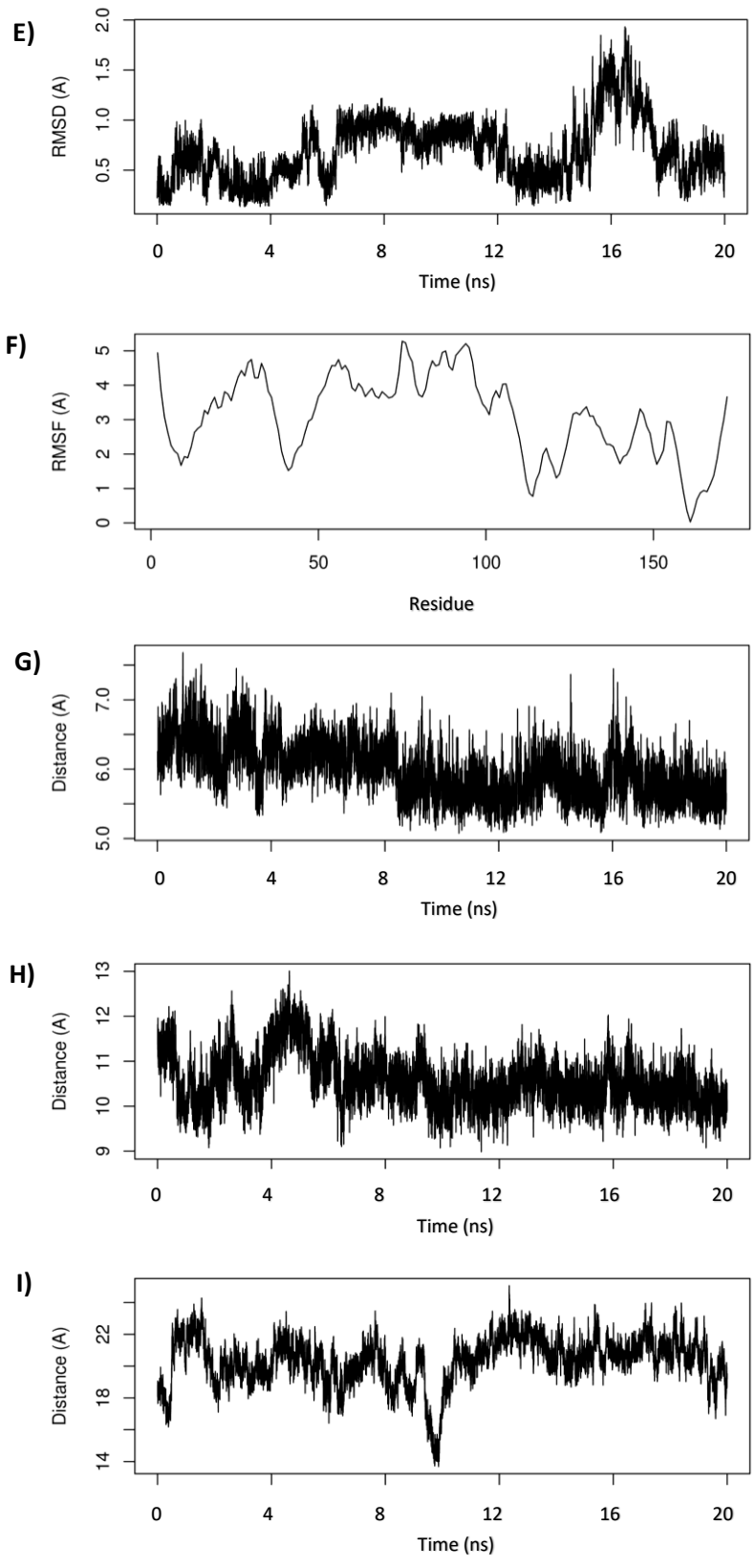


Figure 5.4 (Continued)

E) RMSD of I148M containing loop (residues 147-151)
F) Root mean square fluctuations along the protein chain averaged for full simulation
G) Distance between S46 and D166
H) Distance between S46 and I148
I) Distance between D166 and I148.

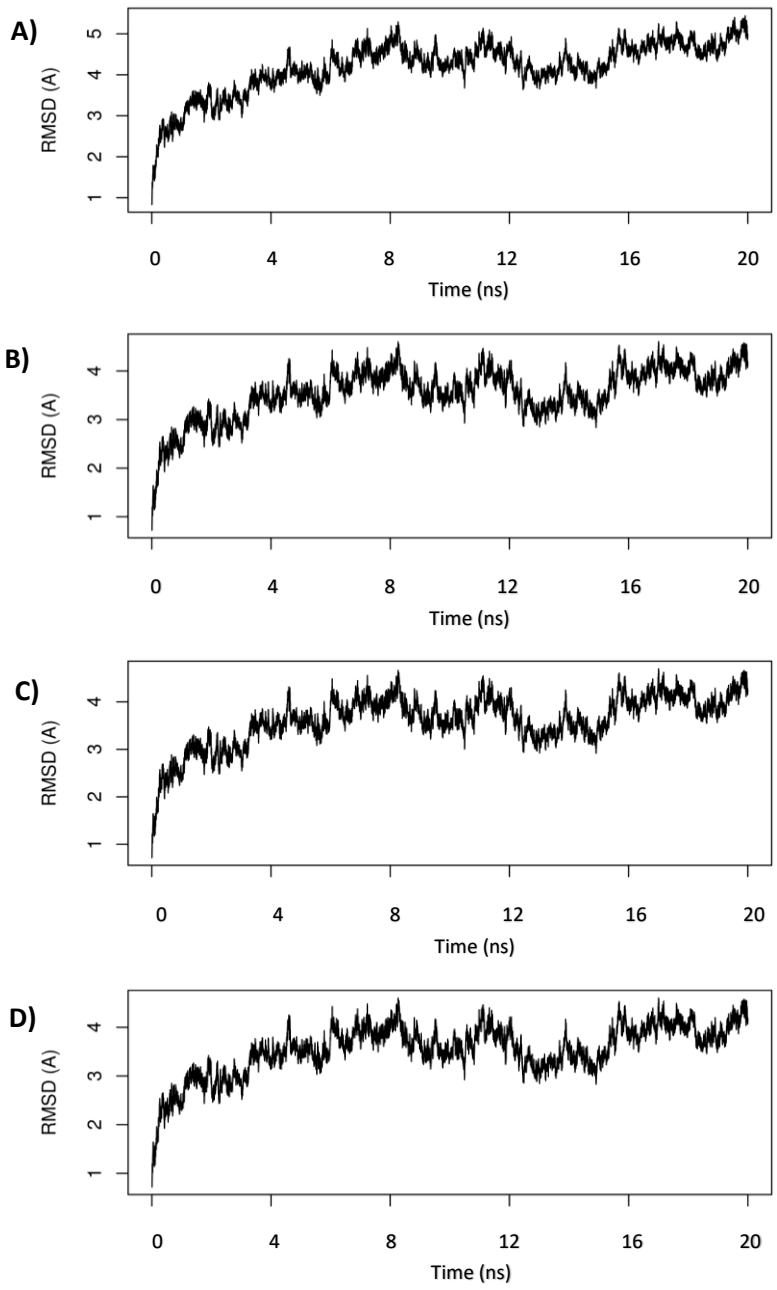


Figure 5.5 Model 3 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.

A) RMSD of entire protein chain

B) RMSD based on backbone chain atoms C, CA and N

C) RMSD based upon α carbons

D) RMSD including only the patatin domain (residues 5-179)

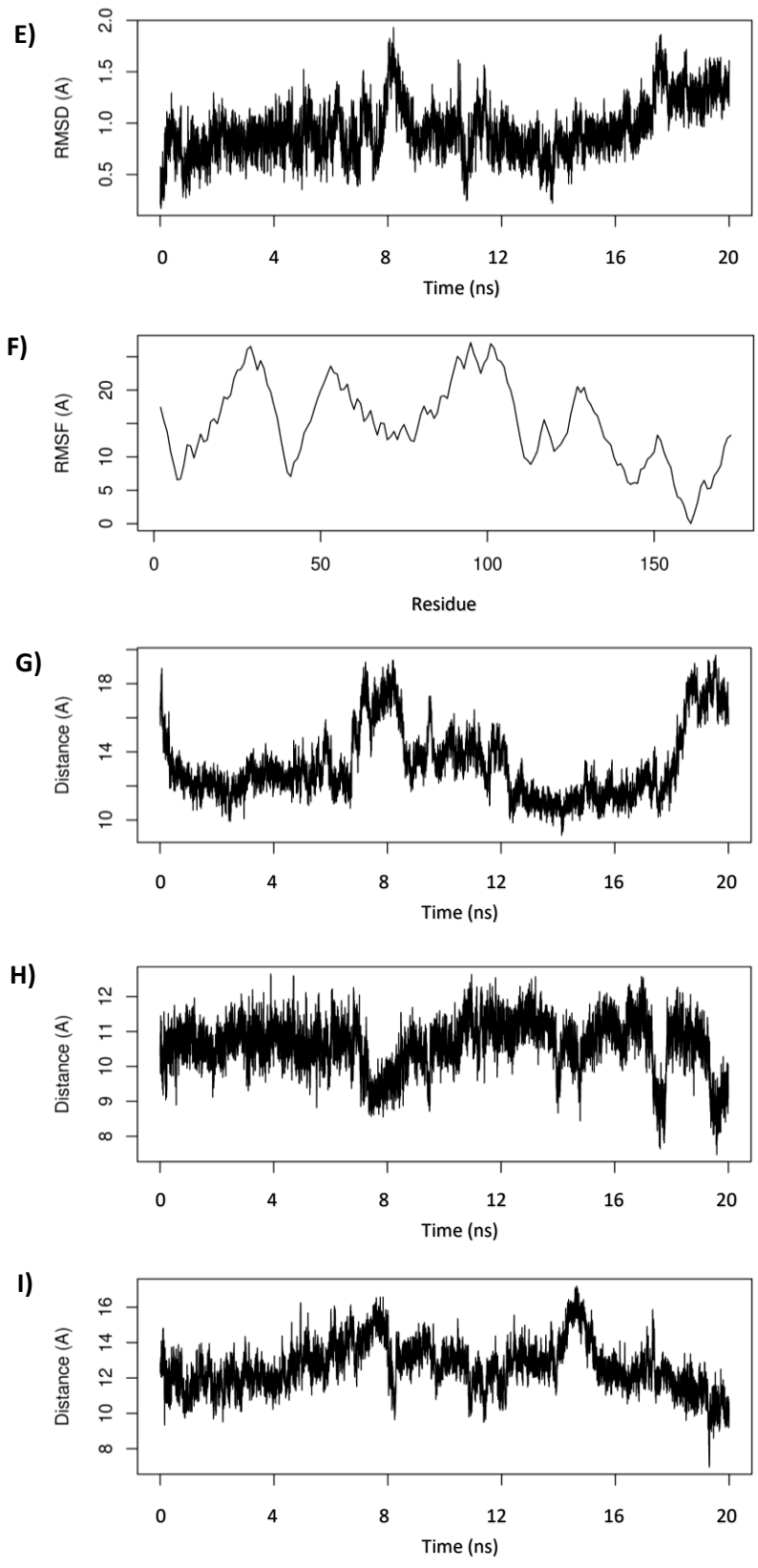


Figure 5.5 (Continued)

E) RMSD of I148M containing loop (residues 147-151)
F) Root mean square fluctuations along the protein chain averaged for full simulation
G) Distance between S46 and D166
H) Distance between S46 and I148
I) Distance between D166 and I148.

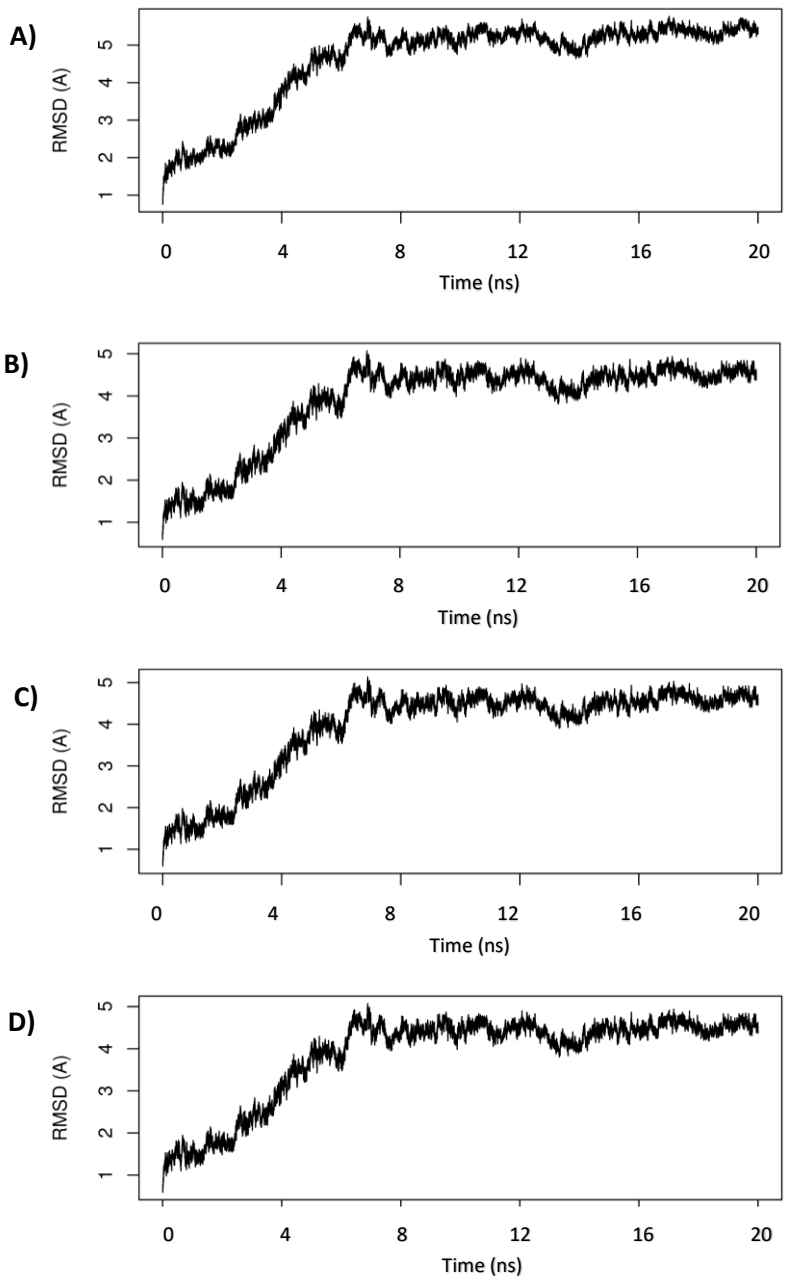


Figure 5.6 Model 4 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.
A) RMSD of entire protein chain
B) RMSD based on backbone chain atoms C, CA and N
C) RMSD based upon α carbons
D) RMSD including only the patatin domain (residues 5-179)

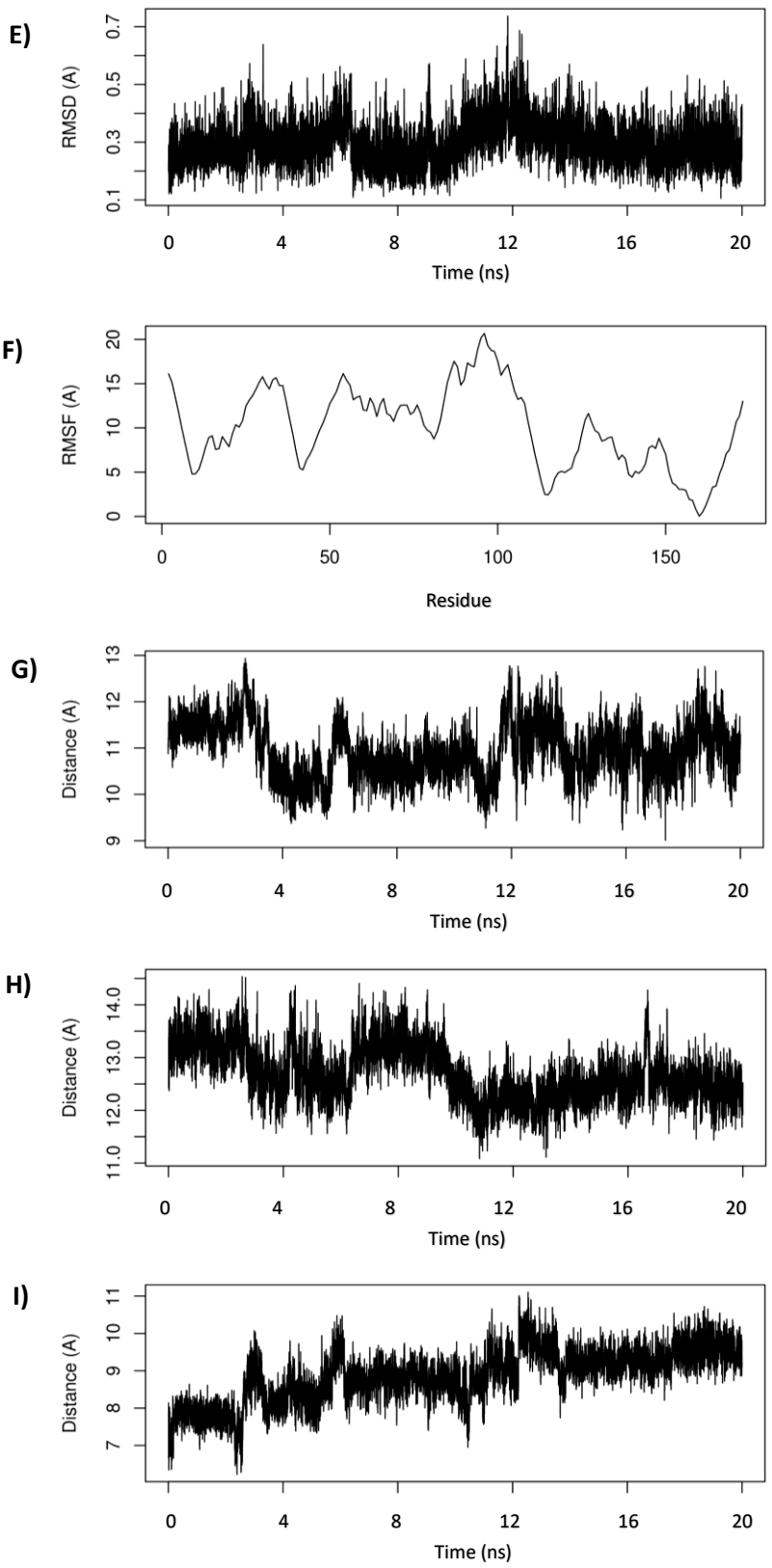


Figure 5.6 (Continued)

E) RMSD of I148M containing loop (residues 147-151)
F) Root mean square fluctuations along the protein chain averaged for full simulation
G) Distance between S46 and D166
H) Distance between S46 and I148
I) Distance between D166 and I148.

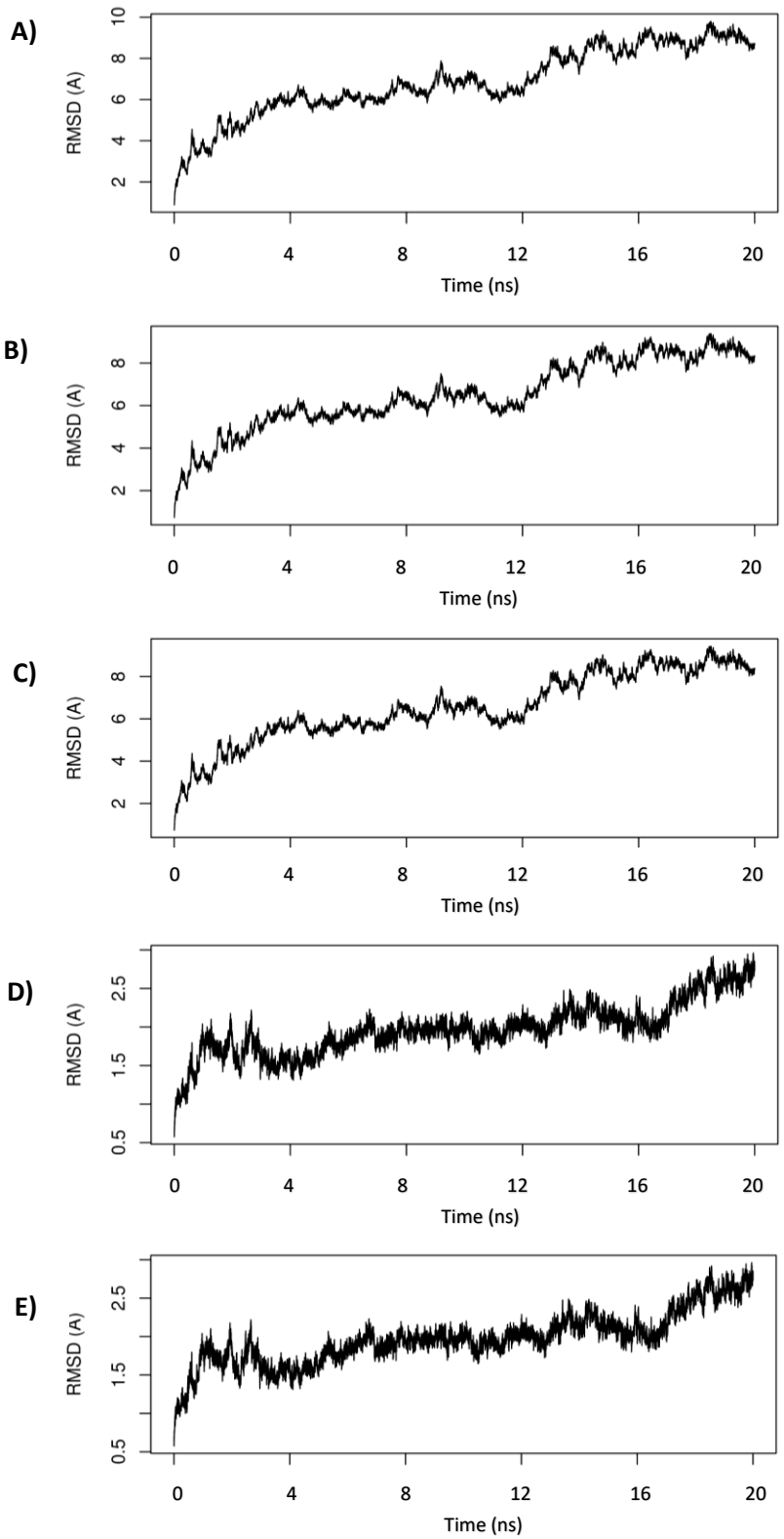


Figure 5.7 Model 5 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.

A) RMSD of entire protein chain

B) RMSD based on backbone chain atoms C, CA and N

C) RMSD based upon α carbons

D) RMSD including only the patatin domain (residues 5-179)

E) RMSD including extended patatin domain (residues 5-239)

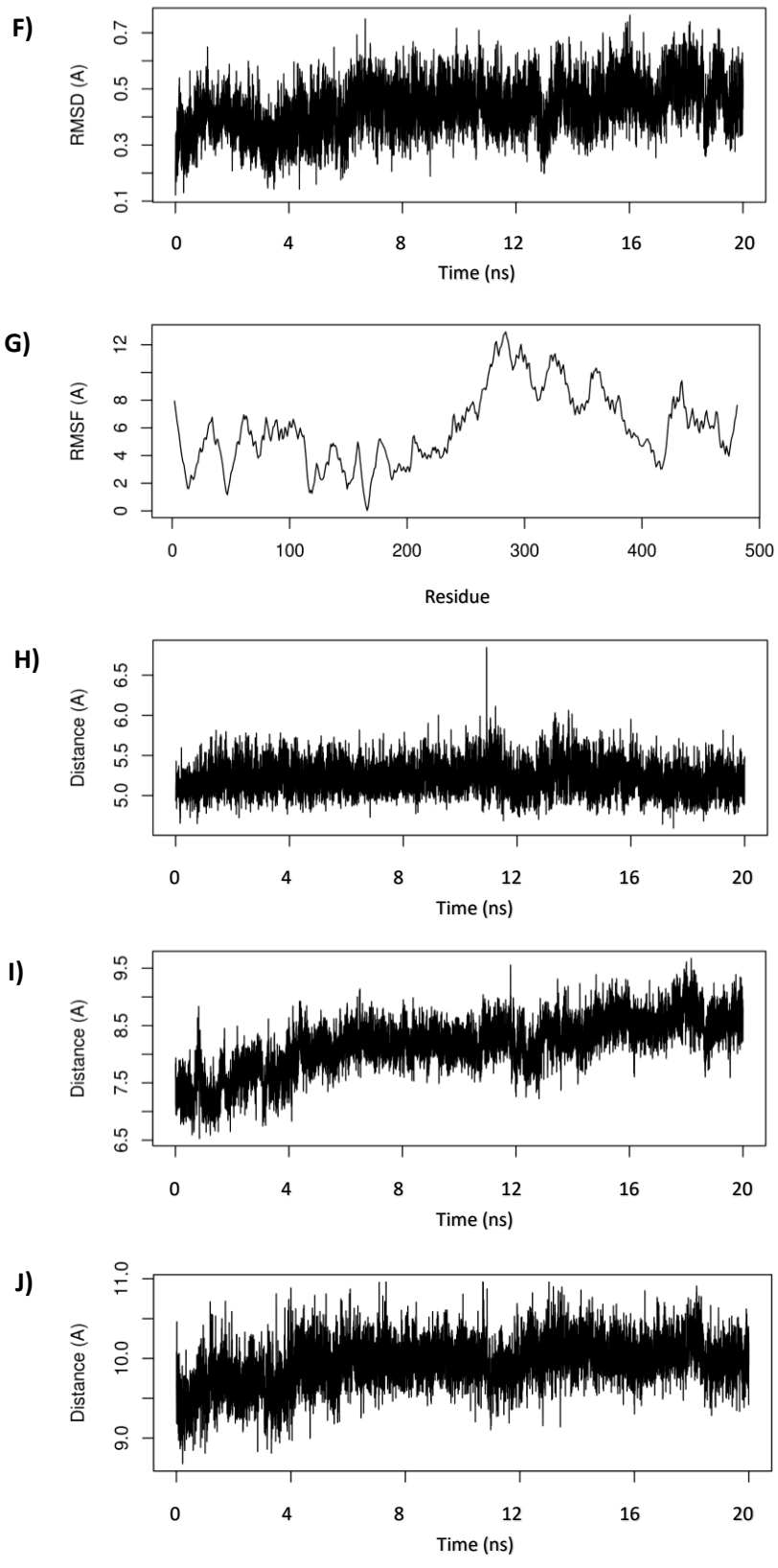


Figure 5.7 (Continued)

F) RMSD of I148M containing loop (residues 147-151)
G) Root mean square fluctuations along the protein chain averaged for full simulation
H) Distance between S46 and D166
I) Distance between S46 and I148
J) Distance between D166 and I148.

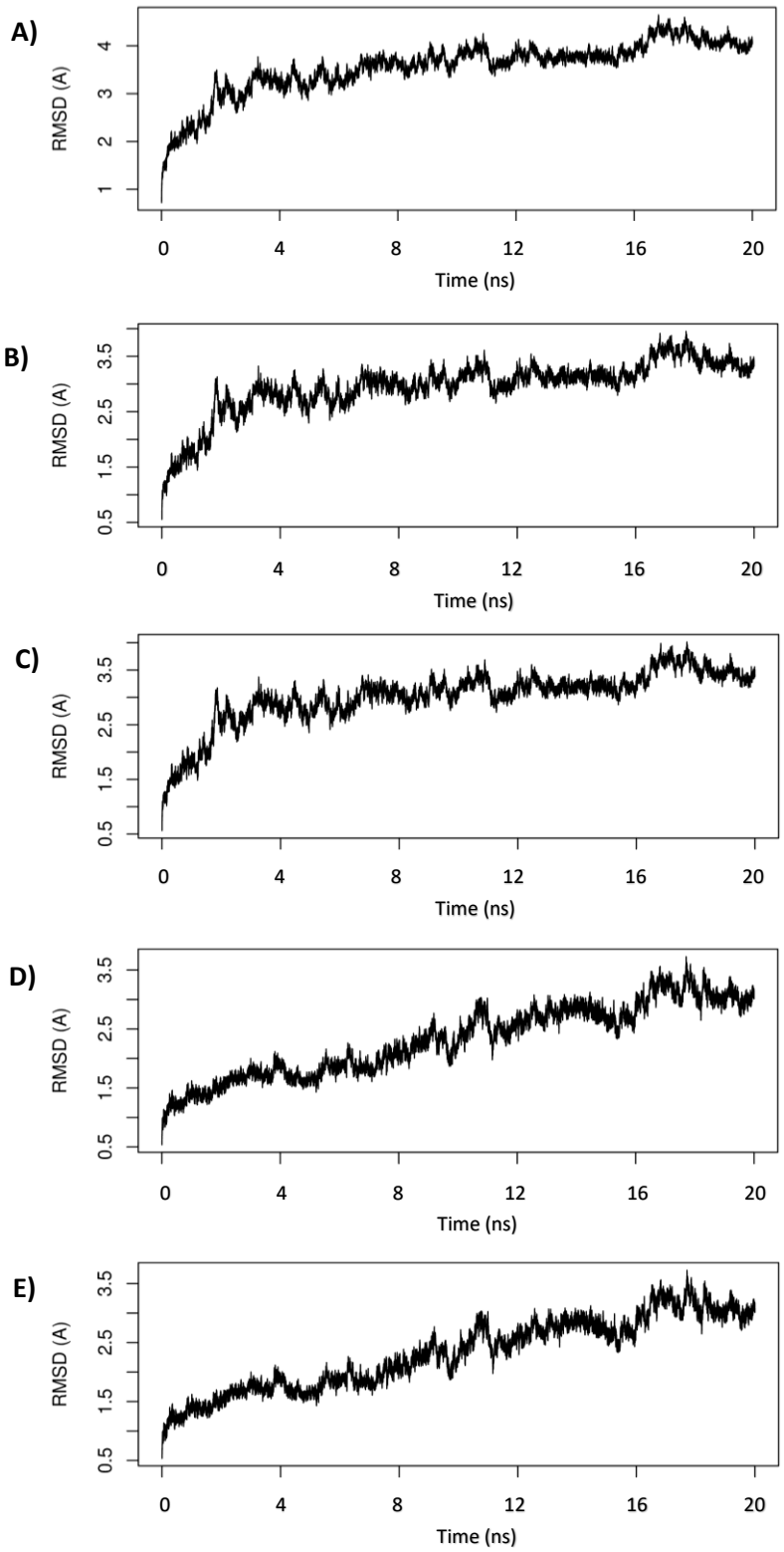


Figure 5.8 Model 6 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.

A) RMSD of entire protein chain

B) RMSD based on backbone chain atoms C, CA and N

C) RMSD based upon α carbons

D) RMSD including only the patatin domain (residues 5-179)

E) RMSD including extended patatin domain (residues 5-239)

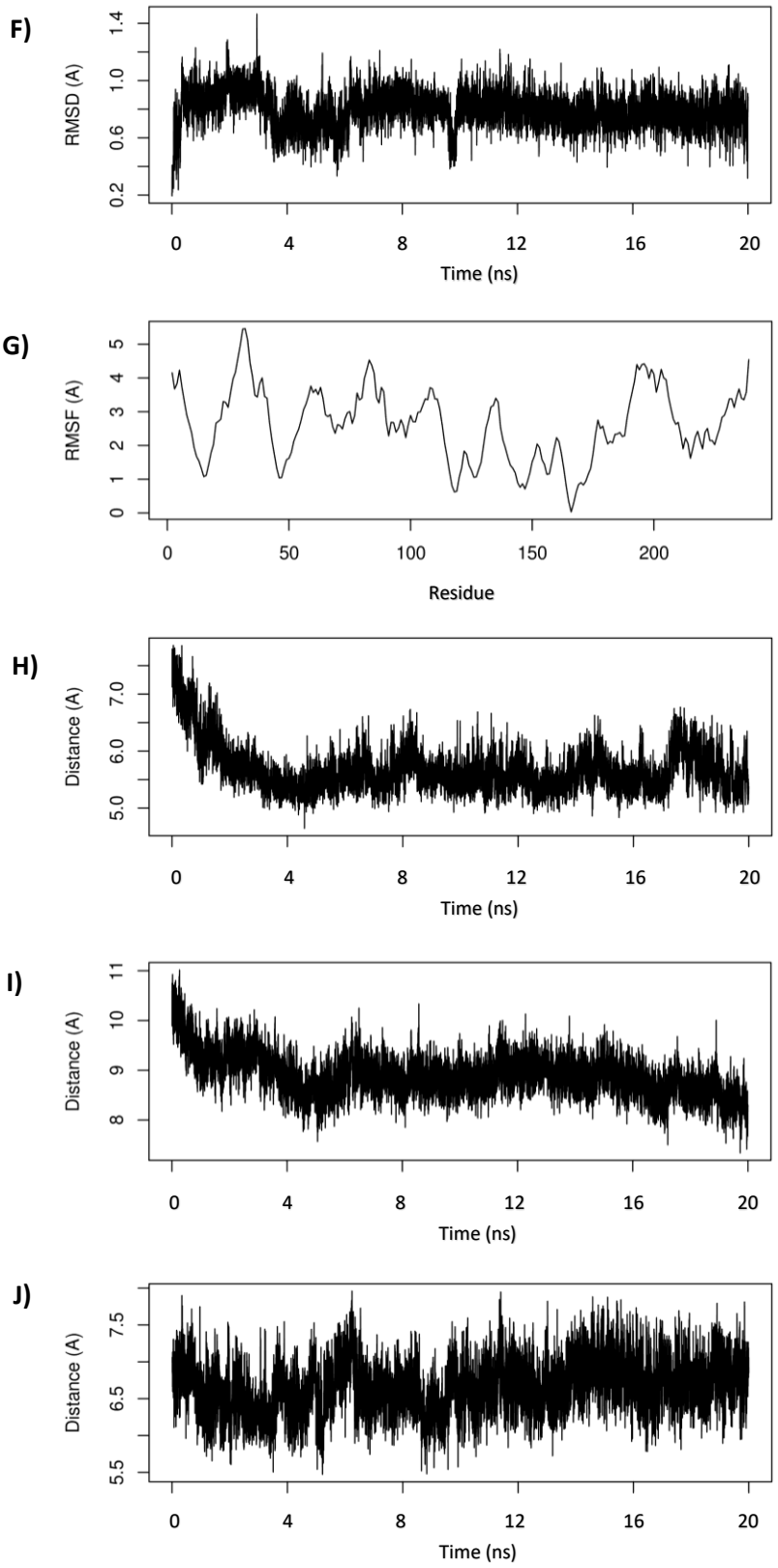


Figure 5.8 (Continued)

F) RMSD of I148M containing loop (residues 147-151)
G) Root mean square fluctuations along the protein chain averaged for full simulation
H) Distance between S46 and D166
I) Distance between S46 and I148
J) Distance between D166 and I148.

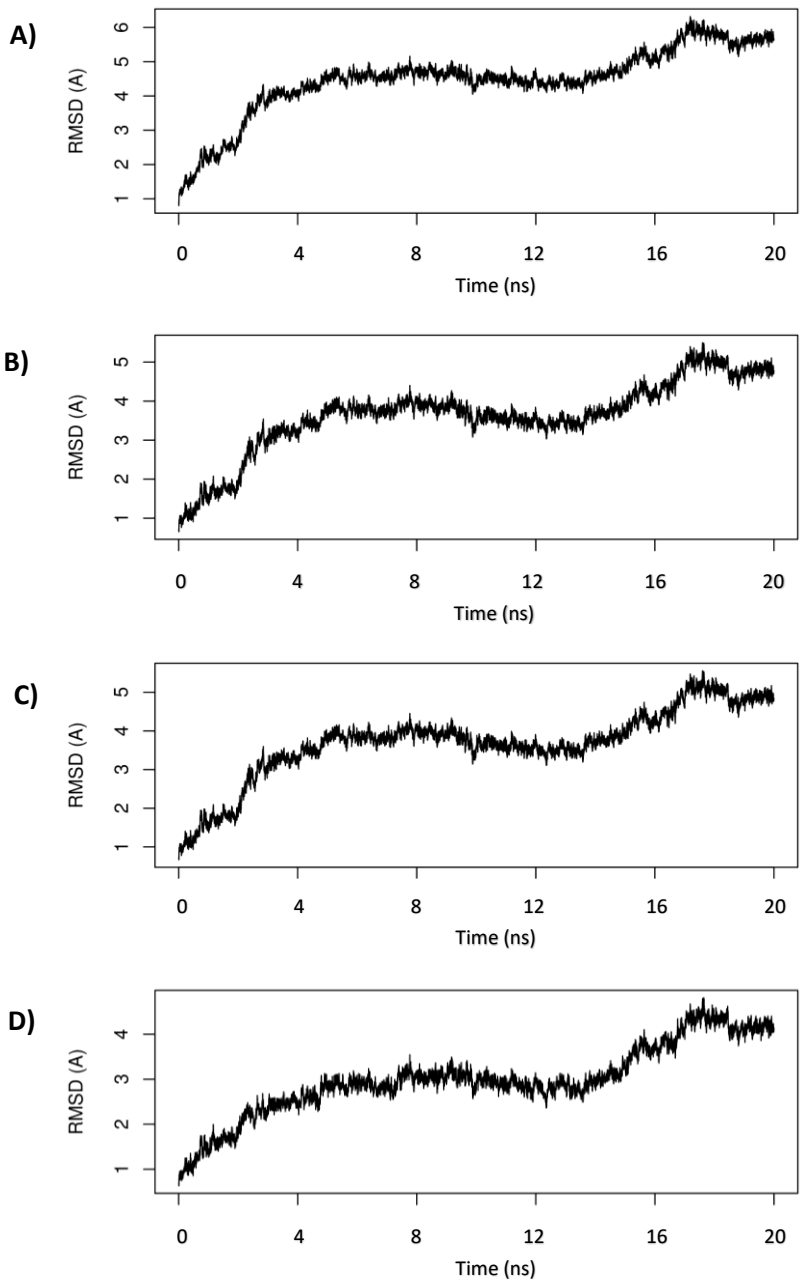


Figure 5.9 Model 7 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.

A) RMSD of entire protein chain

B) RMSD based on backbone chain atoms C, CA and N

C) RMSD based upon α carbons

D) RMSD including only the patatin domain (residues 5-179)

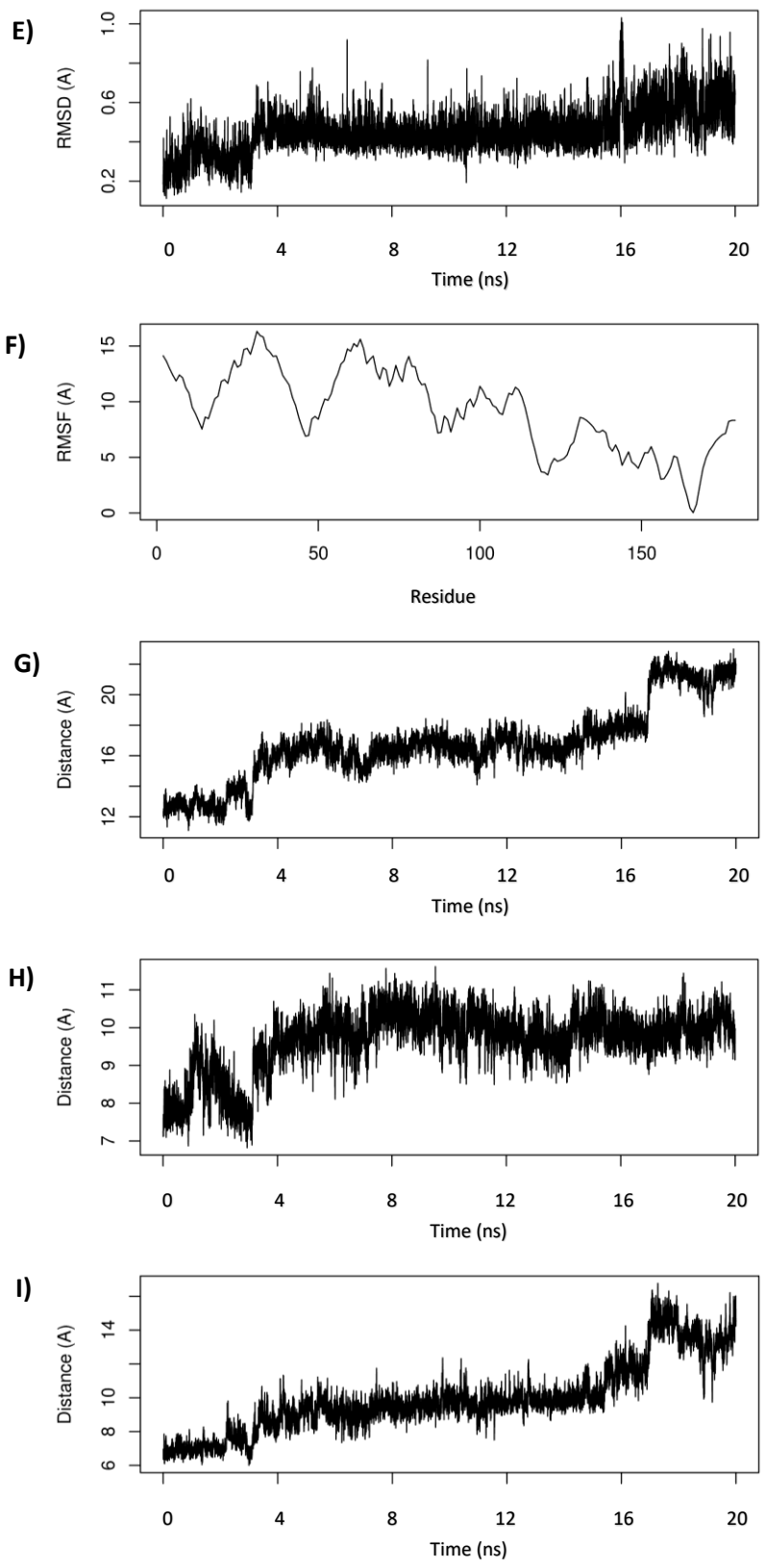


Figure 5.9 (Continued)

E) RMSD of I148M containing loop (residues 147-151)
F) Root mean square fluctuations along the protein chain averaged for full simulation
G) Distance between S46 and D166
H) Distance between S46 and I148
I) Distance between D166 and I148.

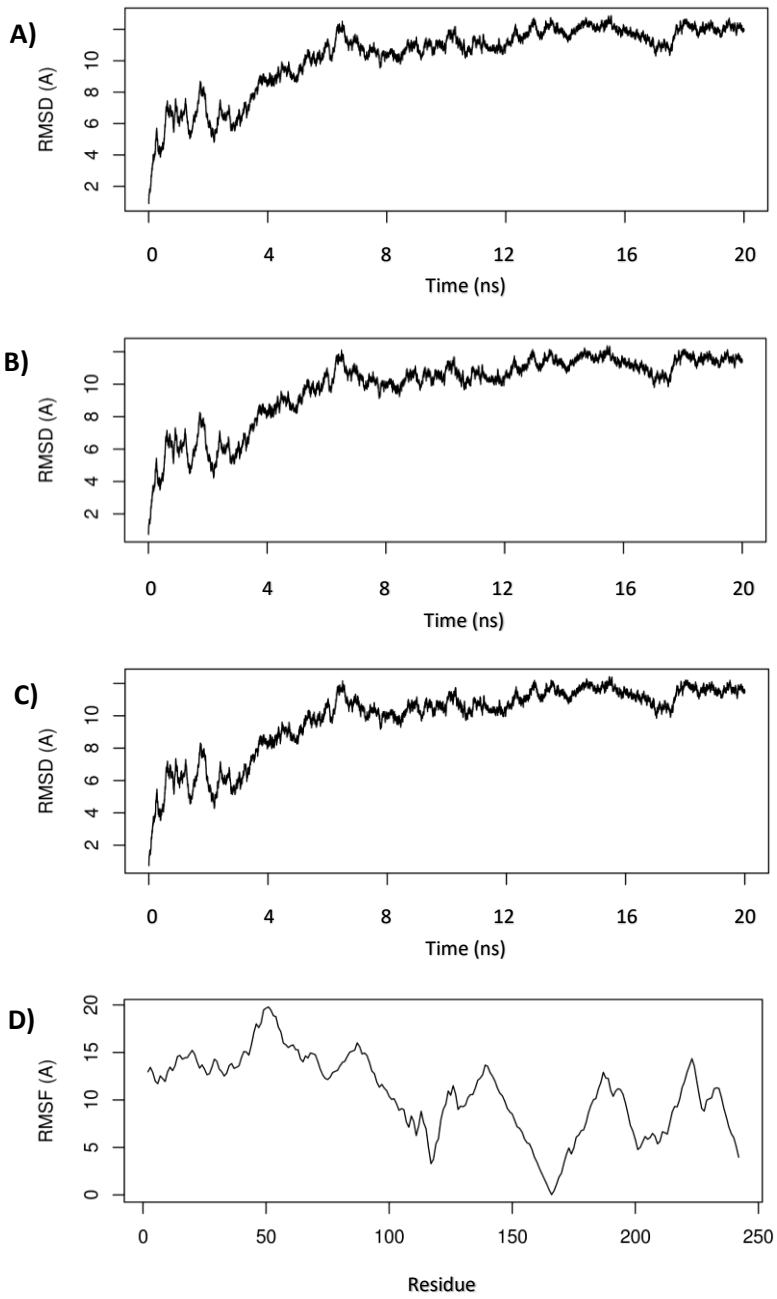


Figure 5.10 Model 8 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.

A) RMSD of entire protein chain

B) RMSD based on backbone chain atoms C, CA and N

C) RMSD based upon α carbons

D) Root mean square fluctuations along the protein chain averaged for full simulation

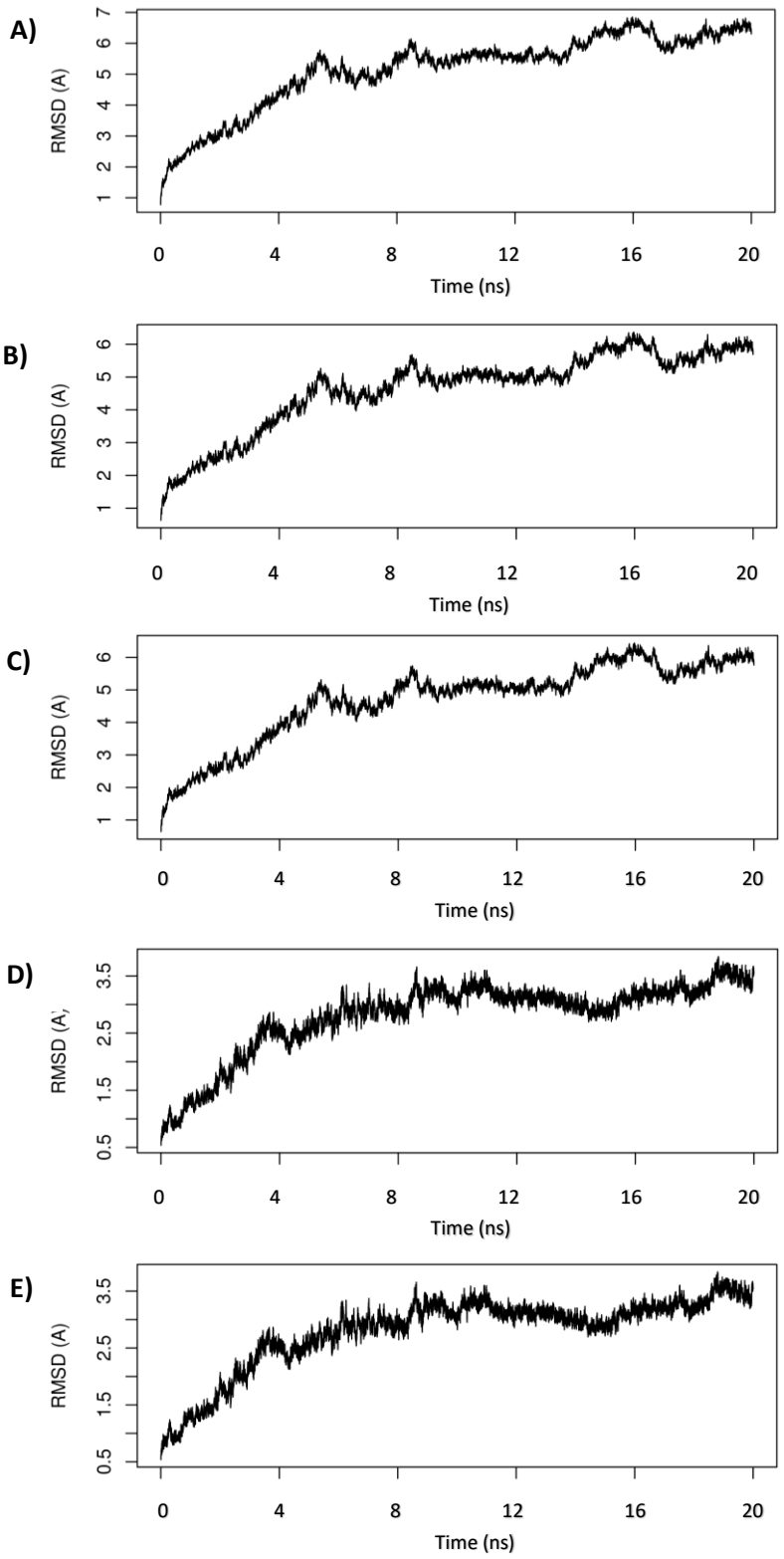


Figure 5.11 Model 9 simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.

A) RMSD of entire protein chain

B) RMSD based on backbone chain atoms C, CA and N

C) RMSD based upon α carbons

D) RMSD including only the patatin domain (residues 5-179)

E) RMSD including extended patatin domain (residues 5-239)

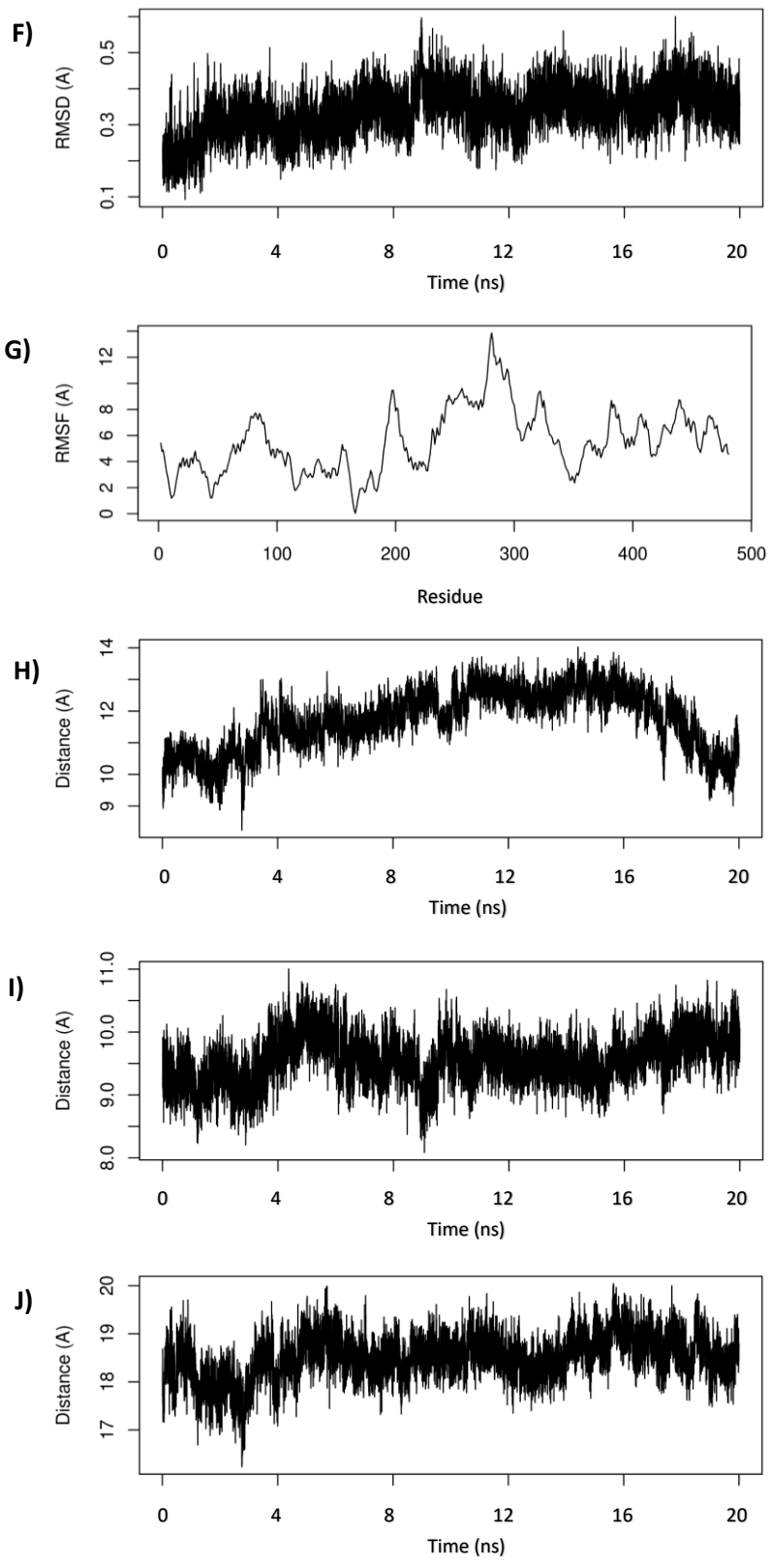


Figure 5.11 (Continued)

F) RMSD of I148M containing loop (residues 147-151)
G) Root mean square fluctuations along the protein chain averaged for full simulation
H) Distance between S46 and D166
I) Distance between S46 and I148
J) Distance between D166 and I148.

5.5.2 Full length simulations

Both proteins remained stable for the duration of the full 100ns simulations. There was a slight decrease in energy across the simulation; however, this did not decrease significantly to dramatically impact the conditions of the systems.

5.5.2.1 Wild type protein:

The RMSD of the entire protein chain showed an increasing RMSD over the first 37,000 frames from 2 to 8A, suggesting a gradual protein structural change. In the last 13,000 frames, a stable final structure appears to have been achieved, with fluctuations of only 0.5A. The RMSD does not decrease significantly when reducing the calculation to only the backbone atoms or α carbons, showing the instability is not due to side chain movements alone (Figure 5.12).

The patatin domain of the protein shows an initial increase to 2.5A in the first 10,000 frames, at which point a stable conformation was achieved in this domain, with fluctuations of only 0.5A. It appears the domain increases in stability along the timeframe of the simulation, reaching a particularly stable point at 30,000 frames. The extended patatin domain has an almost identical RMSD profile.

The RMSD of the flexible loop region, goes through 3 stages during the simulation. It is initially stable and fluctuating between 0.2 and 0.4A over the first 8,500 frames. Between frames 8,500 and 30,000 there is a gradual increase in the RMSD up to 1.4A. At this point the stability is maintained to the end of the simulation.

The RMSF profile is similar to what was observed was similar to the shorter investigatory simulations. The model has a relatively flat RMSF profile along the entire protein chain. Notably the patatin domain of the protein was clearly lower than the C-terminal half of the protein. No regions of the domain were fluctuating abnormally, suggesting a consistent level of stability across the structure. Contrary to earlier investigations, there was a peak of flexibility in the C-terminal 150 residues, which had an RMSF up to 12A.

All of the residues of interest maintained consistent distances from one another throughout the simulation. There was a slight decrease in average distance of the catalytic residues in the first 20,000 frames, and a slight increase between S47 and I148; however, this was less than 1A. The average distance between the catalytic residues was 5A, the distance between S47 and I148 was 8.5A, and between I148 and D166, 10A.

5.5.2.2 Variant protein:

The RMSD of the entire protein chain shows an initial increase over the first 5,000 frames to 4A. At this point, the system appears more stable, with very small fluctuations of 0.5A; however, there was a clear overall increase in RMSD from 4 to 7A over the remainder of the simulation. The RMSD decreased by only 1A when reducing the calculation to only the backbone atoms or α carbons, showing the instability is not due to side chain movements alone (Figure 5.13).

The patatin domain underwent an initial shift over the first 10,000 frames up to 3A RMSD. At this point a highly stable structure was reached, fluctuating by only 0.5A throughout. A slight upward trend in RMSD is still noticeable, but does not appear to be due to significant structural change.

The flexible loop remained stable throughout the simulation, fluctuating between 0.4 and 0.8A RMSD, with a slight upward trend over the final 10,000 frames.

The model has a relatively flat RMSF profile along the entire protein chain. Notably the patatin domain of the protein was clearly lower than the C-terminal half of the protein. No regions of the domain were fluctuating abnormally, suggesting a consistent level of stability across the structure. There was a peak of flexibility in the region of residues 270 to 410, which had an RMSF up to 12A.

The distance between the catalytic residues fluctuated throughout the simulation between 8 and 13A, which appeared to be random movement. Similarly, the distance between S47 and I148, fluctuated between 10 and 14A. The distance between I148 and D166 was much lower, on average only 5.5A, and was consistent throughout the simulation.

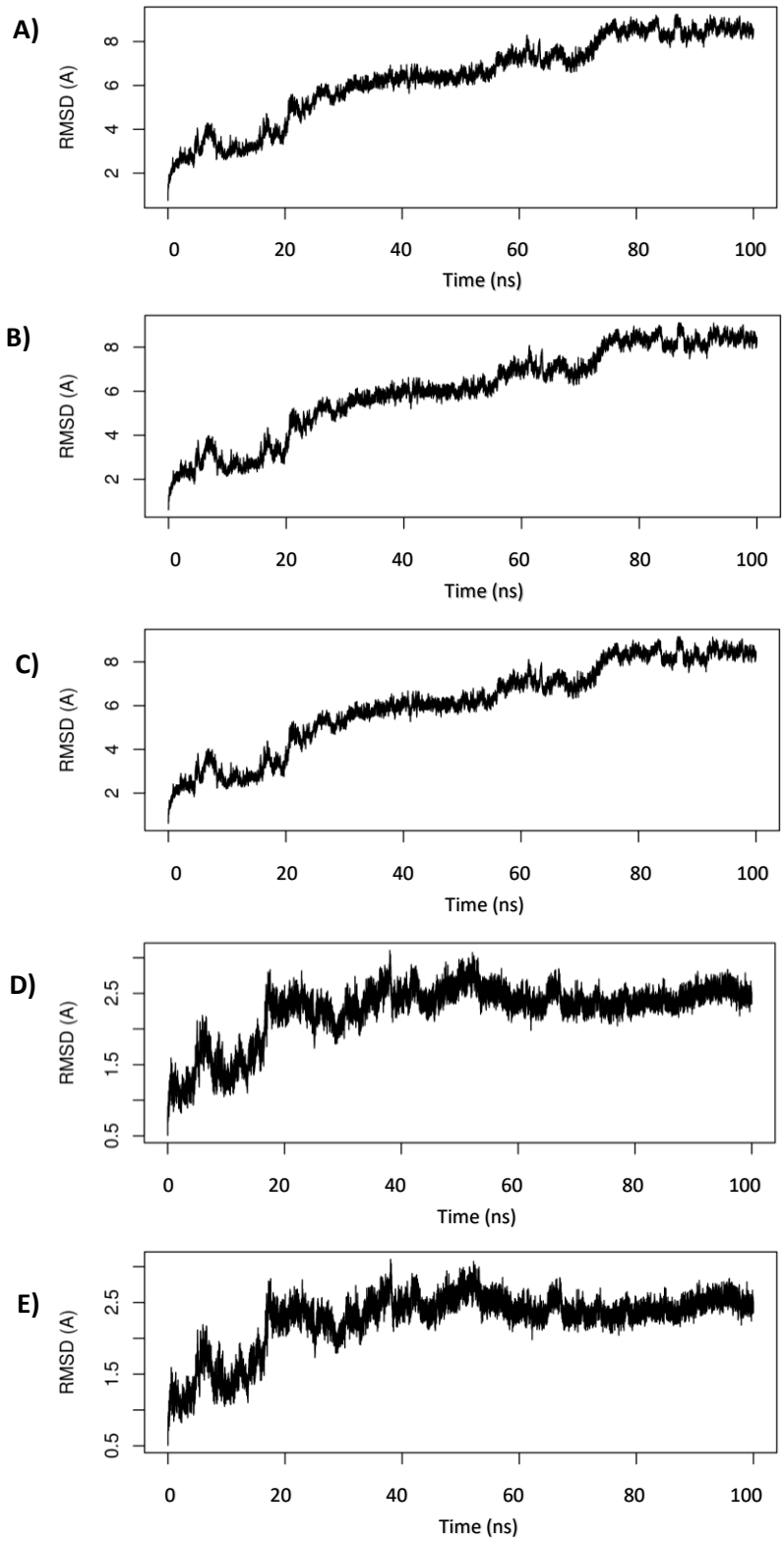


Figure 5.12 Final wild-type simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.

A) RMSD of entire protein chain

B) RMSD based on backbone chain atoms C, CA and N

C) RMSD based upon α carbons

D) RMSD including only the patatin domain (residues 5-179)

E) RMSD including extended patatin domain (residues 5-239)

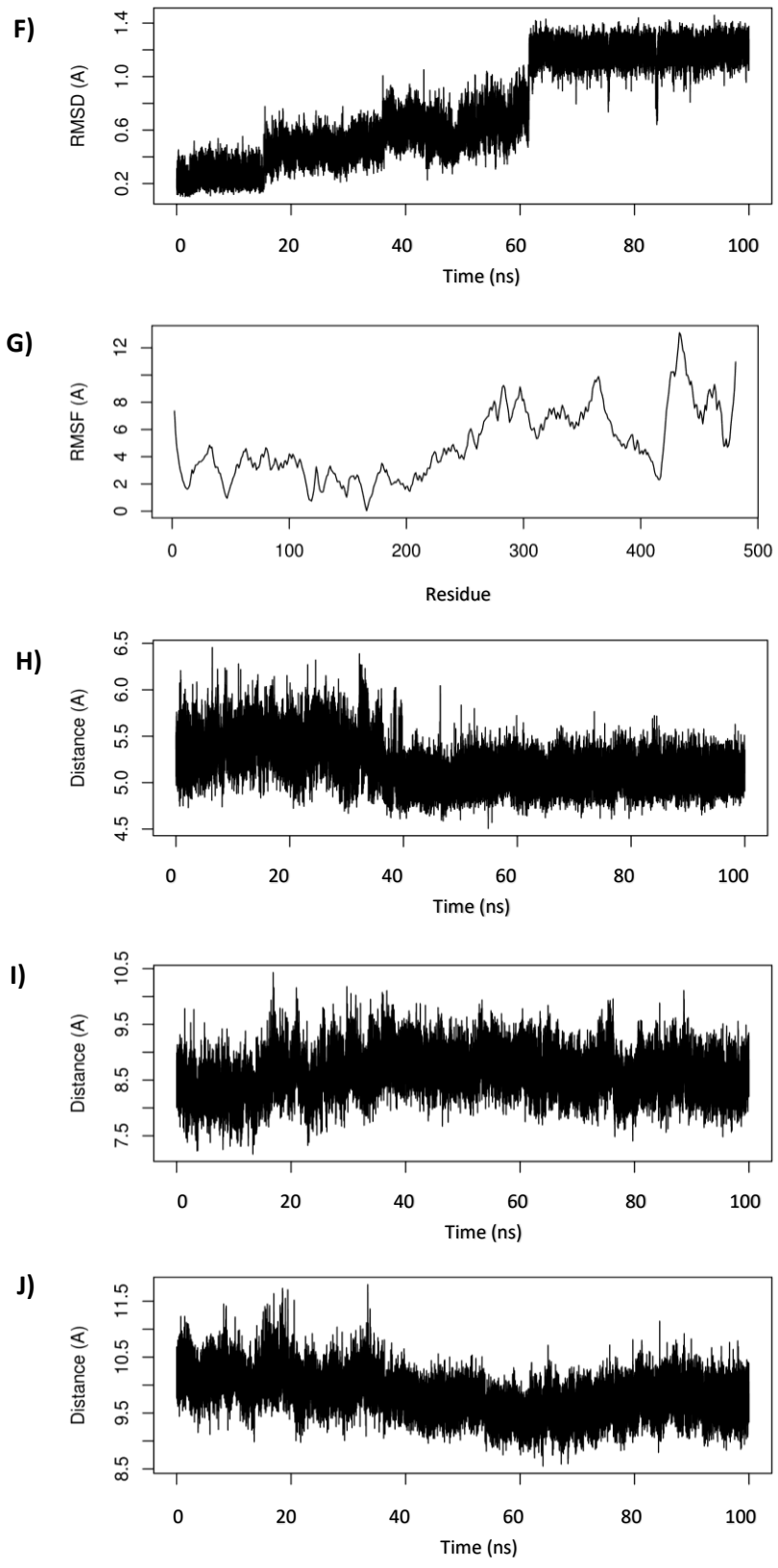


Figure 5.12 (Continued)

F) RMSD of I148M containing loop (residues 147-151)
G) Root mean square fluctuations along the protein chain averaged for full simulation
H) Distance between S46 and D166
I) Distance between S46 and I148
J) Distance between D166 and I148.

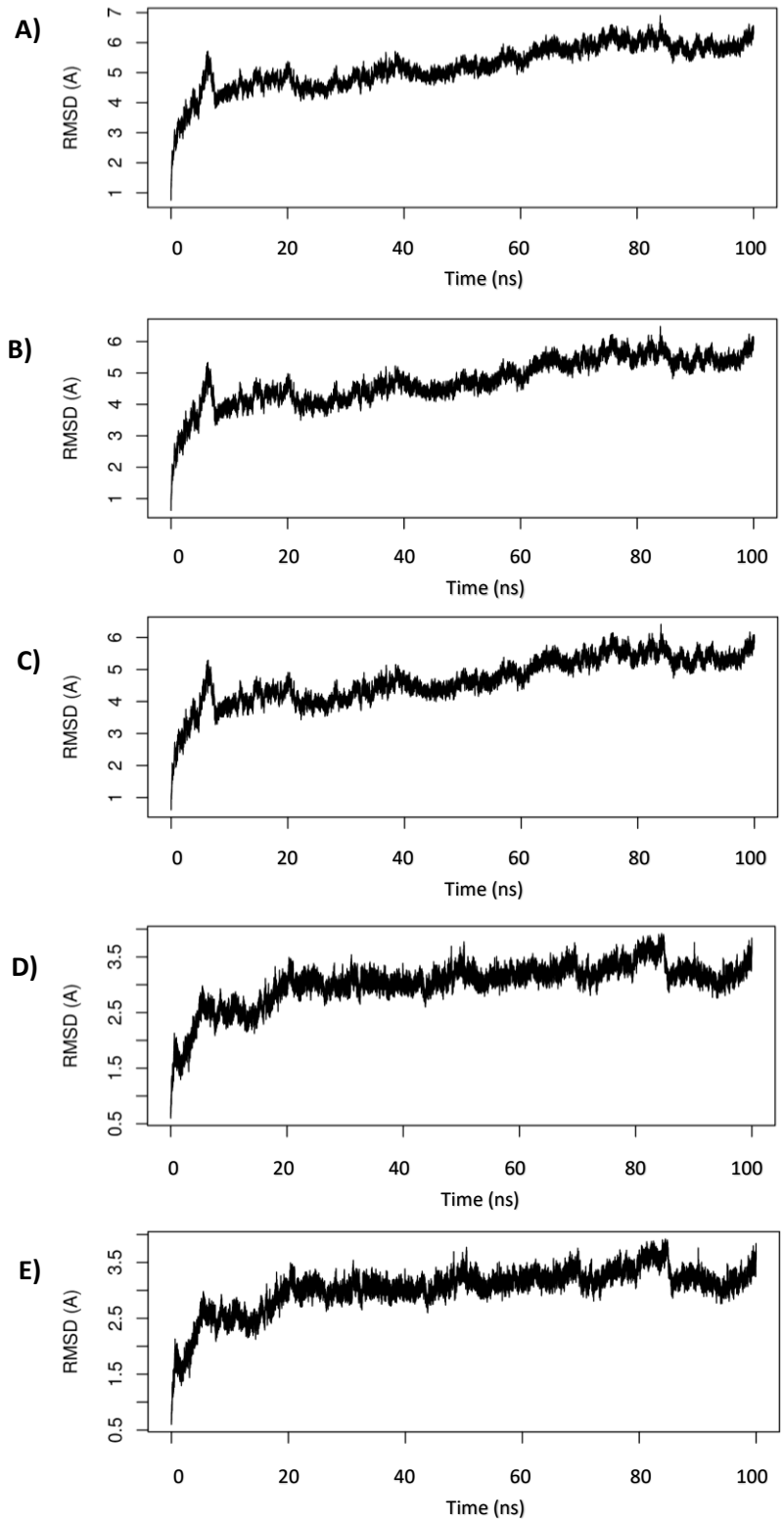


Figure 5.13 Final I148M variant simulation results (cont'd next page)

Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in nanoseconds.

A) RMSD of entire protein chain

B) RMSD based on backbone chain atoms C, CA and N

C) RMSD based upon α carbons

D) RMSD including only the patatin domain (residues 5-179)

E) RMSD including extended patatin domain (residues 5-239)

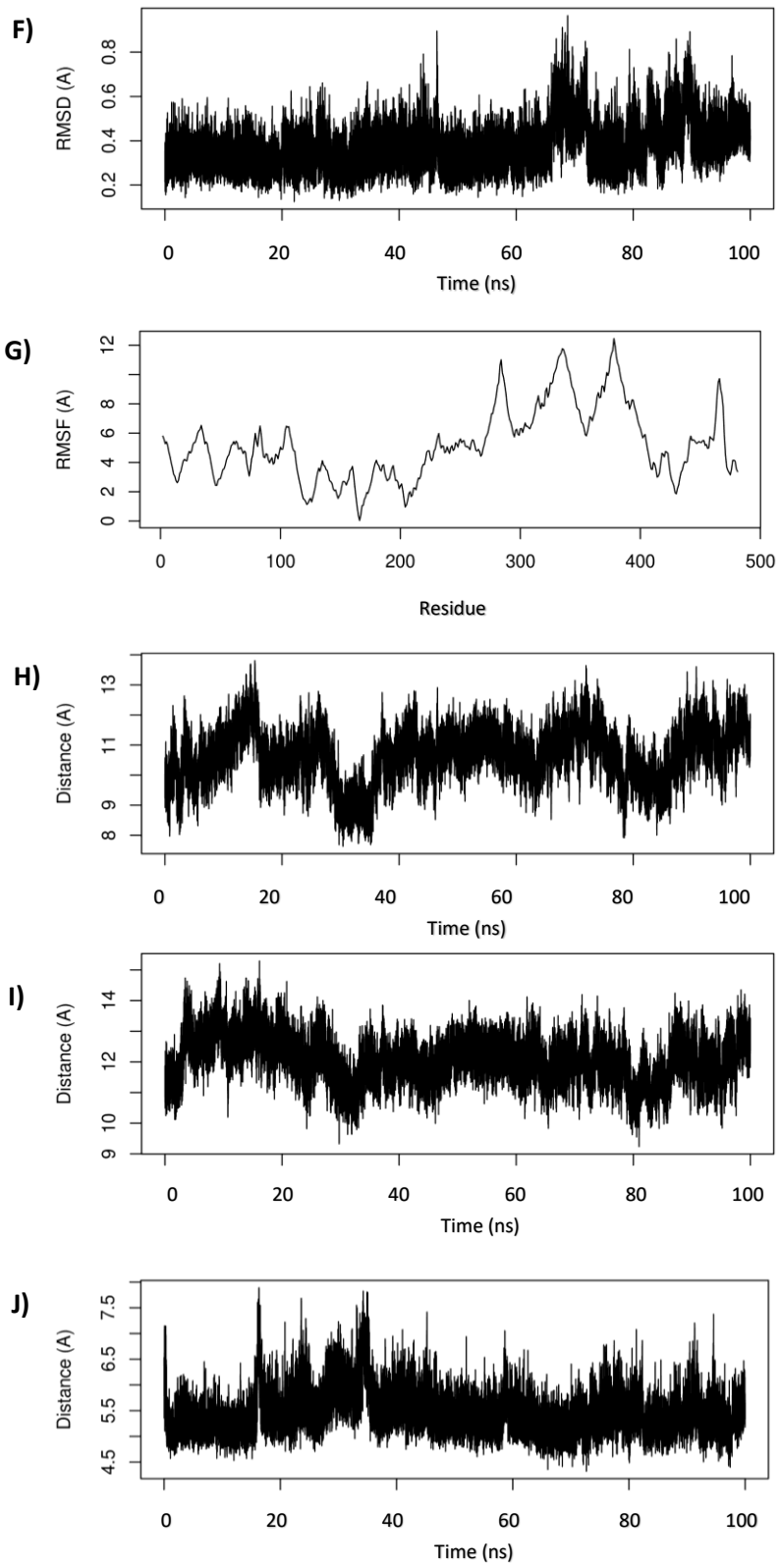


Figure 5.13 (Continued)

F) RMSD of M148 containing loop (residues 147-151)
G) Root mean square fluctuations along the protein chain averaged for full simulation
H) Distance between S46 and D166
I) Distance between S46 and I148
J) Distance between D166 and I148

5.5.3 Simulation structures

5.5.3.1 Domain architecture of PNPLA3 variants

To compare the overarching conformational differences between the wild type and variant protein, the final frame of each simulation was used as the most confident structure for comparison.

When viewed from the front, the structures of PNPLA3 can be said to have a clear separation into two lobes, which are separated by regions of flexible coil. For the purpose of clarity the lobe containing the patatin domain was designated the left lobe, and the other lobe the right lobe.

Both variants of PNPLA3 have similar overall domain architecture, visually consisting of the core patatin domain which makes up the predominant portion of the left lobe, spanning residues 1-256, the right lobe domain, spanning residues 257-390, and the C-terminal domain, spanning residues 391-481, which is folded back to form part of the left lobe (Figure 5.14).

While the overarching structures are similar, there are notably large structural changes between the variants, even when looking at the domain level.

Notably, the wild type protein has a much tighter structure, where the right lobe is compressed into the left lobe, creating a highly globular appearance. Contrary to this, in the I148M variant, the right lobe is extended away from the left lobe of the protein and maintains its own independent α-helix core structure, which is almost independent from the rest of the protein.

This pattern is mirrored in the C-terminal domain, whereby in the wild type it is more compressed into the protein, while the I148M variant extends between the domains, remaining more elongated.

5.5.3.2 Domain movement during simulations

Conformational stability was further assessed visually based on snapshots of the structure taken at 20ns intervals from the start of the simulation. Minimal changes to the domain architecture are observed throughout the simulation between these snapshots for both the wild type or variant structures (Figures 5.15 -5.17).

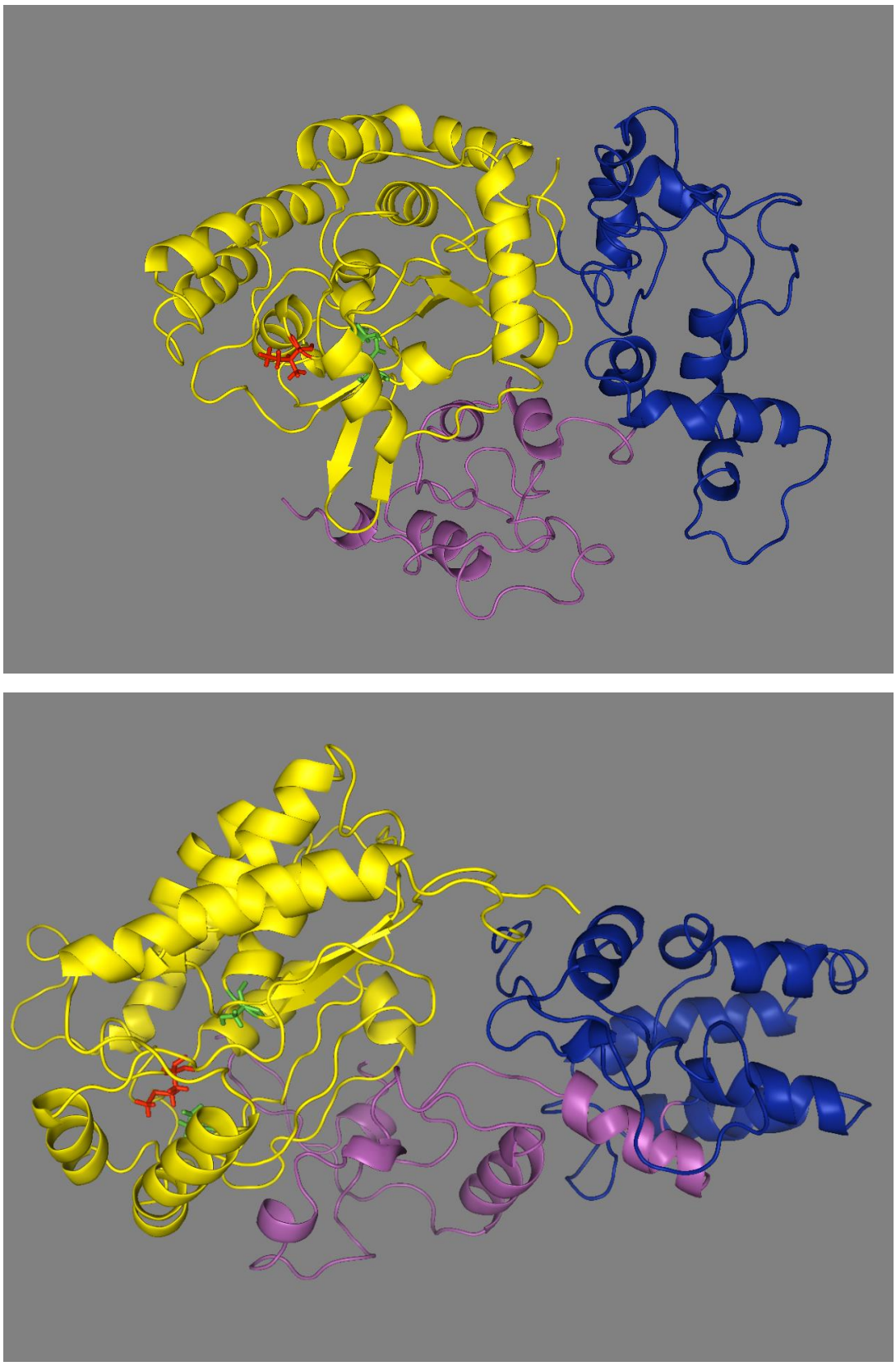


Figure 5.14 Domain architecture of PNPLA3 final conformations

The catalytic residues S47 and D166 are highlighted in green, and residue 148 highlighted in red. The patatin domain (residues 1-256) is coloured in yellow, the right lobe (residues 257-390) are coloured in blue and the C-terminal domain (residues 391-481) coloured magenta.
Top panel; wild type.
Bottom panel; I148M variant.

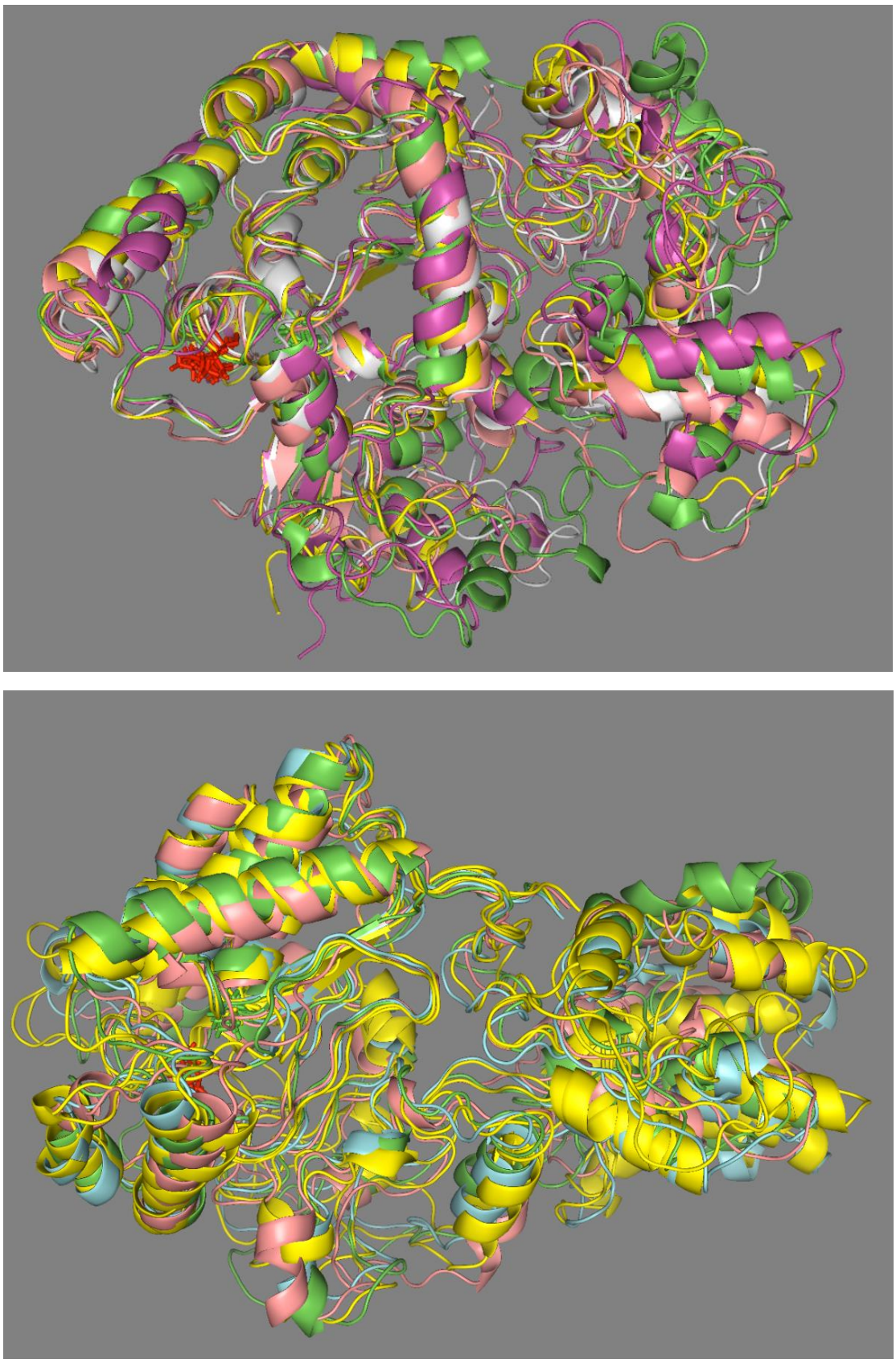


Figure 5.15 Ensemble of simulation snapshots

Snapshots taken at 20, 40, 60, 80 and 100ns timepoints throughout each simulation. The catalytic residues are highlighted in green and residue 148 highlighted in red.
Top panel; wild type protein.
Bottom panel; I148M variant.

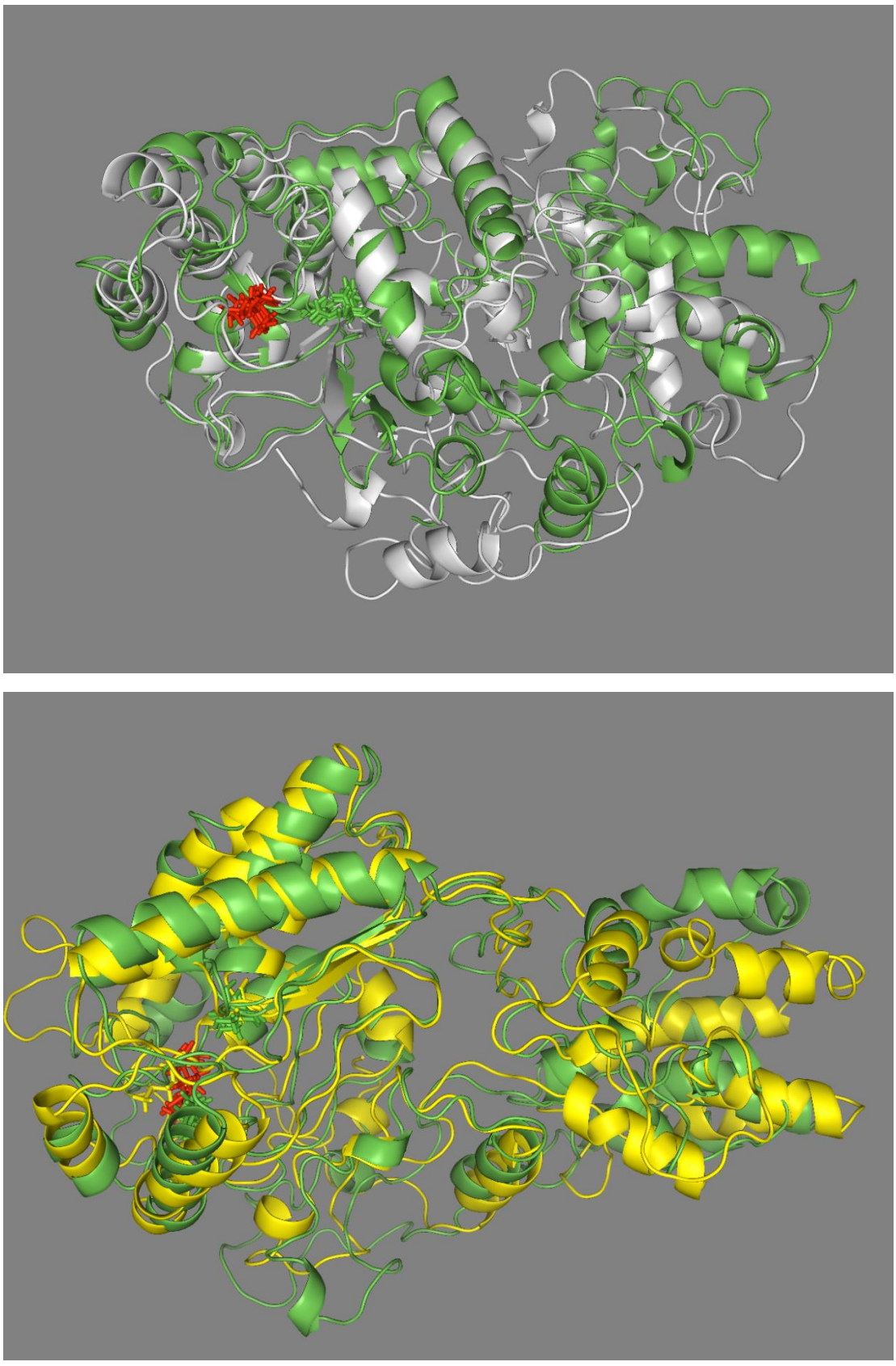


Figure 5.16 Simulation structural changes

Showing the structure of each protein at 20ns and 100ns. The catalytic residues are highlighted in green and residue 148 highlighted in red. The starting structure in each simulation is represented in green.
Top panel; wild type protein
Bottom panel; I148M variant.

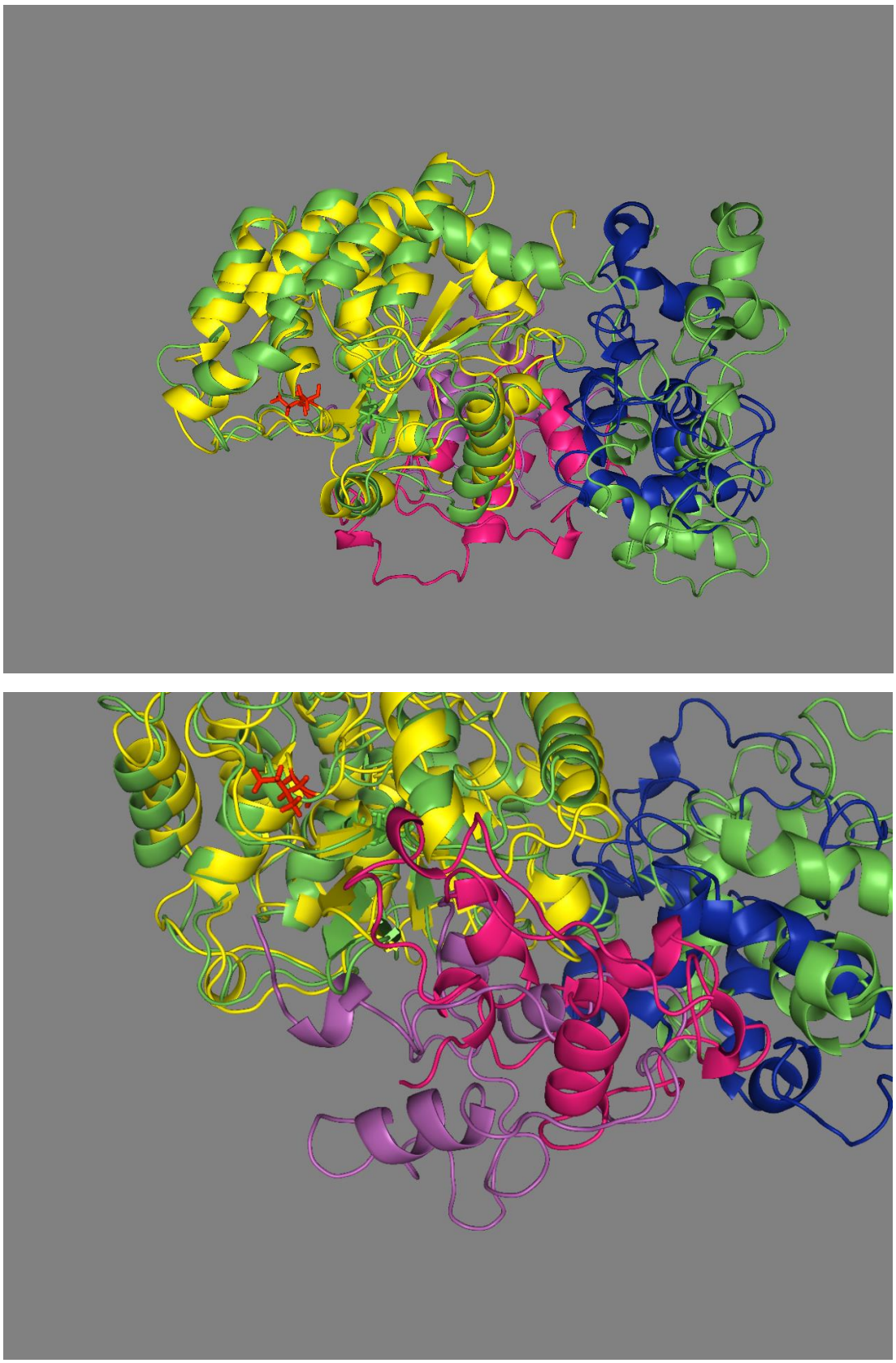


Figure 5.17 Domain movement of wild type PNPLA3 over simulation

The catalytic residues S47 and D166 are highlighted in green, and residue 148 highlighted in red. The patatin domain (residues 1-256) is coloured in yellow, the right lobe (residues 257-390) are coloured in blue and the C-terminal domain (residues 391-481) coloured magenta.
Top panel; wild type.
Bottom panel; I148M variant.

Wild type PNPLA3

The wild type protein displays a high level of stability in the patatin domain, which remains almost identical across snapshots. Notably, there is a more significant domain shift in both the right lobe and C-terminal domains.

Across each snapshot, there is a continuous compression of the right lobe into the patatin domain of the protein. This is coupled with helices in this domain turning to align parallel with the face of the patatin domain.

Additionally, the C-terminal domain transitions across the face of the protein, again moving to form a tighter conformation with the patatin domain, and interfacing between the left and right lobes of the protein.

Notably the C-terminal domain was also shown to transition on the surface of the patatin domain in an alternate simulation; however, to a different degree (results not shown). This suggests the C-terminal domain has inherent flexibility in this structure.

I148M variant

In the I148M variant, the patatin domain also remains stable throughout, and there are no notable changes during the timeframe of the simulation. The I148M variant is overall more stable and displays a much greater degree of stability in the C-terminal domains, where only the flexible loop regions ostensibly alter their position.

20ns structure shown in green. 100 ns structure coloured by domain, the patatin domain (residues 1-256) are coloured in yellow, the right lobe (residues 257-390) are coloured in blue and the C-terminal domain (residues 391-481) coloured magenta.

5.5.3.3 Detailed description of structural differences between variants

When comparing the final structures of each variant, based upon the last frames of the simulation, there are significant differences between the wild type and I148M variants of PNPLA3. These changes range from small alterations around the active site, but also extend to the entire domain structure (Figure 5.18 – 5.19).

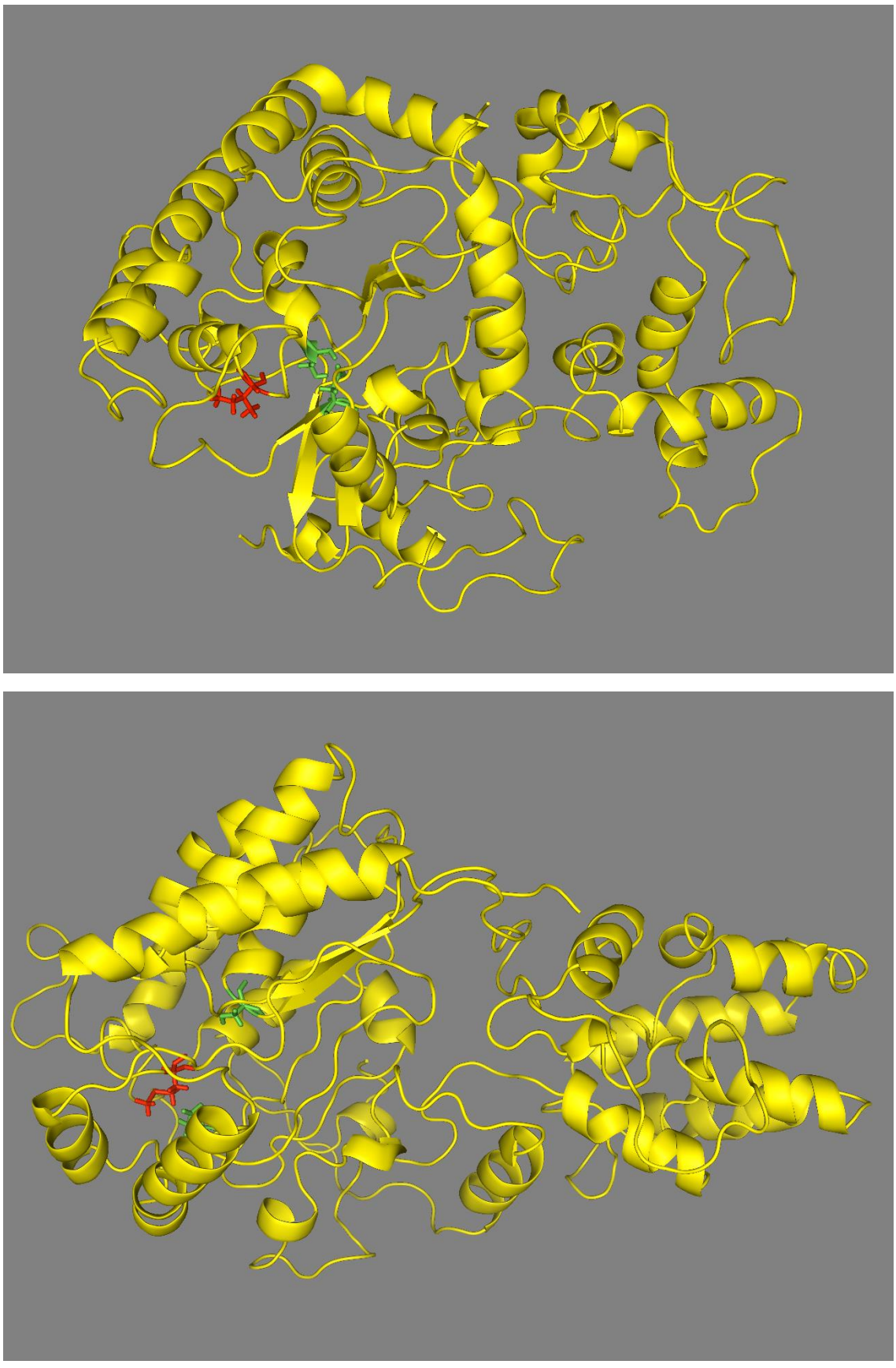


Figure 5.18 Final conformational structures of PNPLA3 after full 100ns simulation

The catalytic residues S47 and D166 are highlighted in green, and residue 148 highlighted in red.
Top panel; wild type.
Bottom panel; I148M variant.

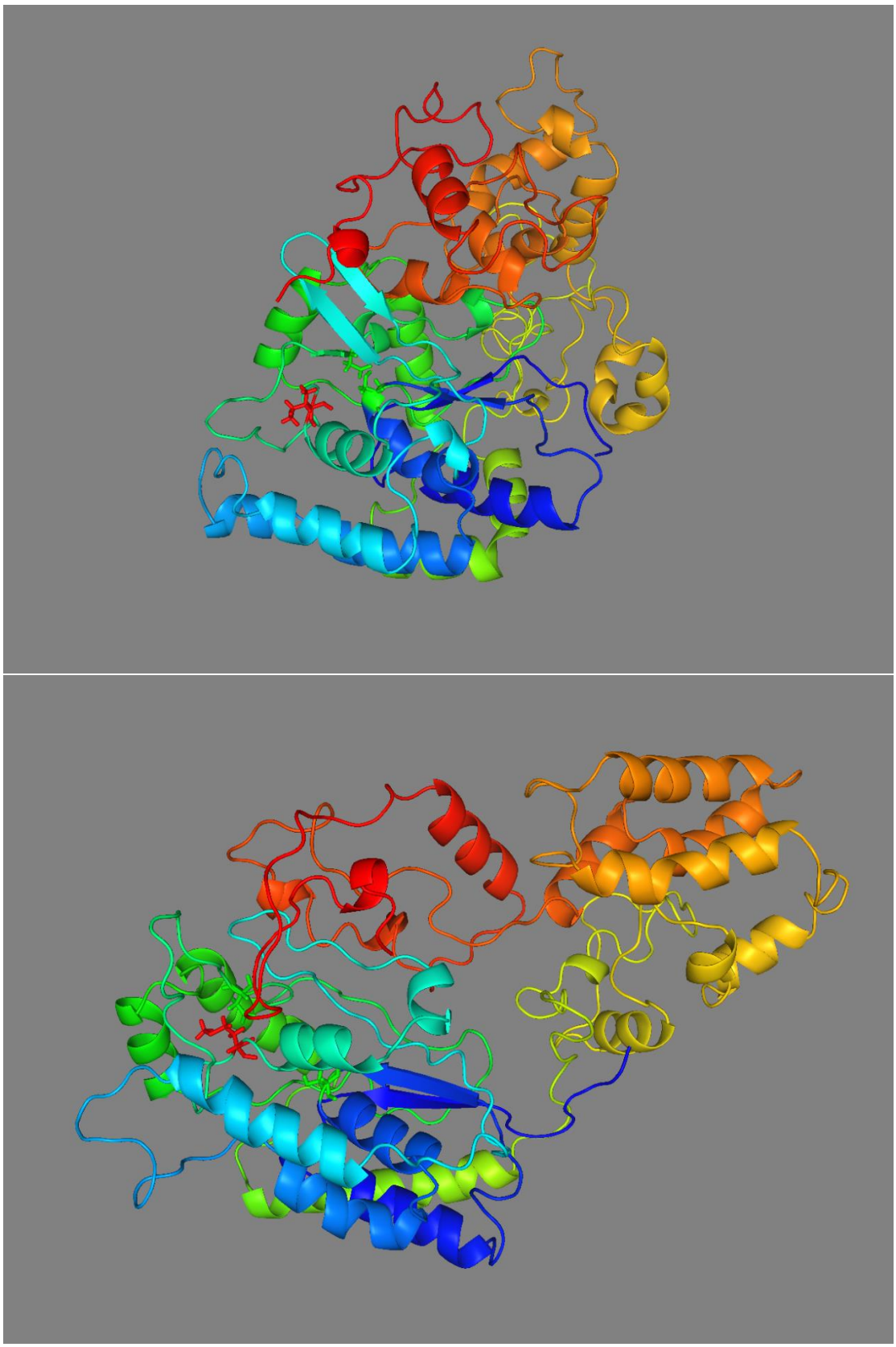


Figure 5.19 Final conformational structures of PNPLA3 after full 100ns simulation

The catalytic residues S47 and D166 are highlighted in green, and residue 148 highlighted in red. Both proteins are coloured from N to C terminal ranging from blue to red respectively.
Top panel; wild type.
Bottom panel; I148M variant.

Patatin domain

In both variants, there is an initial β-strand (strand 1), helix (helix A), β-strand (strand 2), helix structure (helix B). At the hinge region between the strand 2 and helix B is the catalytic serine

In the I148M variant helix B after S47 folds perfectly back on the strands, contributing to the helix cluster in the region, while in the wild type, this helix has lost some helical character and curves strongly upward.

The wild type protein follows by 2 helices heading across the protein surface (Helix C and D), followed by a helix which folds back adding to the helix core structure in the patatin domain (helix E). In the variant protein, helix C and D are a single long helix, which appears more stable than in the wild type. Helix E in the wild type is held further forward from the core because of the upward turn in helix B.

At this point there is a large portion of random coil, which in the wild type loops over helix B, and runs parallel to the β-strand core formed by strand 1 and 2. The coil then juts almost perpendicular out from the catalytic serine and forms 2 anti-parallel β-strands (strands 3 and 4), followed by a loop which traces back on the inside of the previous section.

This is the first significant difference between the proteins, as in the variant protein, the random coil, follows along β-strand 2, but then continues straight forward with a loss of β-sheet character, before folding back to trace back along itself to the base.

In the wild type this is followed by an α-helix (Helix F), which is angled downward running in the opposite direction along the inside of helix E, and above helix B. Residue 148 is located at the terminal end of helix F. Following I148, there is a short flexible loop which coils down to interact with loop adjoining helix D and E, before turning over itself and forming a short β-strand (Strand 5) which sits antiparallel to strand 3. The catalytic aspartate (D166) is the final residue of strand 5. It appears this β-strand unit consisting of Strands 3, 4 and 5 helps form the stable catalytic site, positioning S47 and D166 in very close proximity.

In the variant structure, the loop turns in and helix F is positioned similarly to the wildtype, however, it is not spatially constrained by the upward helix 3, and so extends beyond the active site, again terminating with Met 148. After residue 148 the large flexible loop is more loosely packed, contacting only helix E and folding out to the surface of the protein. Notably it does not form strand 5, and again this strand character is lost from the structure. Instead the loop fold back on itself, holding Met 148 and Asp 166 in close proximity.

Both variants then have a large coil region, which cups around strands 1 and 2 in the structure, followed by 2 antiparallel helices (helix G and H), which cover the active site pocket. In the wild type these helices are tightly in toward the protein centre, packed under strands 3 and 4, whereas in the I148M variant, they come in under the loop between 148 and 166.

An additional helix in the wild type protein (helix I), runs initially parallel to helix A, before curving at a 90^{o} angle around the outside of helix A, to form an almost triangular base with the loop.

In the methionine variant, helix I is completely straight and runs parallel to the first helices throughout.

The difference in the angle of this helix is notable, because it is descriptive of the angle of the other domain of the protein, which differs strongly in direction. In the wild type the right lobe of the protein extends across the face of strands 1 and 2, whereas in the I148M variant, this domain extends directly parallel beyond the base of strand 1 and 2.

Mid domain (right lobe)

The C-terminal portions from the patatin domain of each protein variant are dramatically different spatially.

In the wild type there is a small turn inwards and along helix H, followed by a long coil region which folds away from the patatin domain, before folding back around to form a large curved α-helix (Helix J), which interact with the N-terminal residues. There are then three more small helices which form a small cluster in the right lobe (Helix K, L and M). Notably helices J, K and M are all aligned with helix H, forming a compressed right lobe structure.

In the variant, the initial coil folds much further up along the patatin domain, before folding back to form a small helix J, which barely contacts the N-terminal. Following this there are an additional 6 helices (helix K, L, M, N, O, P) which form an ordered left lobe, all running perpendicular to the helix core of the patatin domain. The overall architecture of the lobe is then far more extended than the wild type protein and has more clear overall secondary structure.

C-terminal domain

In the final C-terminal region of the protein, the wild type has two small helices (helix Q and R) which tuck into the cleft between the right and left lobes and run parallel to helix M. followed

by a coil which runs under β-strands 3 and 4, interacting with the post aspartate coil and helix G. It is clear this is helping to support the elevated β-strand position of strands 2 and 3.

This is followed by a large region of random coil which folds back over helices Q and R, and forms a final helix (helix S), which folds back over the coil maintaining a small clustered structure. This is terminated with an additional long coil region, which runs toward the right lobe of the protein, before folding back to sit on the opposite side of strands 3 and 4. Overall in this model this positions the C-terminal domain sandwiching strand 3 and 4 on both sides offering structural support.

In the variant form of the protein, the coil extends straight along the lost β-strand coil before folding back on itself. A single helix is then formed which sites in the cleft between the lobes of the protein and the final c terminal residues coil over the patatin domain and rest between the helix before residue 148 and the loop after, further stabilising the active site position.

5.5.3.4 Active site changes

Due to the putative catalytic changes caused by the I148M variant, changes to the catalytic region is of particular interest and also one of the largest areas of structural difference between the two variants (Figures 5.20 - 5.21).

In the wild type a very well defined catalytic dyad is observed, whereby the conserved catalytic serine (S47) and aspartate (D166) are in close proximity (< 5A). This contact is maintained throughout the entire simulation (Figure 5.22 – 5.23).

This is similar in structure to the stereotypical patatin site which has been observed in crystal structures and facilitates polar contacts to be maintained between the oxygen of aspartate 166 and serine 47.

Contrasting this, the I148M variant has significantly altered architecture, in which the catalytic residues are now separated by over 10A. In this conformation methionine 148 is in close proximity to the aspartate (D166) throughout.

In addition to the change in distances between these key residues, there is an additional change in the secondary structure in the local area. The wild type forms a short β-strand from residues 163 to 166, there is a clear loss of β-sheet character upon changing the variant to methionine. This is evidenced by a change in recognition of the PyMOL secondary structure.

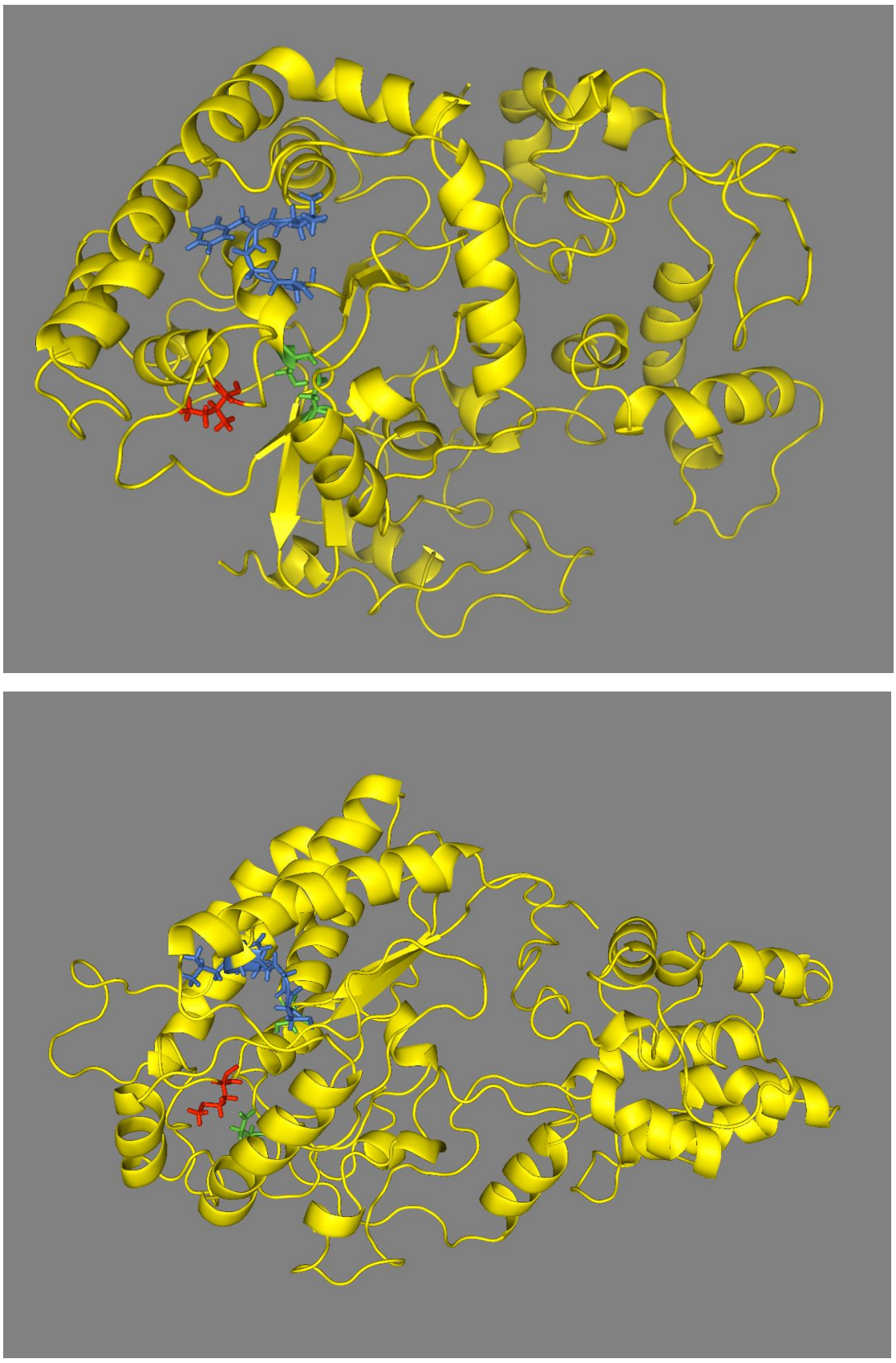


Figure 5.20 Final conformational structures of PNPLA3 highlighting oxyanion hole

The catalytic residues S47 and D166 are highlighted in green, and residue 148 highlighted in red and the residues contributing to oxyanion hole in blue.
Top panel; wild type.
Bottom panel; I148M variant.

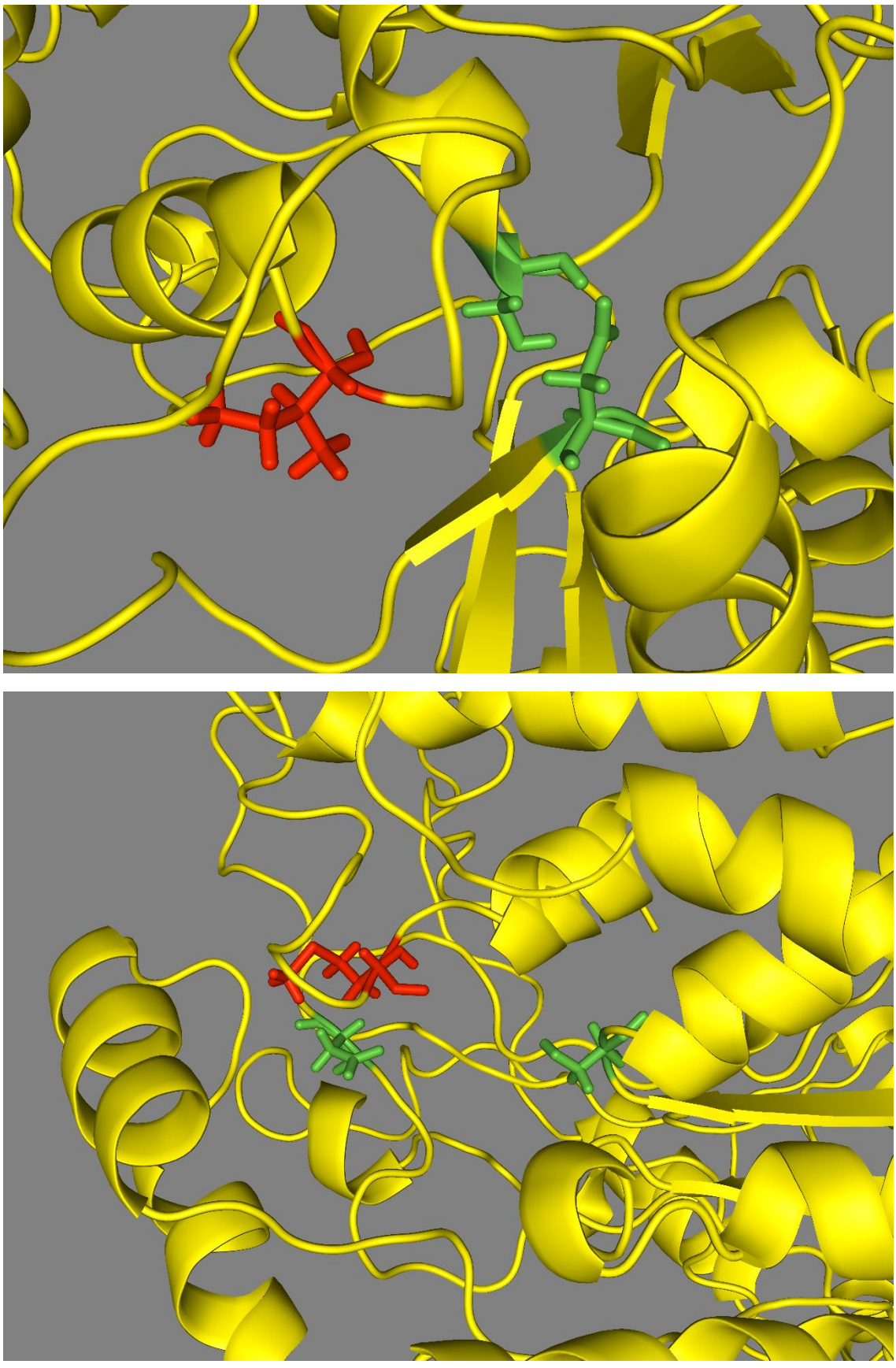


Figure 5.21 Active site of final simulated structures

The catalytic residues (S47 and D166) are highlighted in green and residue 148 highlighted in red.
Top panel; wild type protein.
Bottom panel; I148M variant.

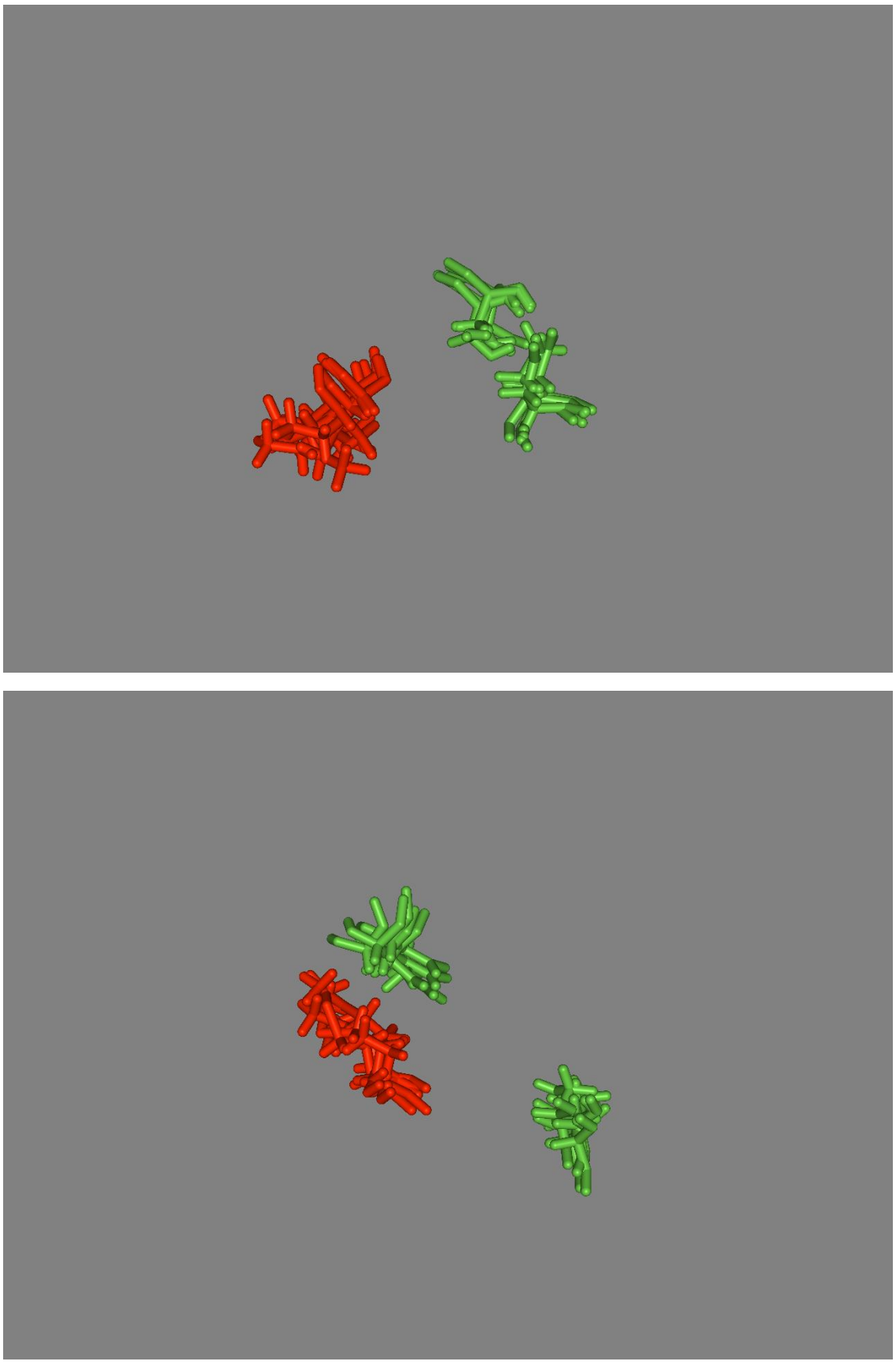


Figure 5.22 Ensemble of active site residue snapshots, taken at 20, 40, 60, 80 and 100ns

The catalytic residues (S47 and D166) are highlighted in green and residue 148 highlighted in red.
Top panel; wild type protein.
Bottom panel; I148M variant.

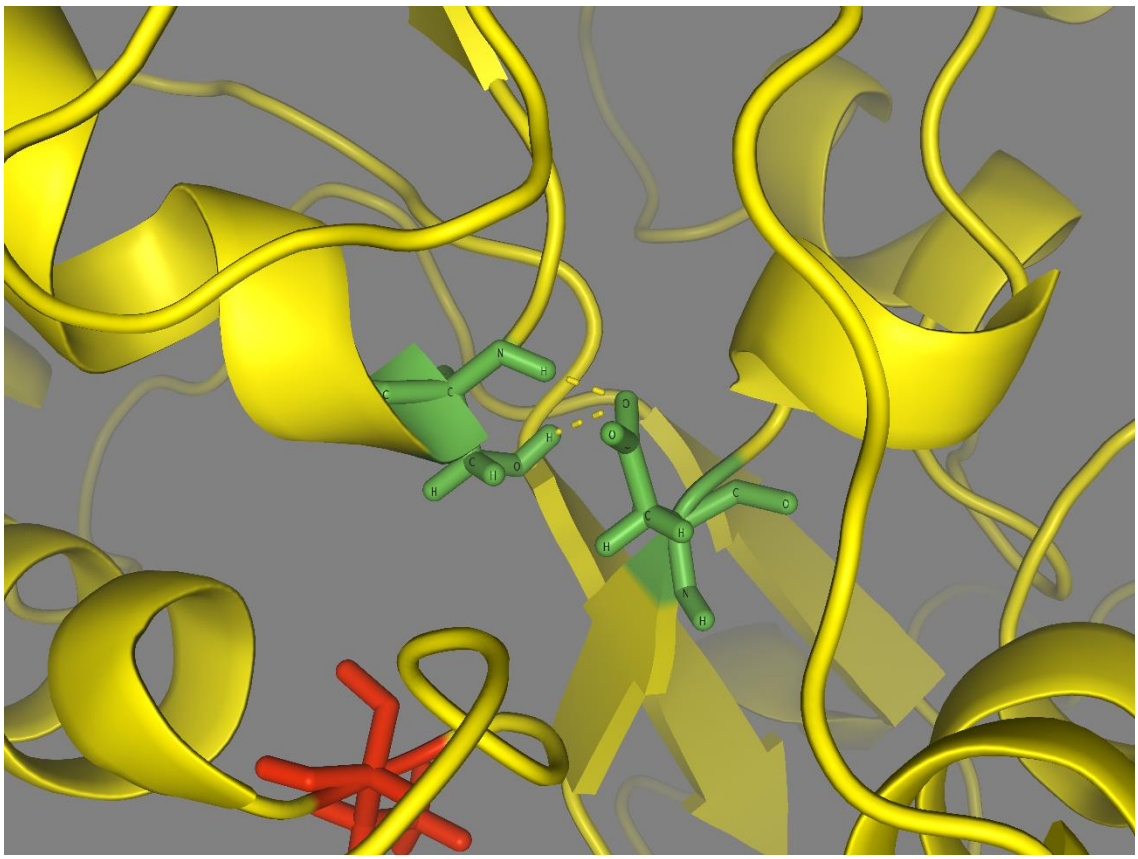


Figure 5.23 Interaction between catalytic residues in wild type PNPLA3 final structure
The catalytic residues S47 and D166 are highlighted in green, and residue 148 highlighted in red. Polar contacts are highlighted with yellow dashes.

Both of these simulations have the highest stability and highest confidence in the patatin domain. In both simulations, these three key residues show very little spatial fluctuation, as evidenced by both the RMSF and imaging of the simulation.

5.5.3.5 Position of the oxyanion hole

Like the catalytic residues, the putative oxyanion hole, formed from the conserved sequence of residues 15-18 is altered between the PNPLA3 variants (Figure 5.20).

In the wild type protein, these residues are positioned close to the active site, with the positive back bone amides are positioned toward the catalytic binding cavity.

In the I148M variant these residues are also positioned within the binding cavity, however are positioned with the positive amides turned away from the catalytic residues.

5.5.3.6 Initial changes in simulated structures

As the active sites of both simulations during production were extremely stable, earlier stages in the simulation were investigated to detect when differences in the conformations of each variant were observed.

Both systems start in the same position, with no notable issues with positioning or spatial conflicts. However, the simulations of both active site regions behave differently almost immediately on simulation, beginning with the initial heating of the system.

On heating, the wild type protein displays an improvement in the active site architecture, in which the isoleucine moves away from the catalytic residues, which simultaneously move closer together forming the basis of the catalytic dyad (Figure 5.24).

However, in the I148M variant, the methionine transitions toward the catalytic residues, disrupting the interaction between them, instead forming its own interaction with the catalytic aspartate, leading to increase distance between the catalytic residues (Figure 5.25).

Along with the rotation observed in the wild type about D166, there is already a secondary structural change, where we can observe the short β-strand (strand 5) up to D166. We do not see this strand formation in the I148 variant.

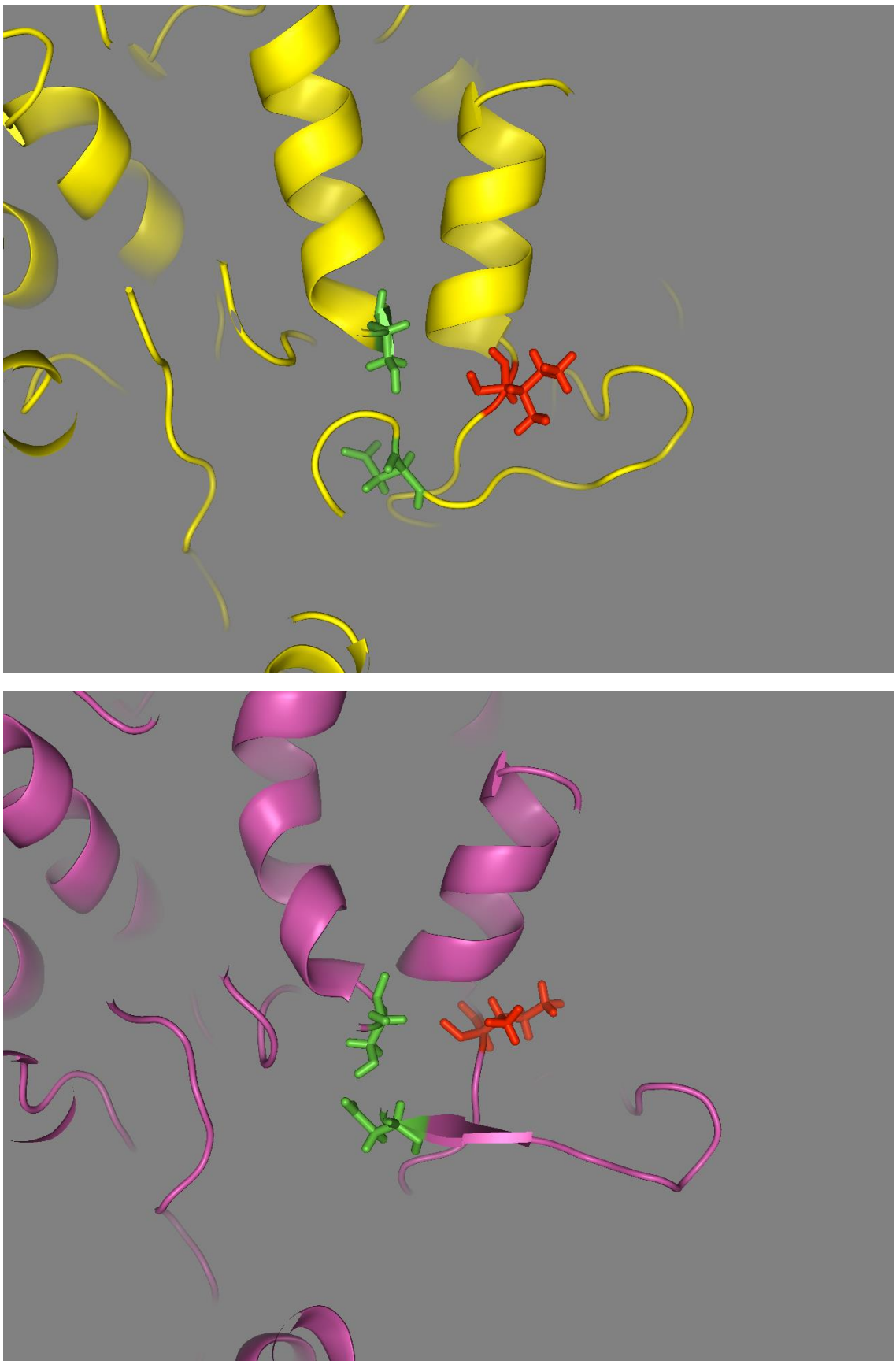


Figure 5.24 Minimised and heated wild type PNPLA3 active site

The catalytic residues (S47 and D166) are highlighted in green and residue 148 highlighted in red.
Top panel; The active site structure after minimisation.
Bottom panel; The active site structure after heating.

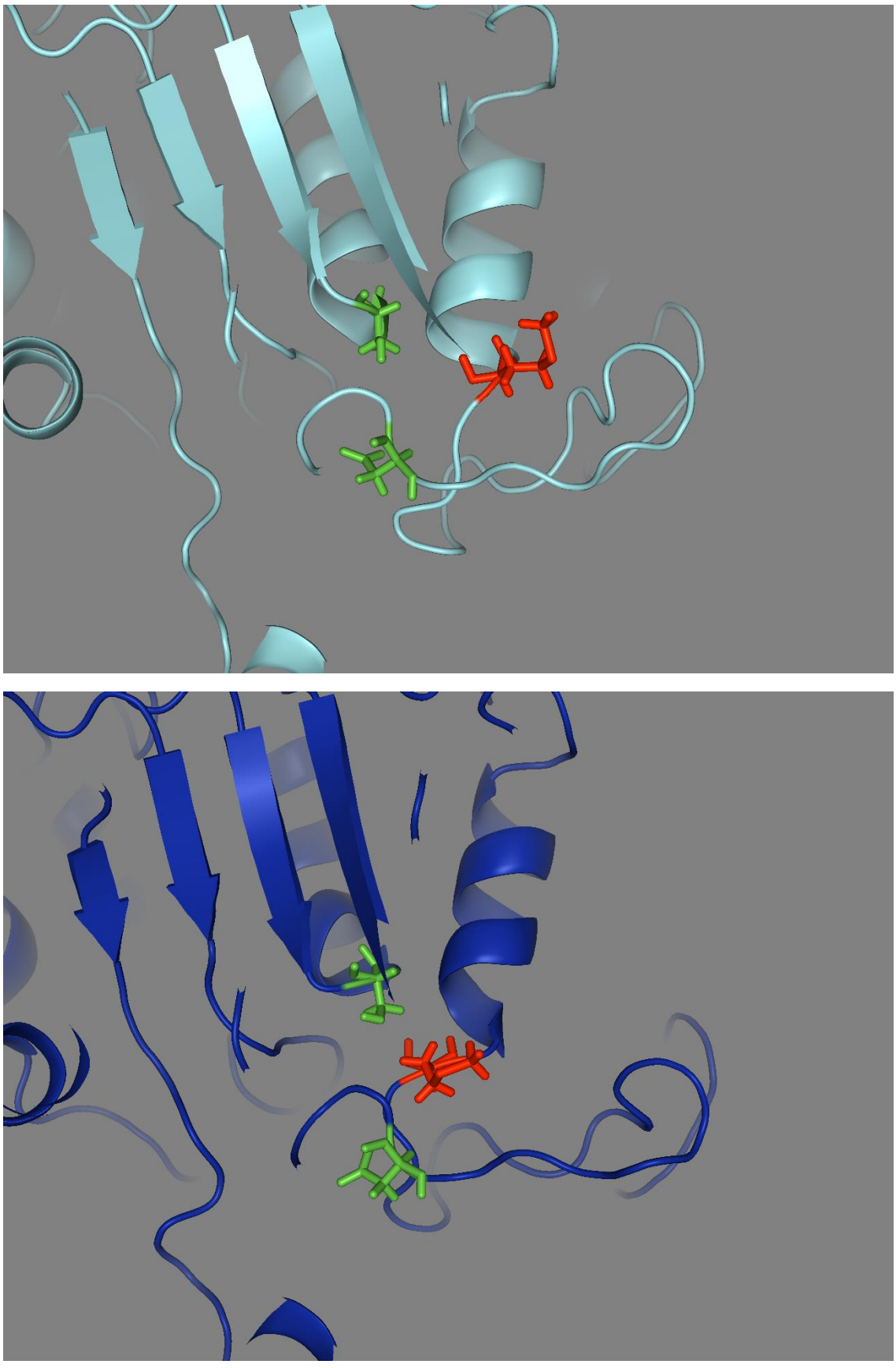


Figure 5.25 Minimised and heated PNPLA3 I148M variant active site

The catalytic residues (S47 and D166) are highlighted in green and residue 148 highlighted in red.
Top panel; The active site structure after minimisation.
Bottom panel; The active site structure after heating.

5.5.4 Tunnels

Tunnels to the catalytic serine were predicted to further assess the impact of differences observed between PNPLA3 variants on access to the catalytic residues.

Both variants had clear access to the catalytic serine, and the wild type had four tunnels which led to S47, while the I148M variant had two tunnels (Figure 5.26).

5.5.4.1 Structural tunnel locations

Wild type PNPLA3

The first tunnel (tunnel 1) was the shortest tunnel and is positioned on the far side of the coil located after D166 in the structure. This Forms an exit out of a large opening before helix I, the C-terminal of coil F and coil C. This is largest opening to the active site in the wild type protein and is positioned directly facing the conserved oxyanion hole (Figure 5.27).

The second tunnel (tunnel 2) is a substantially longer tunnel located again at the loop after D166. In fact, the first portion of this tunnel is shared with tunnel 1, however the tunnel then branches off in the opposite direction. The tunnel is framed by the coiling loop after D166 and between helices G and H.

The third tunnel (Tunnel 3) is located on the opposite side to tunnels 1 and 2. It is positioned below strand 3 and 4 and above helix F, being framed by these secondary structures.

The final tunnel (tunnel 4) is located just before strand 5, however this tunnel has less access to the active site of the protein, beginning 4A away from the serine.

I148M variant PNPLA3

The tunnels in the I148M variant are all located at the large active site cavity in the structure. Specifically, they are framed by the Loop between helix C and D on the right, helix A at the bottom, helix G and preceding loop on the left, and the loop containing Met 148 on the top.

The first tunnel (tunnel 1) in the variant, protrudes straight out of the cavity and past the conserved oxyanion hole, while the second tunnel (tunnel 2) is actually oriented slightly in toward the cavity. Ultimately both tunnels lie in the same cavity, very close to the surface of the protein (Figure 5.28).

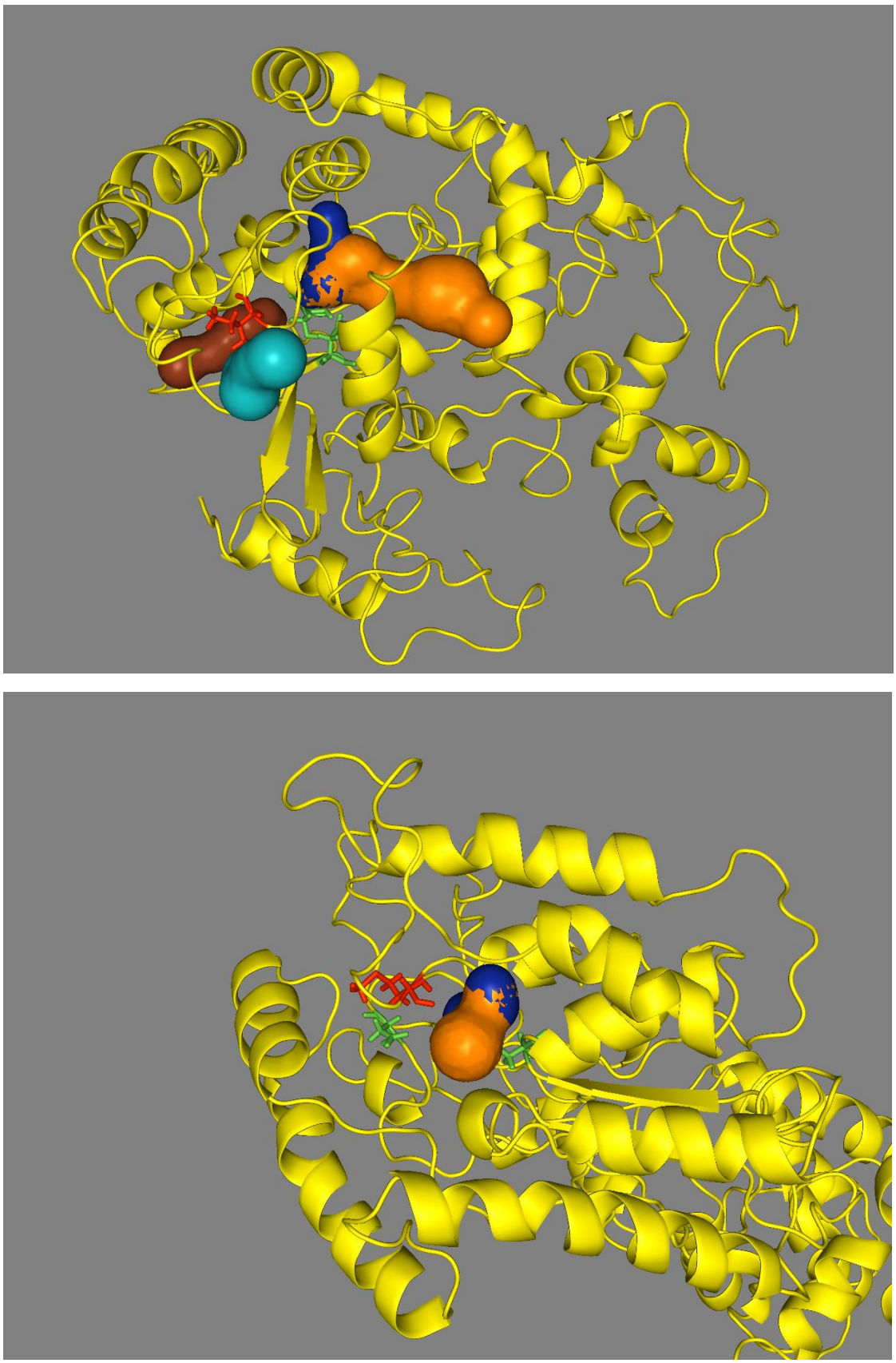


Figure 5.26 Tunnels to catalytic serine in final PNPLA3 structure

The catalytic residues S47 and D166 are highlighted in green, and residue 148 highlighted in red. Tunnel 1 is coloured blue, tunnel 2 orange, tunnel 3 brown and tunnel 4 turquoise.
Top panel; wild type protein.
Bottom panel; I148M variant.

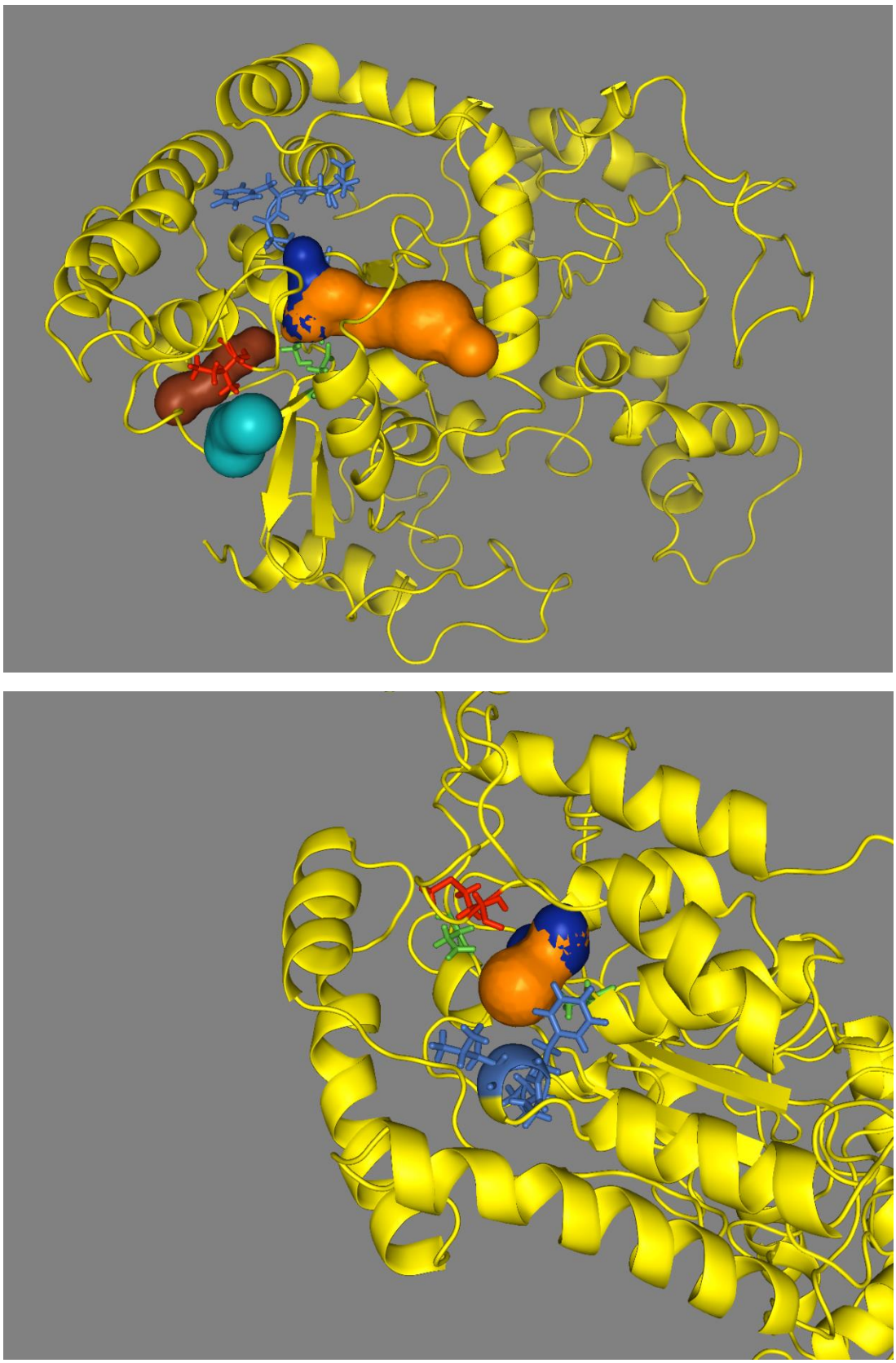


Figure 5.27 Tunnels in relation to oxyanion hole in final PNPLA3 structure

The catalytic residues S47 and D166 are highlighted in green, residue 148 highlighted in red and the oxyanion hole residues in blue. Tunnel 1 is coloured blue, tunnel 2 orange, tunnel 3 brown and tunnel 4 turquoise.

Top panel; wild type protein

Bottom panel; I148M variant.

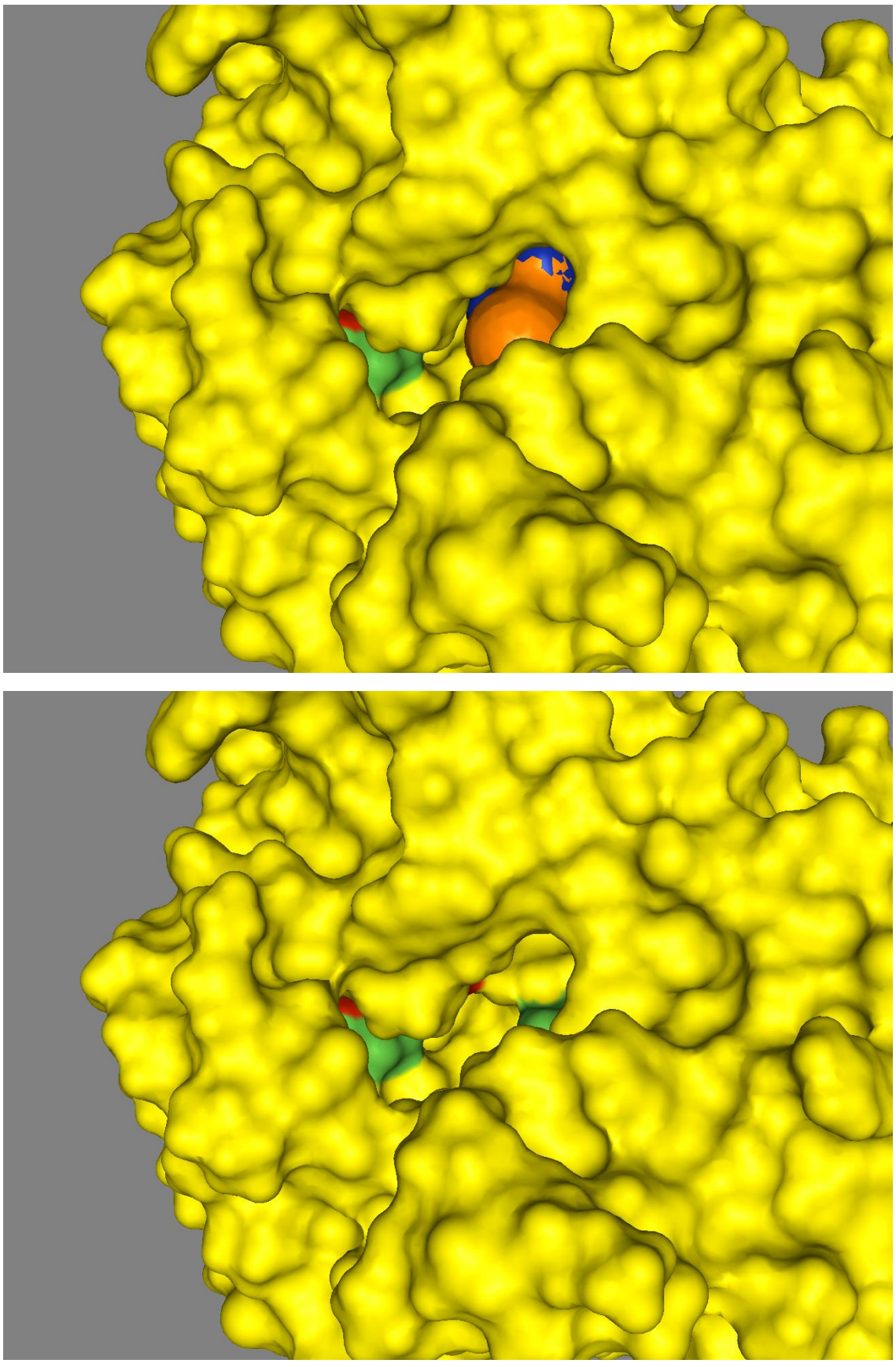


Figure 5.28 Active site cavity in I148M variant

The catalytic residues S47 and D166 are highlighted in green, and residue 148 highlighted in red. Both panels highlight large cavity in catalytic pocket. The tunnels are visible in the top panel.

Notably, while there were only two tunnels to the catalytic serine predicted in the I148M variant, the active site cavity is dramatically larger than that observed in the wildtype. In fact, the tunnels predicted account than less than half of the overall available cavity opening.

5.5.4.2 Tunnel properties

Wild type PNPLA3

Tunnel 1 is 10.25A in length, with a radius of above 1.5A along the tunnel length. The broadest portion of the tunnel is the middle and the radius spans to 3A at its widest part. This tunnel has the strongest negative net charge of -2 and is weakly hydrophilic the centre of the tunnel, with more hydrophobicity toward the ends. The tunnel is weakly polar, and the mutability dips to 0 at layer 14 because all tunnel facing atoms in that layer are backbone atoms (Tables 5.2 and 5.3; Figures 5.29 – 5.33).

Tunnel 2 is the longest tunnel at 23.65A in length. The radius fluctuates between 2 and 4A along the first half of the tunnel and 2 and 4A in the second half. The overall charge is negative and carries a -1 net charge, although the latter half of the tunnel is net positive. Additionally, the second half of the tunnel is significantly more polar and causes a polarity score of 16.13.

Tunnel 3 is 13.25A in length, with a radius which gradually increases from 1 to 3A along the length of the tunnel. There is a net charge of -1 and is weakly polar with a polarity score of 4.91. This tunnel is the most hydrophobic tunnel, with a hydrophobicity score of 0.43 and hydropathy score of 1.05. Along the length of the tunnel, it gradually becomes more polar, positively charged and hydrophilic in nature.

Tunnel 4 is 11.77A in length, with a radius that gradually increases from 1 to 5A along the length of the tunnel; making this the widest tunnel in the wild type protein. The tunnel is the only wild type tunnel with a positive net charge of +3, and the most polar tunnel scoring 18.67 in polarity. Both polarity and charge increase along the length of the protein as well as becoming more hydrophilic in nature.

I148M variant PNPLA3

Tunnel 1 is the shortest tunnel at 5.79A in length. IT has a freeradius above 2.3 along the entire tunnel with a rapid increase toward the end. It carries a net positive charge of +1 and is slightly hydrophilic in nature while being highly polar with a polarity score of 17.70. Interestingly is more polar toward the beginning of the tunnel rather than the outside.

Tunnel 2 is 7.09A in length, and while generally having a radius of above 2A has a narrow bottleneck toward the end of the tunnel. It is positively charged with a charge of +1, is very slightly hydrophobic and highly polar with a polarity score of 15.08. Like tunnel 1, the polarity decreases toward the exit of the tunnel.

Table 5.2 Average tunnel properties to catalytic serine in both PNPLA3 variants

Tunnel	Length	Charge	Hydrophobicity	Hydropathy	Polarity	Mutability
Wild type tunnel 1	10.25	-2	-0.15	-0.64	6.65	81
Wild type tunnel 2	23.65	-1	0.16	-1.14	16.13	77
Wild type tunnel 3	13.25	-1	0.43	1.05	4.91	81
Wild type tunnel 4	11.77	3	0.28	-0.21	18.67	78
I148M variant tunnel 1	5.79	1	-0.03	-0.29	17.70	88
I148M variant tunnel 2	7.09	1	-0.14	-0.47	15.08	76

Table 5.3 The residue flow through tunnels to catalytic serine in both PNPLA3 variants

Tunnel	Residue flow
Wild type tunnel 1	SER 47, ALA 48, PRO 149, PRO 149 Backbone, ASP 166, TYR 191, TYR 151, GLU 14 Backbone, CYS 15 Backbone, SER 152 Backbone, GLU 14, SER 152
Wild type tunnel 2	SER 47, ALA 48, PRO 149, PRO 149 Backbone, ASP 166, TYR 191, TYR 151, GLU 14 Backbone, CYS 15 Backbone, SER 152 Backbone, GLU 14, SER 152, ALA 13, VAL 169, ARG 213, GLU 189 Backbone, LYS 196, TYR 188, GLU 190, LEU 214, LEU 210
Wild type tunnel 3	SER 47, LEU 118, SER 145, CYS 144, ALA 141 Backbone, ALA 141, TYR 164, ASP 140
Wild type tunnel 4	ILE 148, PHE 150, TYR 164 Backbone, ARG 163, VAL 162 Backbone, ARG 160 Backbone, ARG 160, LYS 198
I148M variant tunnel 1	LEU 51, ARG 74, ALA 48, SER 47
I148M variant tunnel 2	LEU 51, ARG 74, SER 47, ALA 48, CYS 146, SER 145 Backbone, MET 148 Backbone, PRO 149

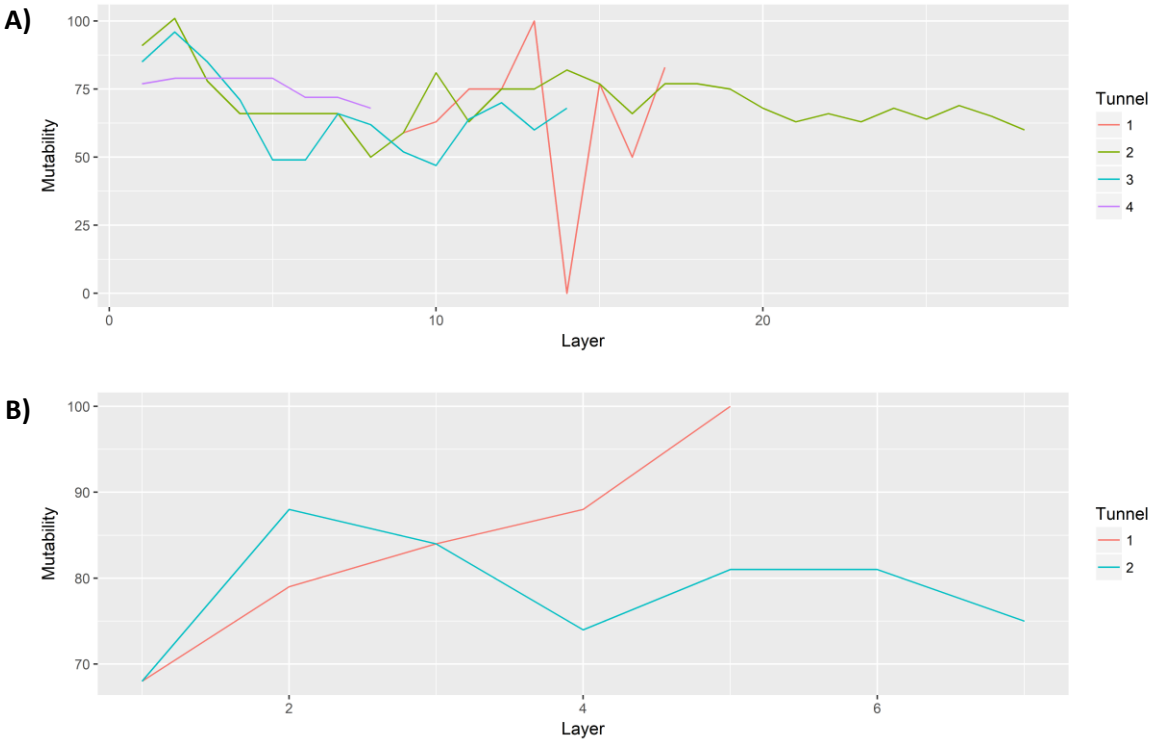


Figure 5.29 Mutability of the active site tunnels of PNPLA3 variants

A) Mutability of wild type PNPLA3.
B) Mutability of I148M variant.

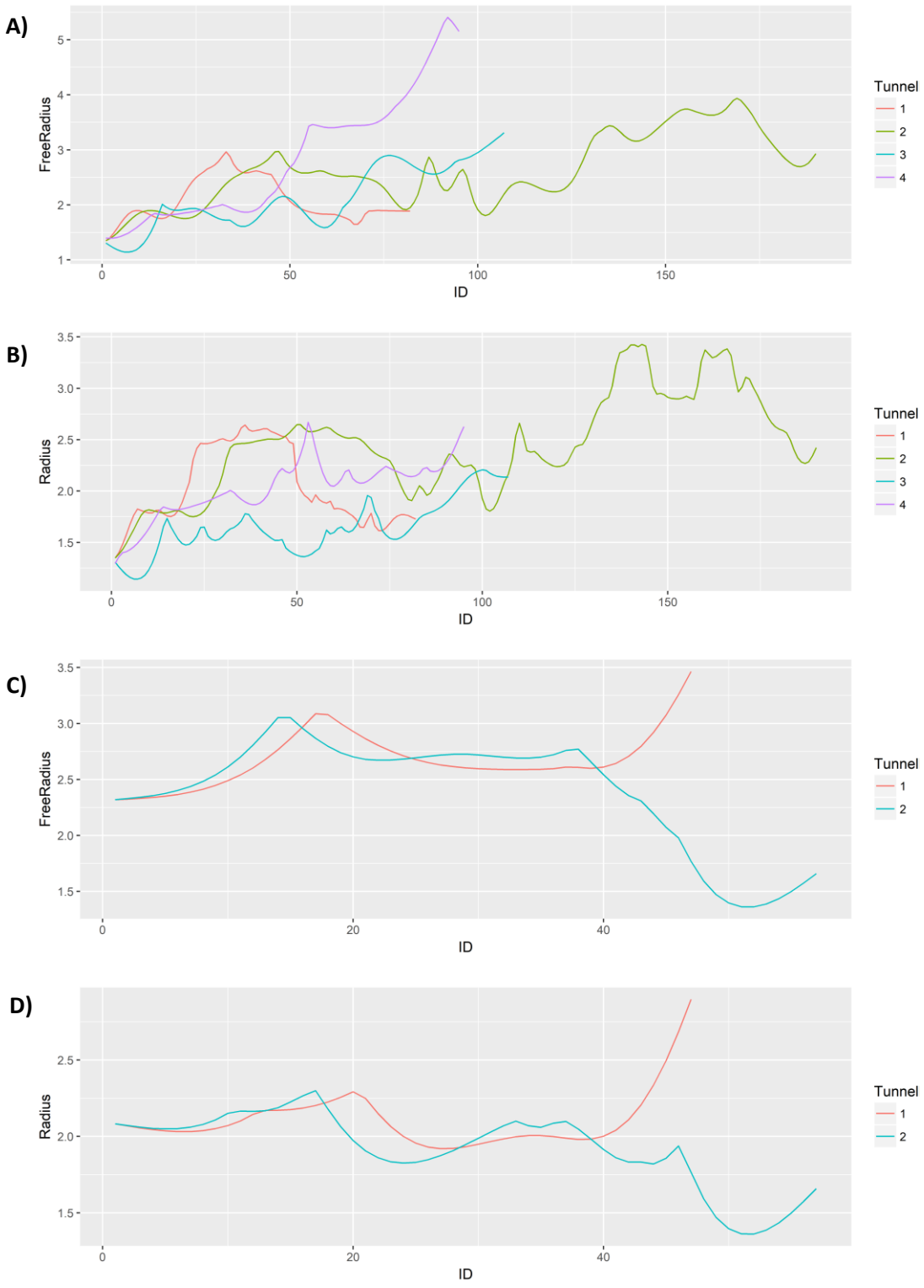


Figure 5.30 Radius of the active site tunnels of PNPLA3 variants

A) Free radius of wild type PNPLA3.
B) radius of wild type PNPLA3.
C) Free radius of I148M variant.
D) radius of I148M variant.

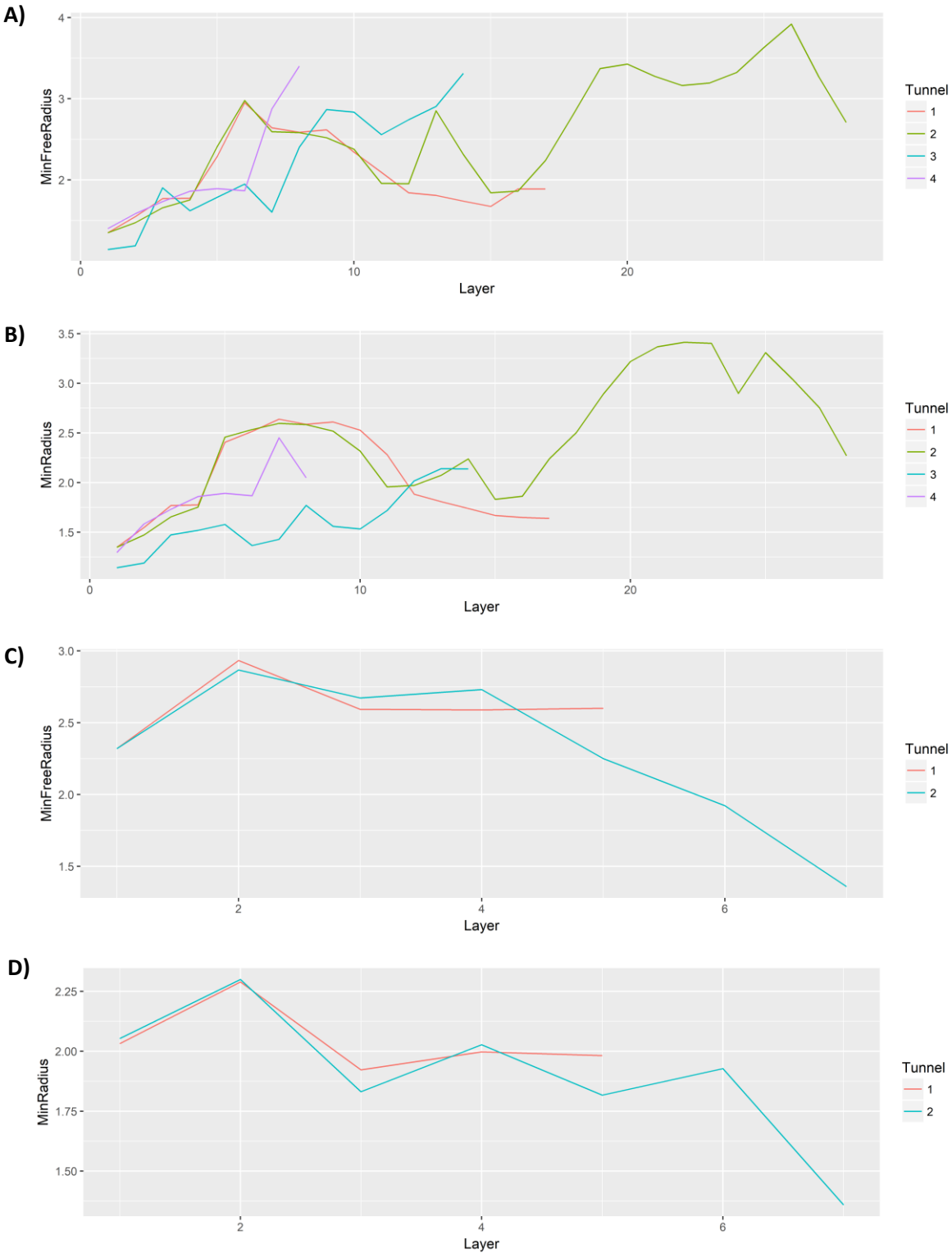


Figure 5.31 Radius of the active site tunnels of PNPLA3 variants by layer

A) Free radius of wild type PNPLA3.
B) radius of wild type PNPLA3.
C) Free radius of I148M variant.
D) radius of I148M variant.

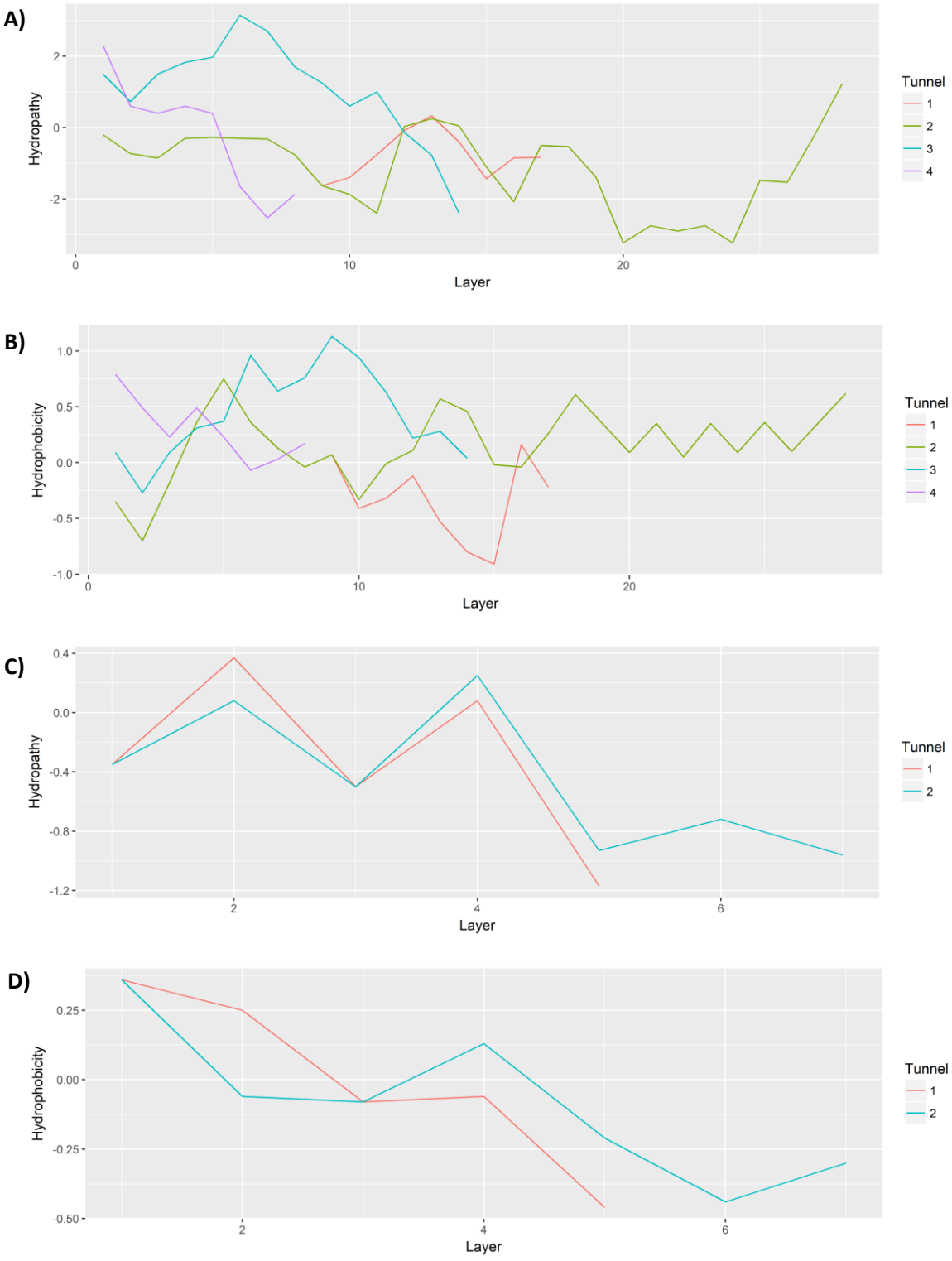


Figure 5.32 Hydropathy and hydrophobicity of active site tunnels of PNPLA3 variants

A) hydropathy of wild type PNPLA3.
B) Hydrophobicity of wild type PNPLA3.
C) Hydropathy of I148M variant.
D) hydrophobicity of I148M variant.

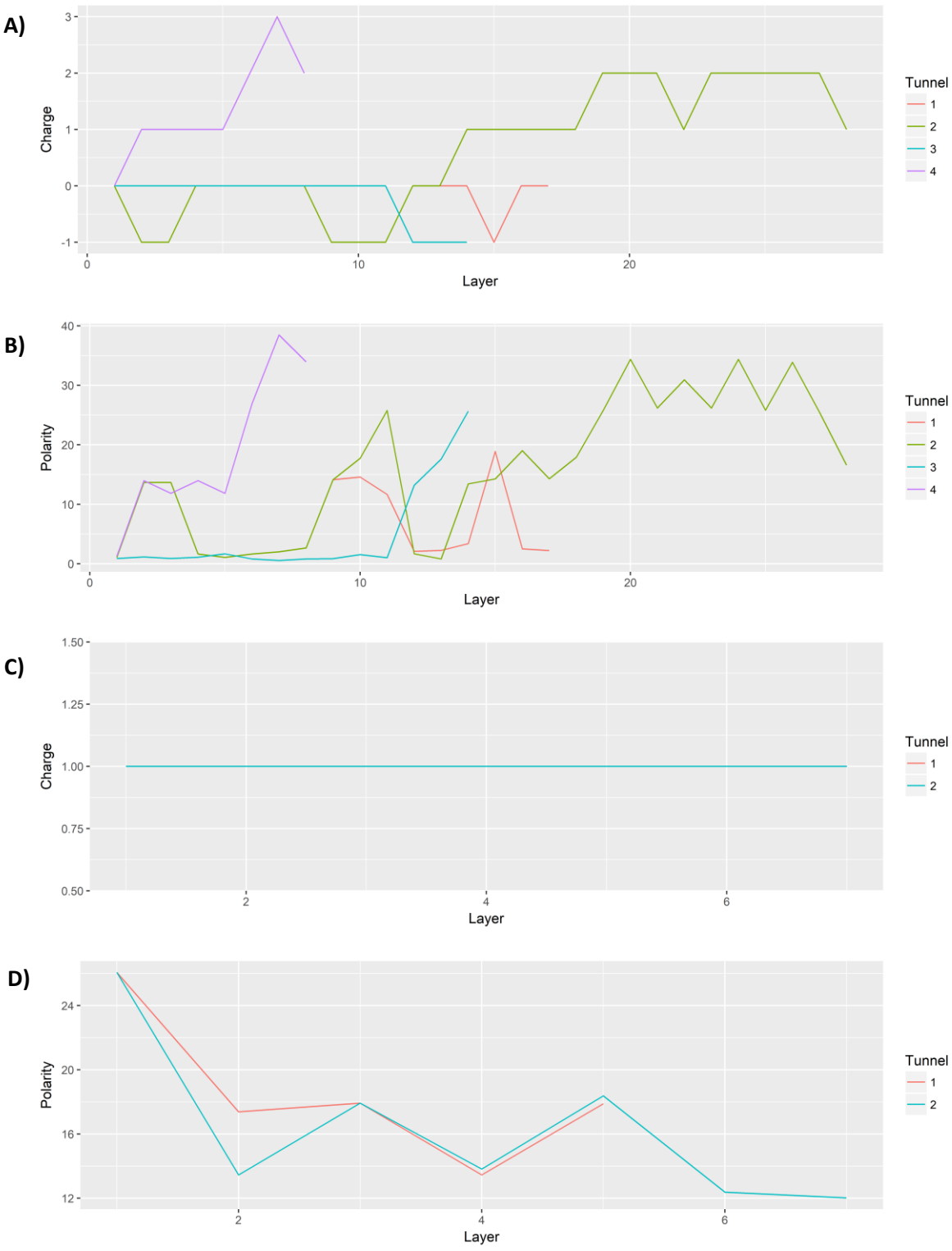


Figure 5.33 Charge and polarity of the active site tunnels of PNPLA3 variants

A) Charge of wild type PNPLA3.
B) Polarity of wild type PNPLA3.
C) Charge of I148M variant.
D) Polarity of I148M variant.

5.5.5 Docking

All ligands were successfully docked into both of the PNPLA3 variants without error. The primary docking mode for all ligands is shown in (Figures 5.34 – 5.36).

5.5.5.1 Wild type binding sites

The ligands all docked into three binding sites in the wild type protein. These three sites were located on different regions around the patatin domain.

The most common binding site (Binding site A) is positioned in the active site cavity of tunnel 1 and 2. This site is confined by helix G and H similarly to tunnels 1 and 2, but also supported by and extending out all the way to helix C and D (Figure 5.34).

The second binding site (Binding site B) is positioned in between left and right lobes of protein, below tunnel 3. It embeds deep into the protein, up to strands 1 and 2, helix N and O from C-terminal domain, as well as the loop after D166, and loop after strand 4. This is a deep cleft and is dependent on the position of C-terminal domain.

The third binding site (binding site C) is positioned just outside of the loop connecting helix G and H, and so below tunnel 2. Like binding site B, it is formed between the right lobe helices, and the C-terminal domain based on the sequence connecting the two domains, but on the opposite face of the protein. It is located in another deep pocket which supports binding (Figure 5.35).

5.5.5.2 I148M variant binding sites

In the I148M variant two binding sites were observed in the docking results. The first (binding site A) was located in the large cavity opening of the active site. Much like the tunnels, the active site cavity in the structure is framed by the Loop between helix C and D on the right, helix A at the bottom, helix G and preceding loop on the left, and the loop containing Met 148 on the top (Figure 5.36).

There is a small secondary binding site (binding site B), which is located slightly above binding site A. This site is framed by helix H, the loop region between helices C and D and the loop region between M148 and D166.

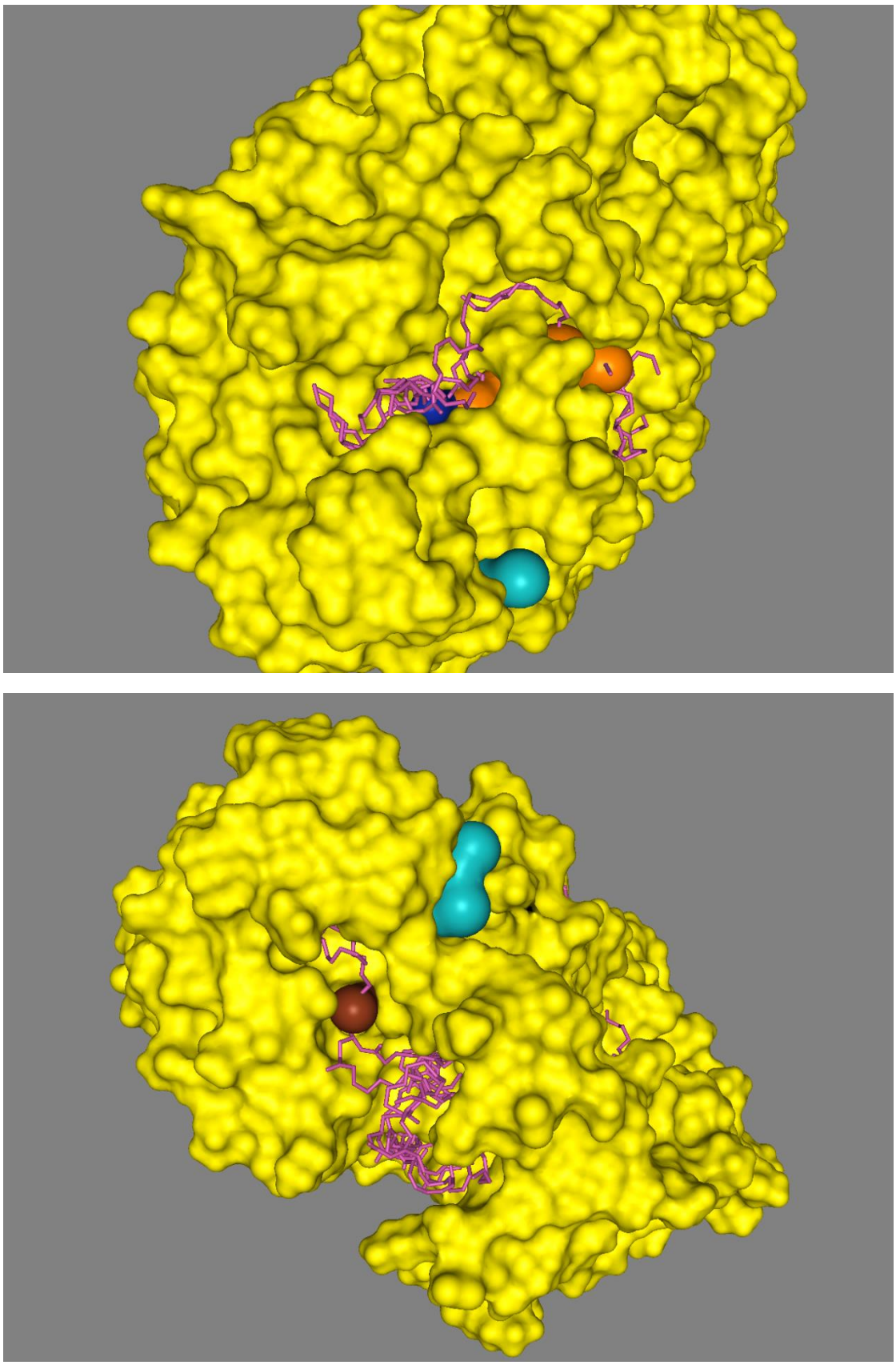


Figure 5.34 Binding sites a and B in wild type PNPLA3

Surface of PNPLA3 in yellow, tunnel 1, 2, 3 and 4 in blue, orange, brown and turquoise respectively. All docked ligands highlighted in magenta.
Top panel; Binding site cavity A
Bottom panel; Binding site cavity B.

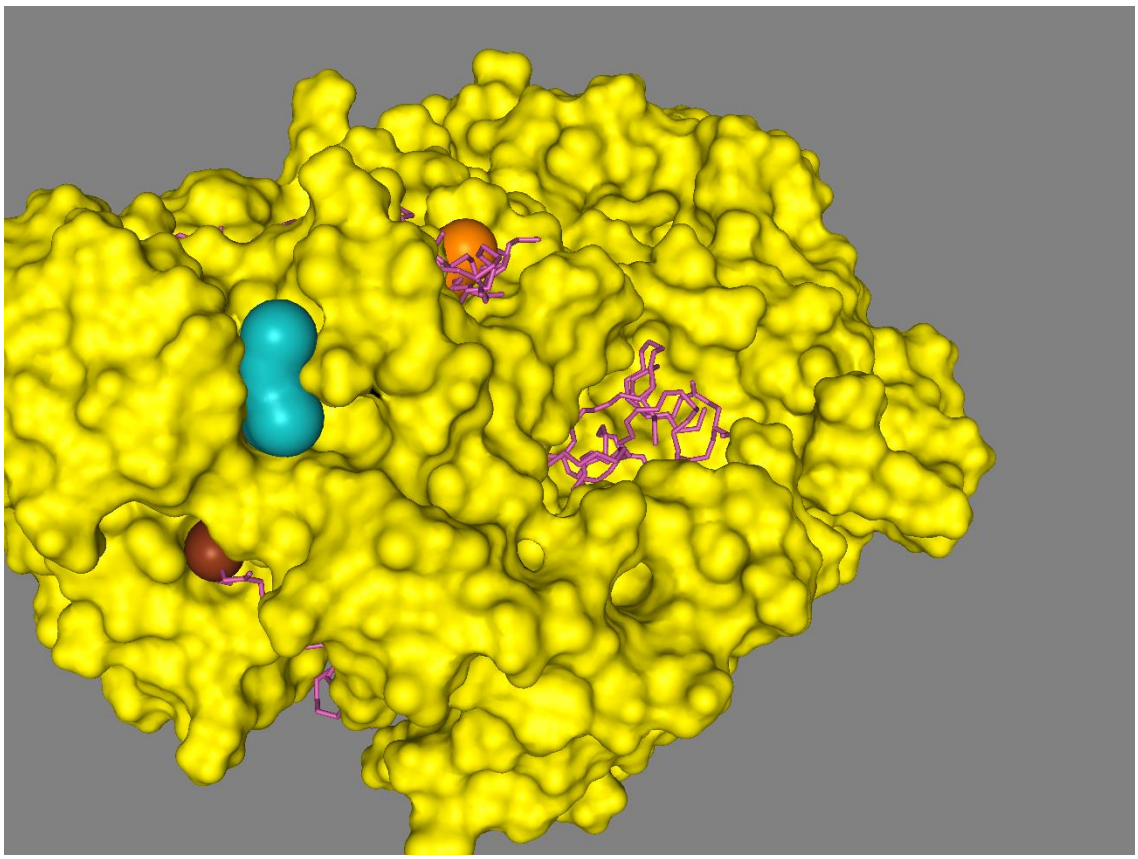


Figure 5.35 Binding sites C in wild type PNPLA3

Surface of PNPLA3 in yellow, tunnel 1, 2, 3 and 4 in blue, orange, brown and turquoise respectively. All docked ligands highlighted in magenta.

5.5.5.3 Wild type docking

The WT clearly has much more complex binding patterns than the variant and seems to facilitate more interaction with the lipids in other surfaces of the protein (Figures 5.37 – 5.45).

The main binding site for most ligands is binding site A, where the majority of the ligands had their best binding mode. Within binding site A, palmitic acid, retinoic acid, retinol and trimyristin all had the best binding sites occupying space predominantly in binding tunnel 2, and the top 3 binding modes were all found in this tunnel. Tunnel 2 seemed to perform particularly well in docking of short chain fatty acids.

Tripalmitin, 1,2-diolein, 1,2-dipalmitin, 1,3-diolein, and 1,3-dilinolein, while also located in binding site A, but docked into both tunnels 1 and 2 through the protein. In this way, tunnels 1 and 2 seem to act as a protein channel, allowing ligands to dock through the protein. This allows the larger fatty acids to present to the active site and would appear to be the most important binding mode for putative catalysis.

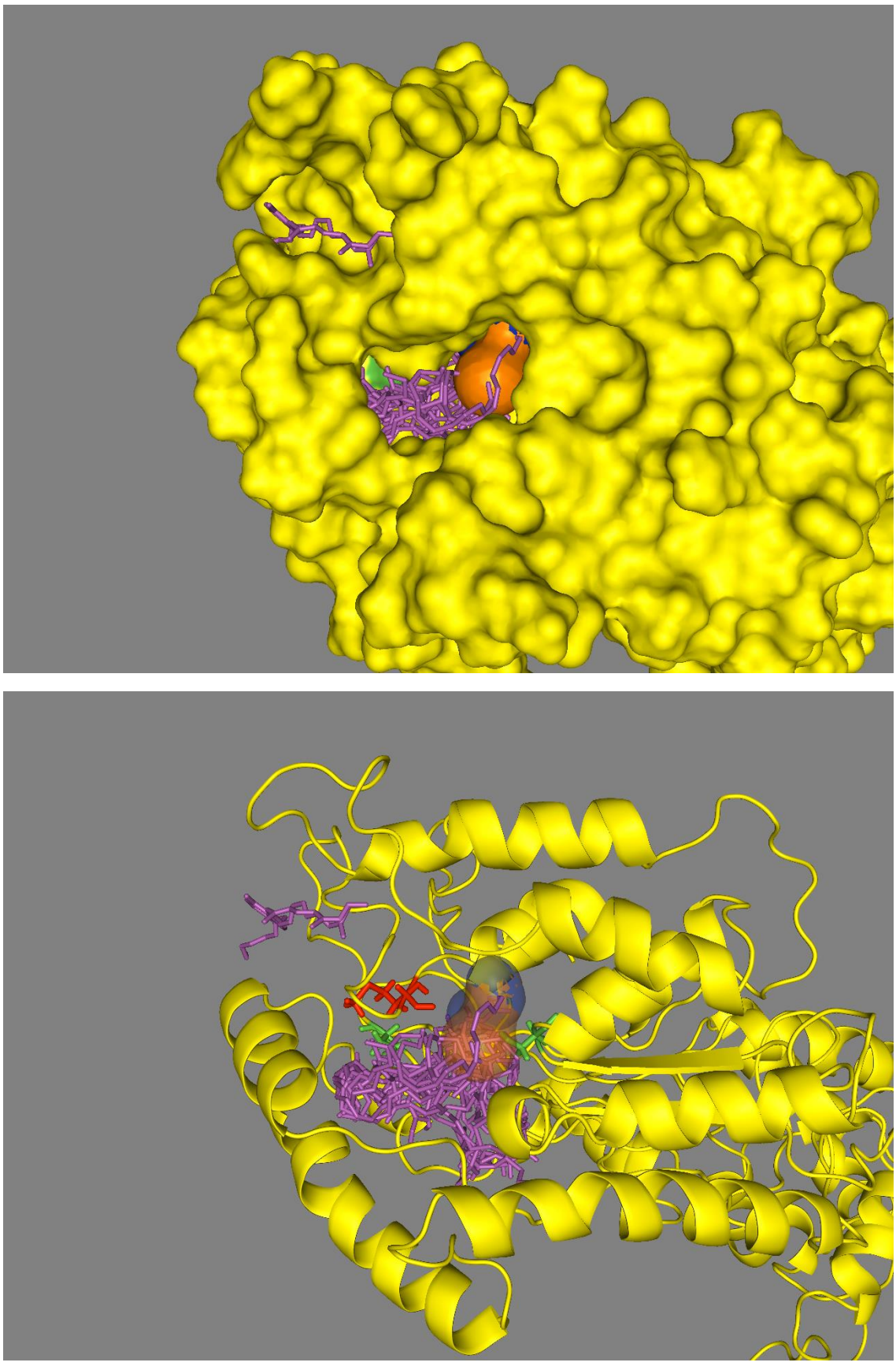


Figure 5.36 Primary binding site in the 148M variant

Tunnel 1 and 2 highlighted in blue and orange respectively. All docked ligands highlighted in magenta.
Top panel; Surface map of PNPLA3
Bottom panel; Cartoon representation of PNPLA3.

The top three binding modes for all of these ligands are in this binding site except tripalmitin, which moves to cavity B. Oleic acid also binds in cavity A, however the best binding mode is on the outer edge of tunnel 1.

The best binding mode for triolein and linoleic acid are located in binding Site B, however the second binding mode of linoleic acid is located in tunnel 2, within binding site A, like the other short chain fatty acids. The secondary binding mode of triolein is in binding site C. In its third binding mode triolein does appear to be entering tunnel 3, but had no other top results suggesting binding toward the active site.

Triarachidonin and trilinolein both have their best binding modes in binding site C, and all four top binding modes for these ligands lie in this site. The fifth binding mode for trilinolein fits very well in binding site A, going through tunnels 1 and 2, as the other TAGs. Triarachidonin however fits in cavity B in its fifth binding mode and has no top binding modes in binding site A at all.

5.5.5.4 I148M variant binding sites

When docking into the I148M variant, nearly all ligands docked with best binding modes in binding site A, which formed the large open active site cavity. Only palmitic acid and retinol preferentially bound to the small surface cavity at binding site B (Figures 5.46 – 5.52).

The second binding mode of retinol however, and the fifth of palmitic acid were also in the active site cavity. Throughout the binding modes other ligands were also docked to binding site B, suggesting although a small site, it may be a common site of interaction on entry to site A.

All of the ligands sit differently within the active site cavity, with no obvious pattern of binding. The pocket appears to provide a large hydrophobic space to facilitate binding of lipids but has no clear potential for catalysis based on the disruption of the active site.

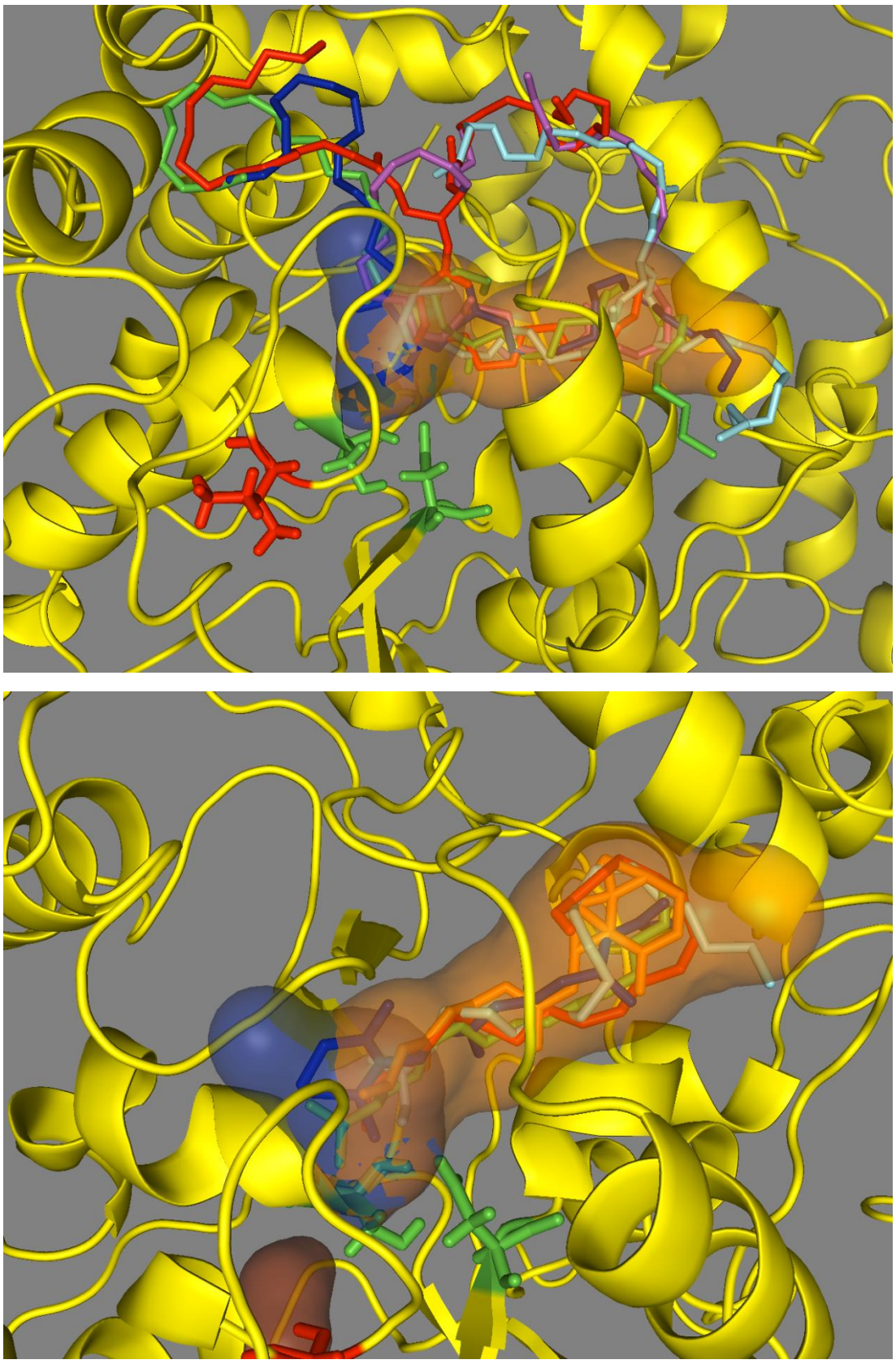


Figure 5.37 Primary binding site in the PNPLA3 wild type

PNPLA3 shown in yellow, tunnels 1 and 2 highlighted in translucent blue and orange respectively.
Top panel; docked TAGs and DAGs binding mode 1. tripalmitin in red, 1,2-diolein in blue, 1,2-dipalmitin in green, 1,3-dilinolein in magenta and trimyristin in cyan.
Bottom panel; Docked short chain fatty acids in binding mode 3. Oleic acid in red, palmitic acid in green, retinoic acid in blue, retinol in orange and linoleic acid in cyan.

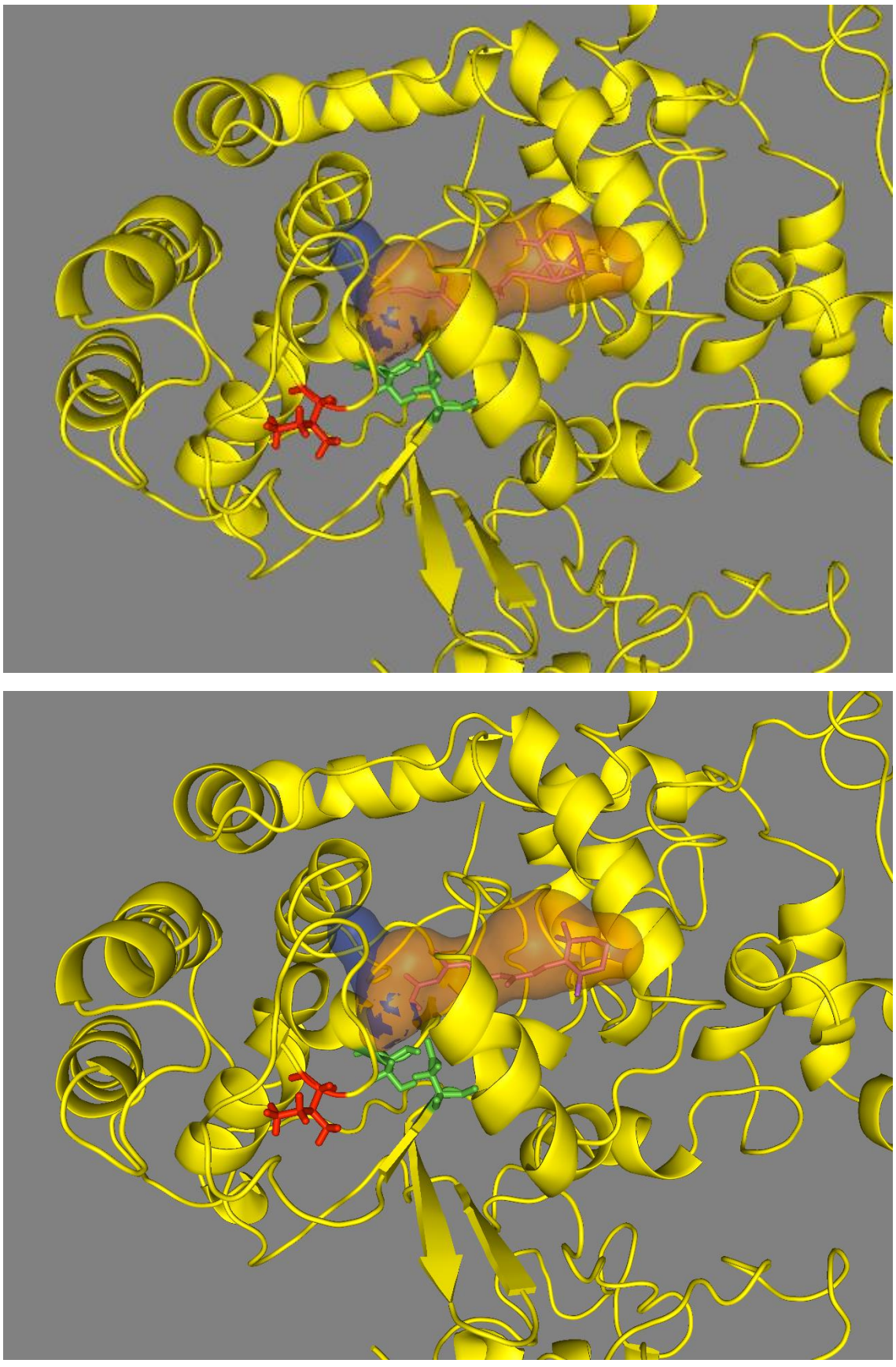


Figure 5.38 Primary docking mode of retinol and retinoic acid to wild type PNPLA3
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; Retinol.
Bottom panel; retinoic acid.

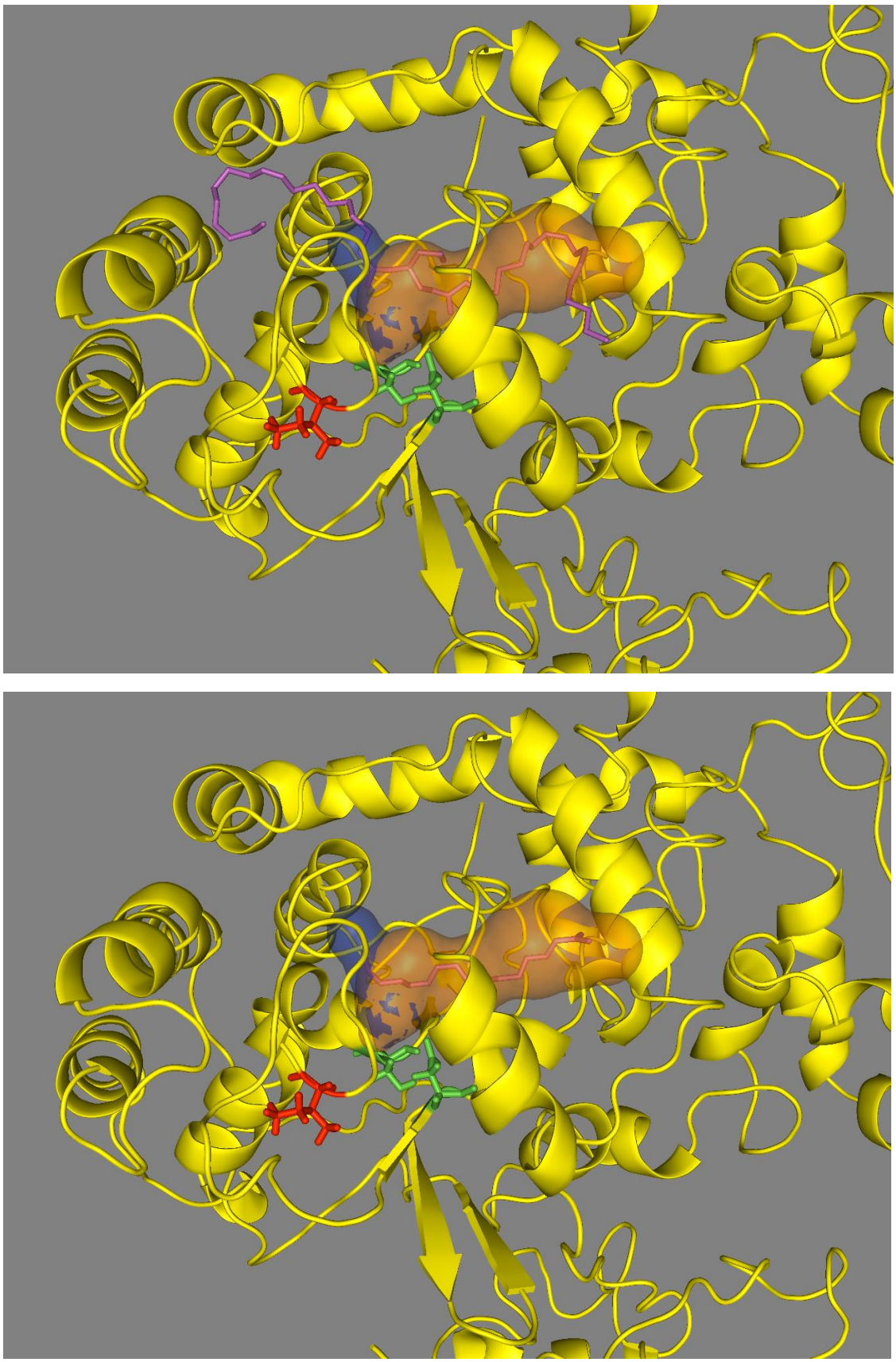


Figure 5.39 Primary docking mode of 1,2-diolein and palmitic acid to wild type PNPLA3

Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; 1,2-diolein.
Bottom panel; palmitic acid.

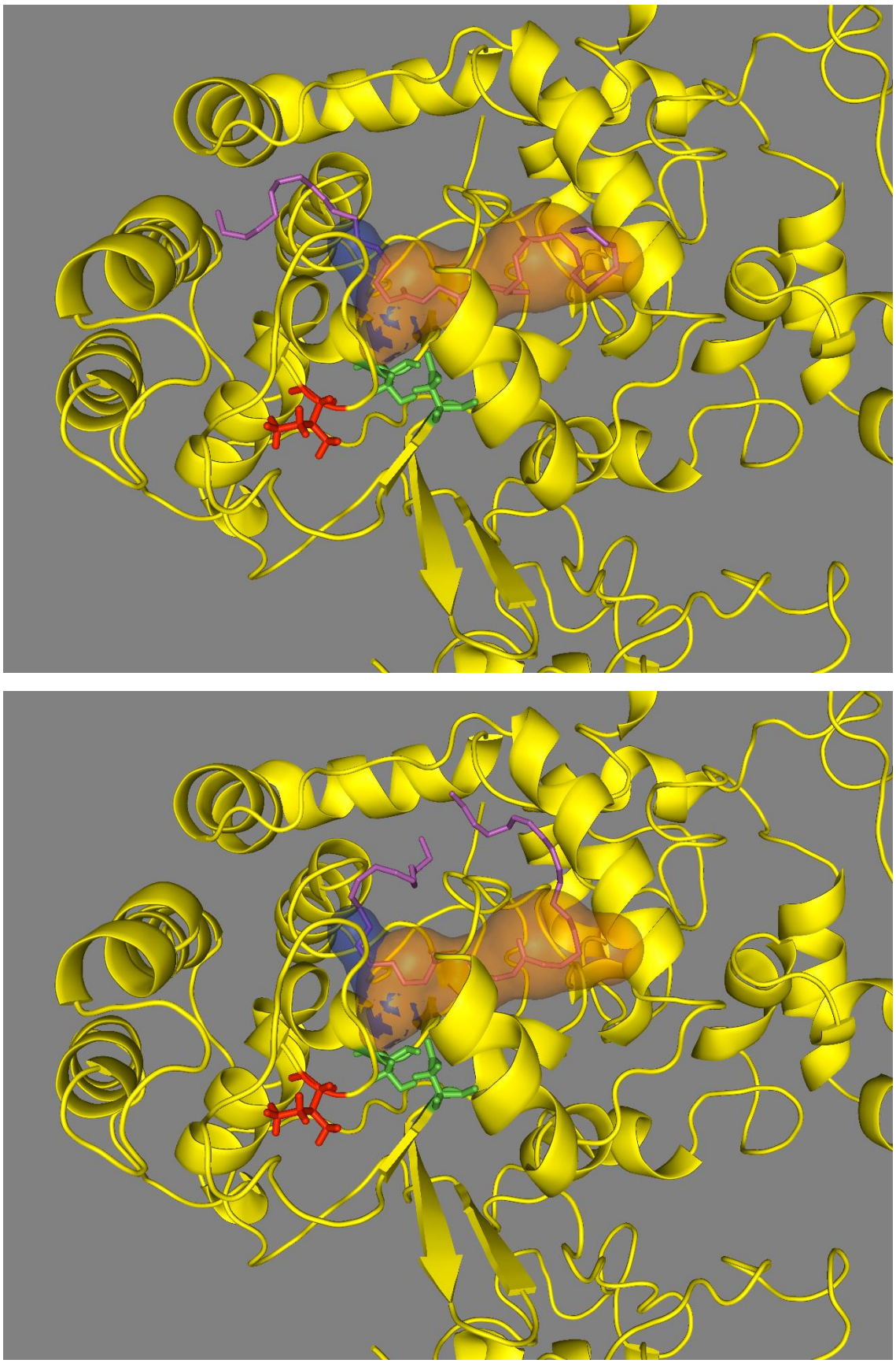


Figure 5.40 Primary docking mode of 1,3-diolein and 1,3-dilinolein to wild type PNPLA3
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; 1,3-diolein.
Bottom panel; 1,3-dilinolein.

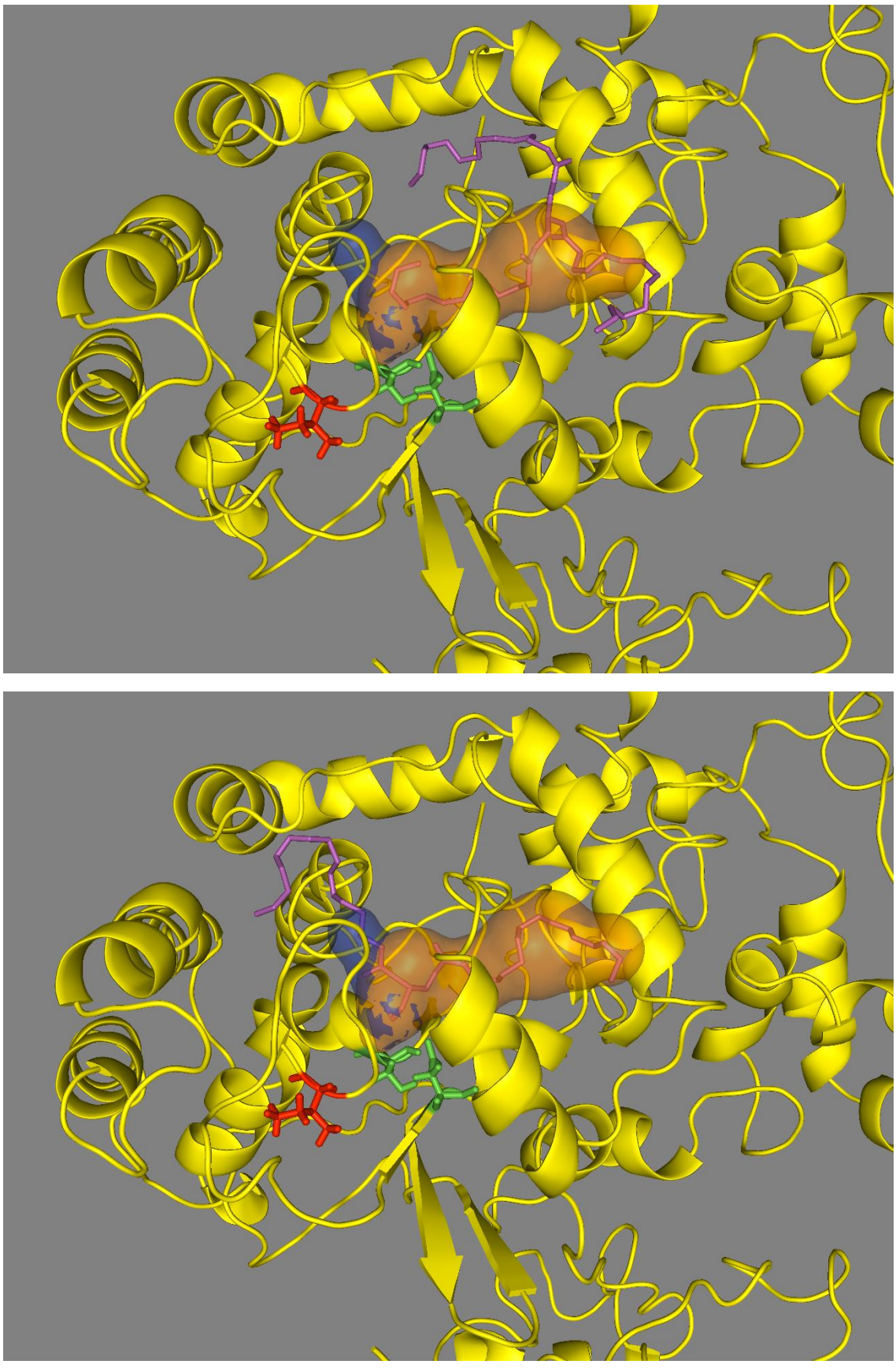


Figure 5.41 Primary docking mode of trimyristin and 1,2-dipalmitin to wild type PNPLA3
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; trimyristin.
Bottom panel; 1,2-dipalmitin.

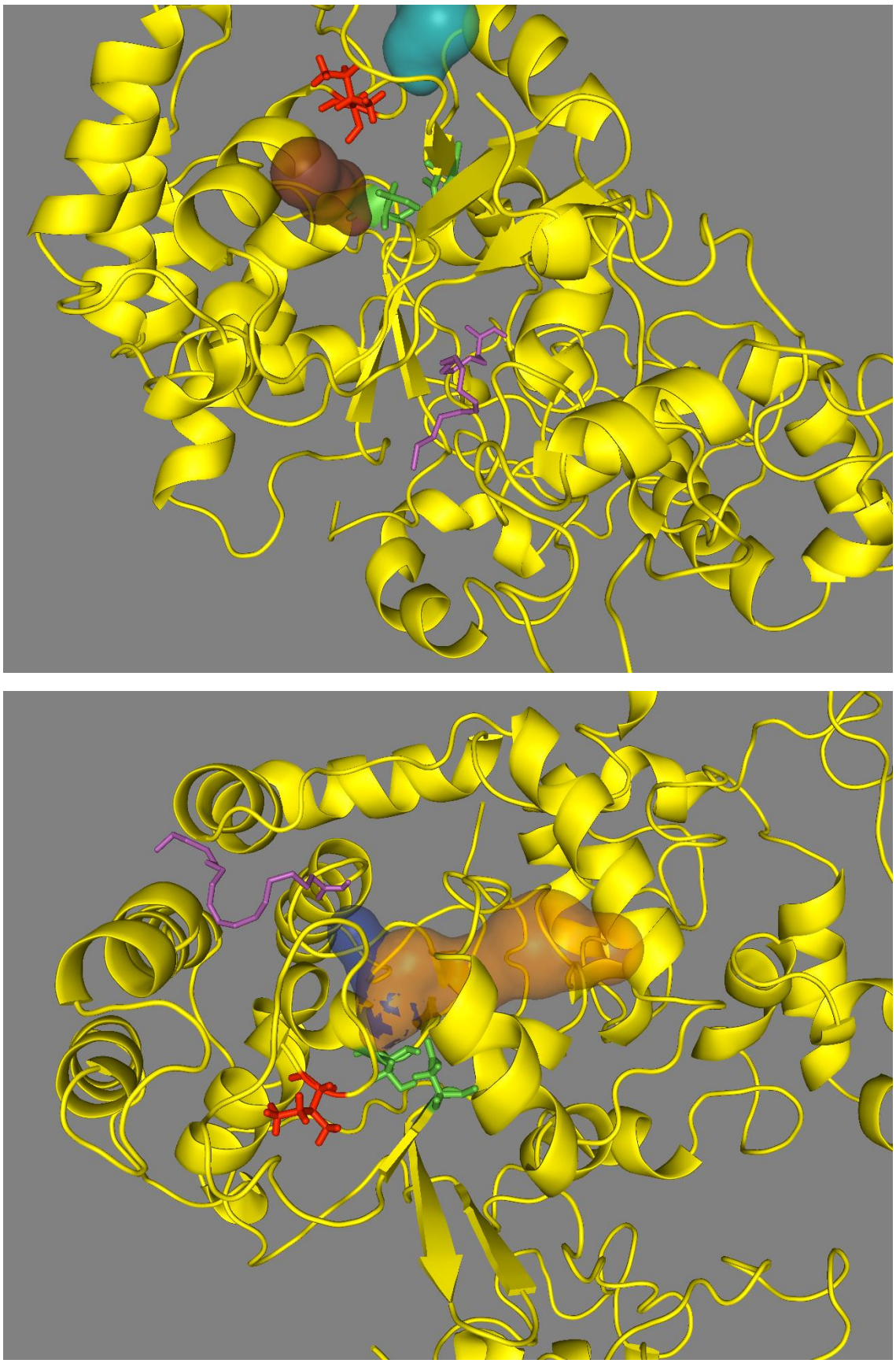


Figure 5.42 Primary docking mode of linoleic acid and oleic acid to wild type PNPLA3
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1, 2, 3 and 4 in translucent blue, orange, brown and turquoise respectively. The docked ligands highlighted in magenta.
Top panel; linoleic acid.
Bottom panel; oleic acid.

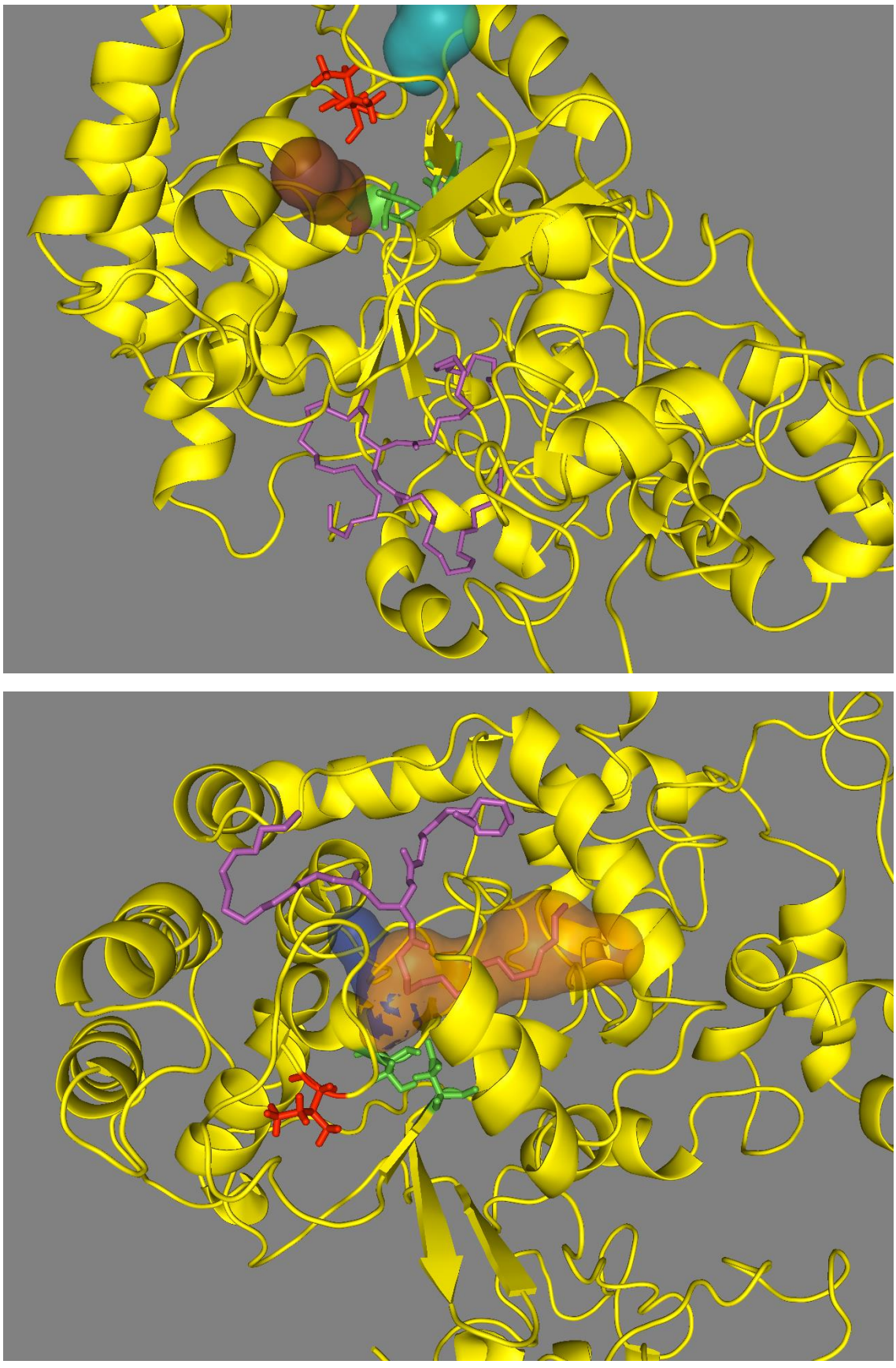


Figure 5.43 Primary docking mode of triolein and tripalmitin to wild type PNPLA3
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1, 2, 3 and 4 in translucent blue, orange, brown and turquoise respectively. The docked ligands highlighted in magenta.
Top panel; triolein.
Bottom panel; tripalmitin.

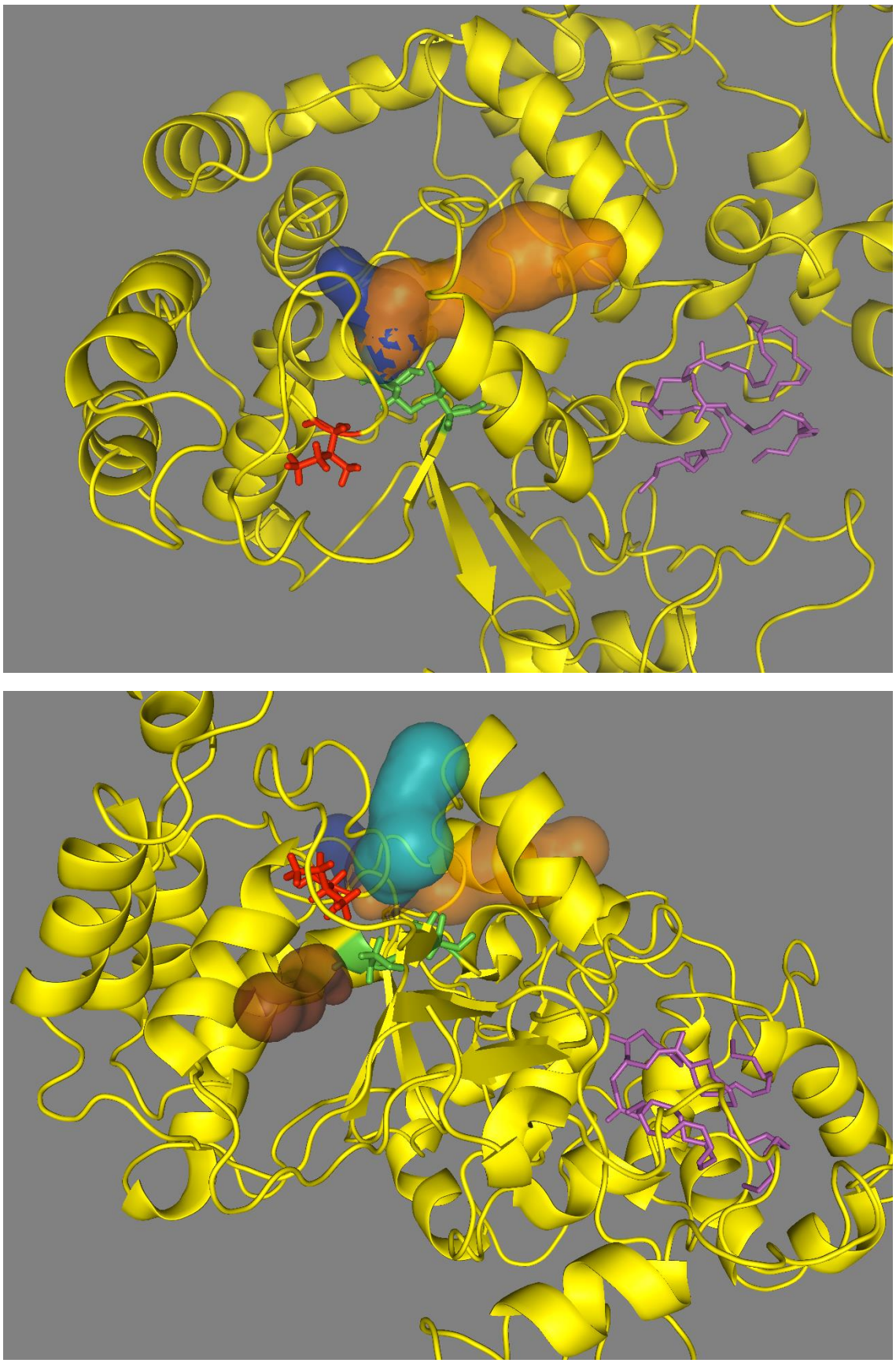


Figure 5.44 Primary docking mode of triarachidonin and trilinolein to wild type PNPLA3
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1, 2, 3 and 4 in translucent blue, orange, brown and turquoise respectively. The docked ligands highlighted in magenta.
Top panel; triarachidonin.
Bottom panel; trilinolein.

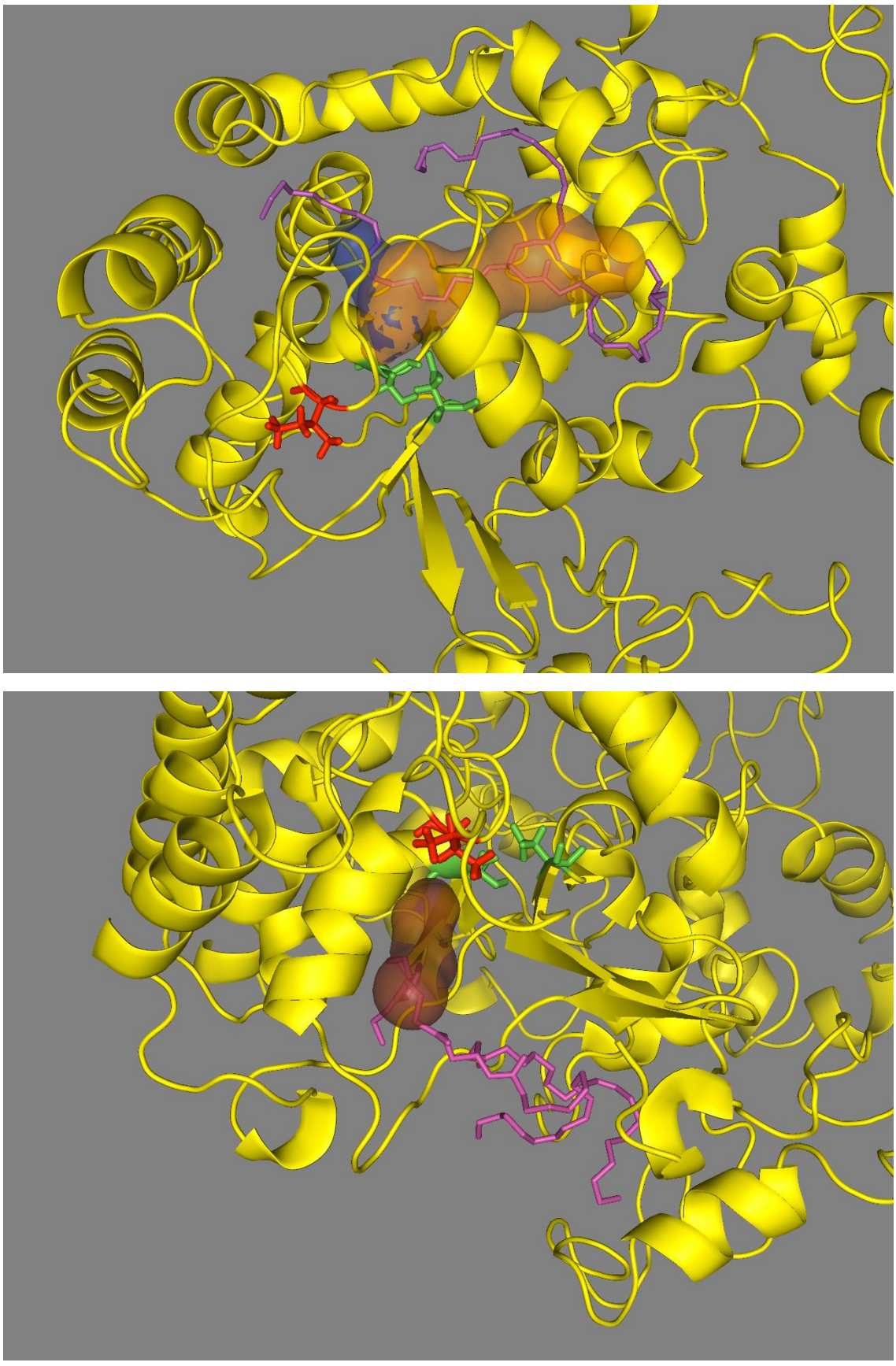


Figure 5.45 Binding mode 3 of trilinolein and triolein to wild type PNPLA3

Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1, 2, 3 and 4 in translucent blue, orange, brown and turquoise respectively. The docked ligands highlighted in magenta.

Top panel; triarachidonin.

Bottom panel; trilinolein.

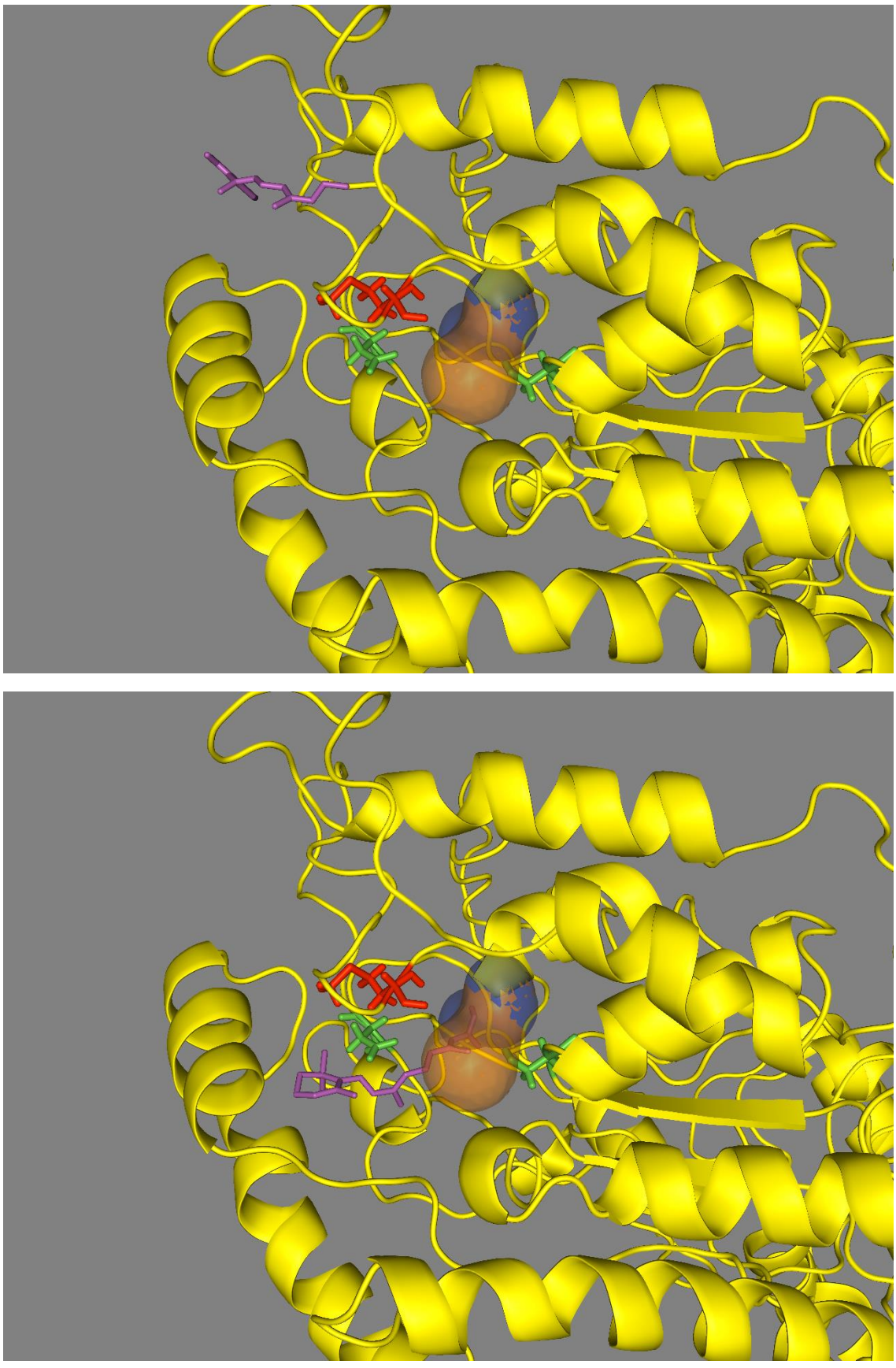


Figure 5.46 Primary docking mode of retinol and retinoic acid to I148M variant

Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.

Top panel; Retinol.

Bottom panel; retinoic acid.

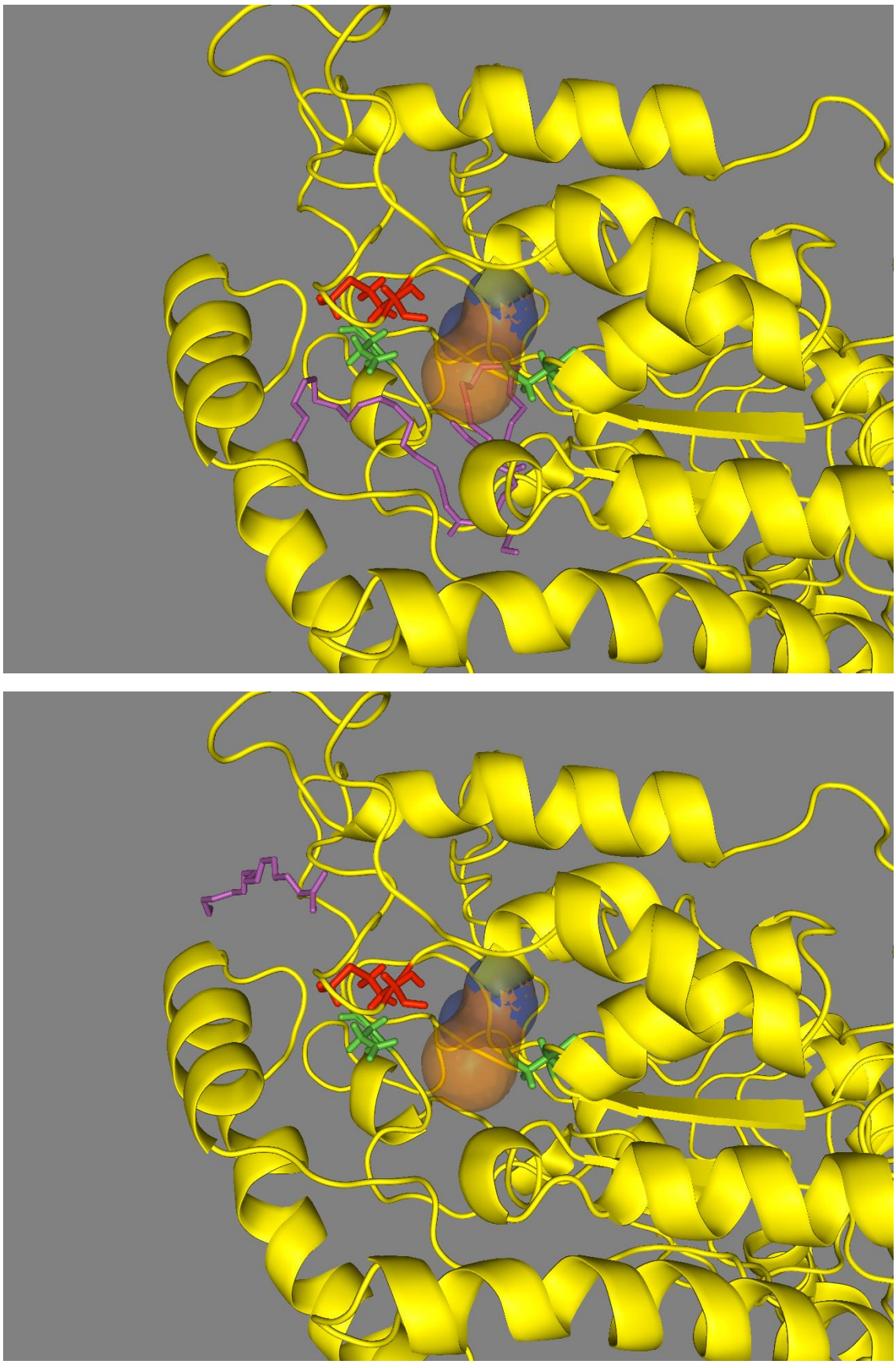


Figure 5.47 Primary docking mode of 1,2-diolein and palmitic acid to I148M variant
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; 1,2-diolein.
Bottom panel; palmitic acid.

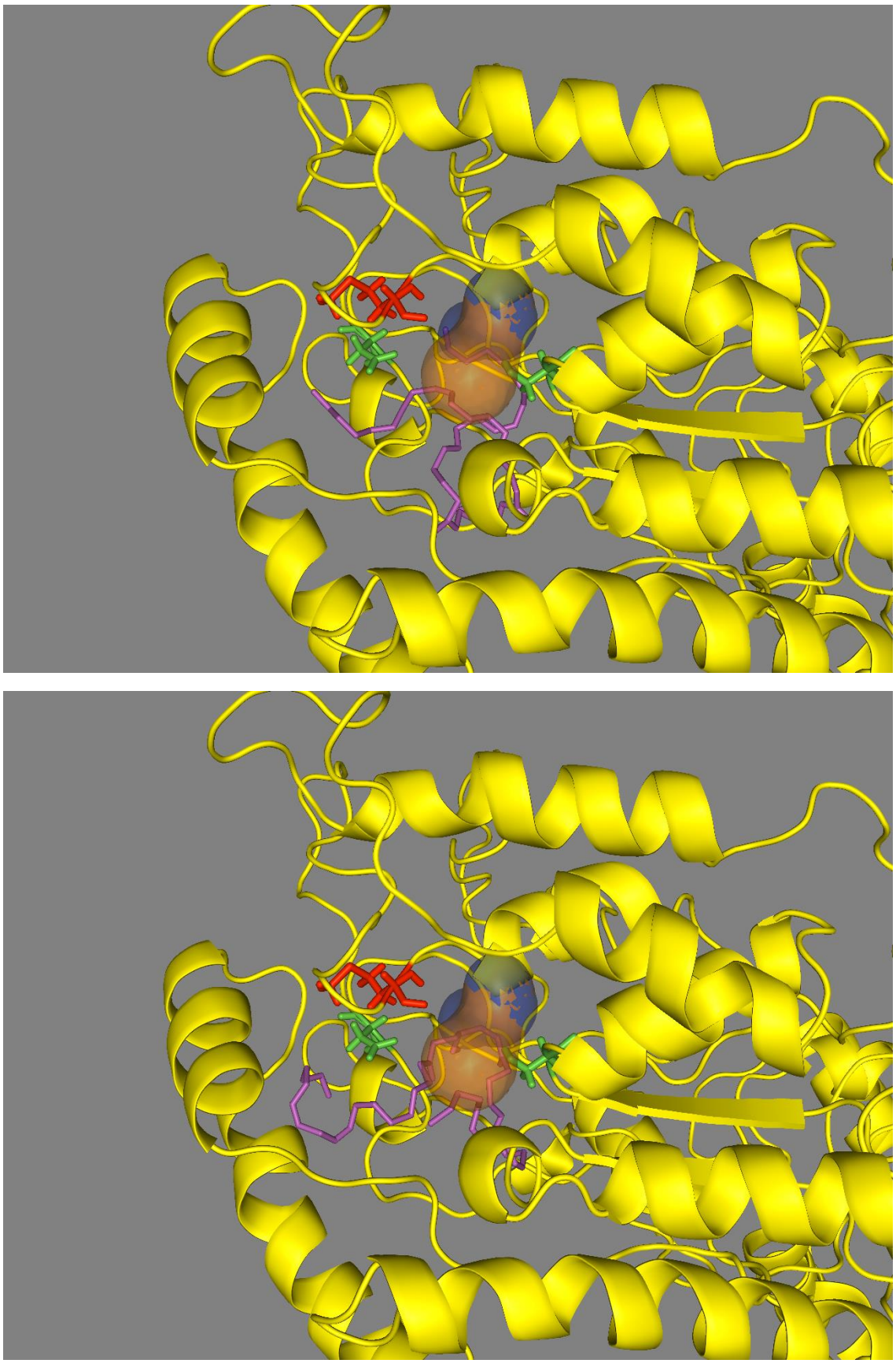


Figure 5.48 Primary docking mode of 1,3-diolein and 1,3-dilinolein to I148M variant
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; 1,3-diolein.
Bottom panel; 1,3-dilinolein.

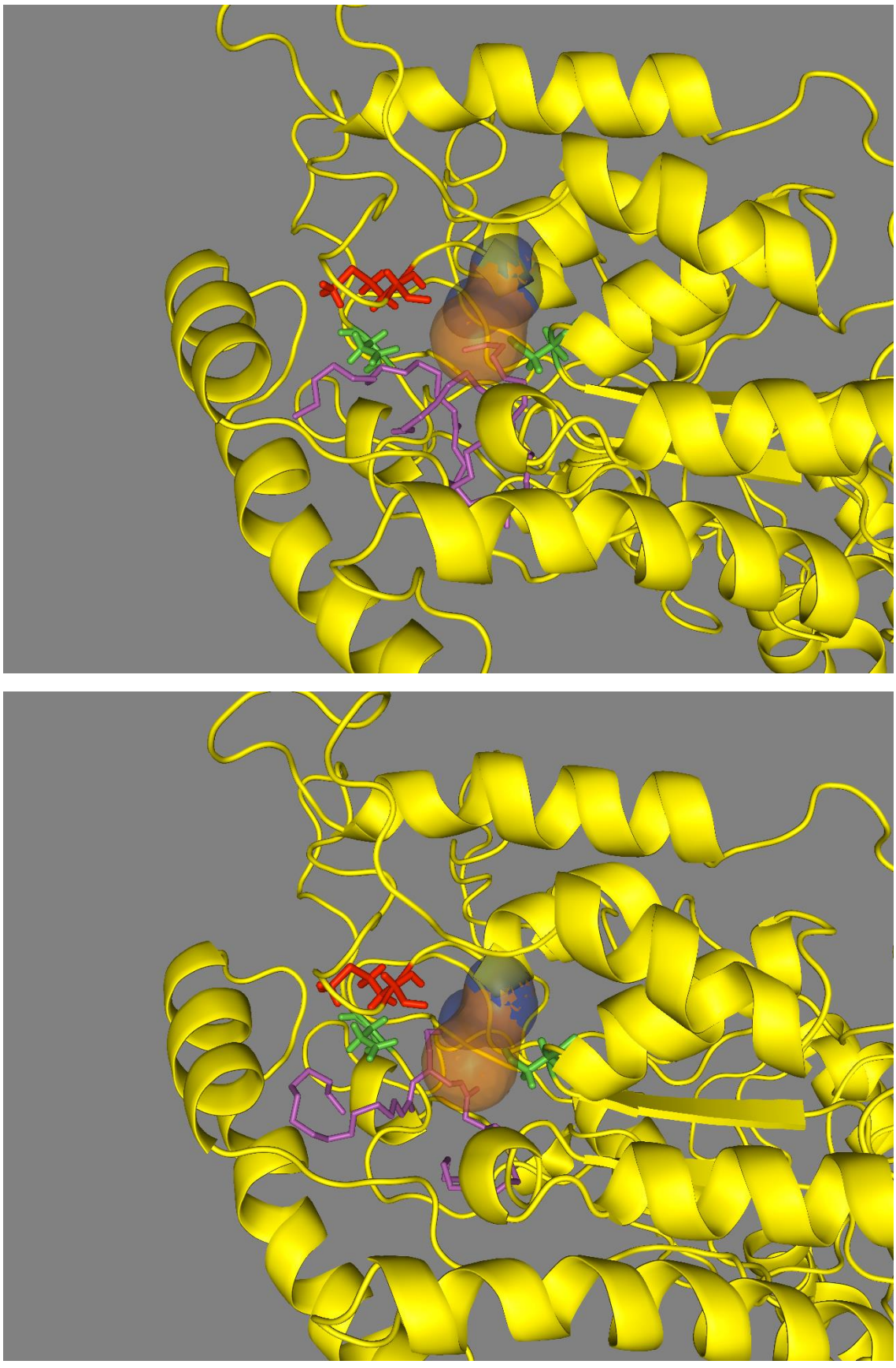


Figure 5.49 Primary docking mode of trimyristin and 1,2-dipalmitin to I148M variant
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; trimyristin.
Bottom panel; 1,2-dipalmitin.

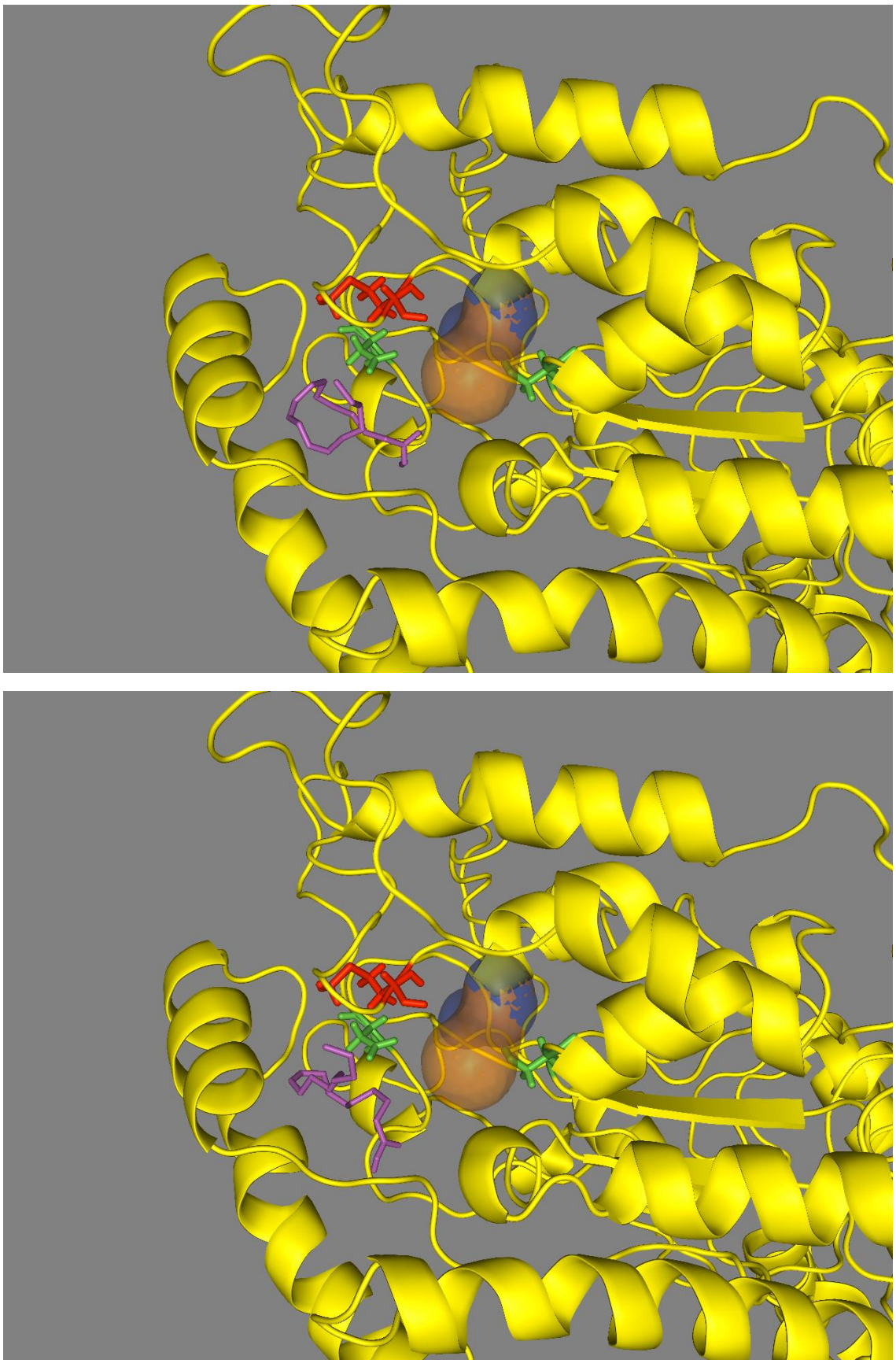


Figure 5.50 Primary docking mode of linoleic acid and oleic acid to I148M variant
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; linoleic acid.
Bottom panel; oleic acid.

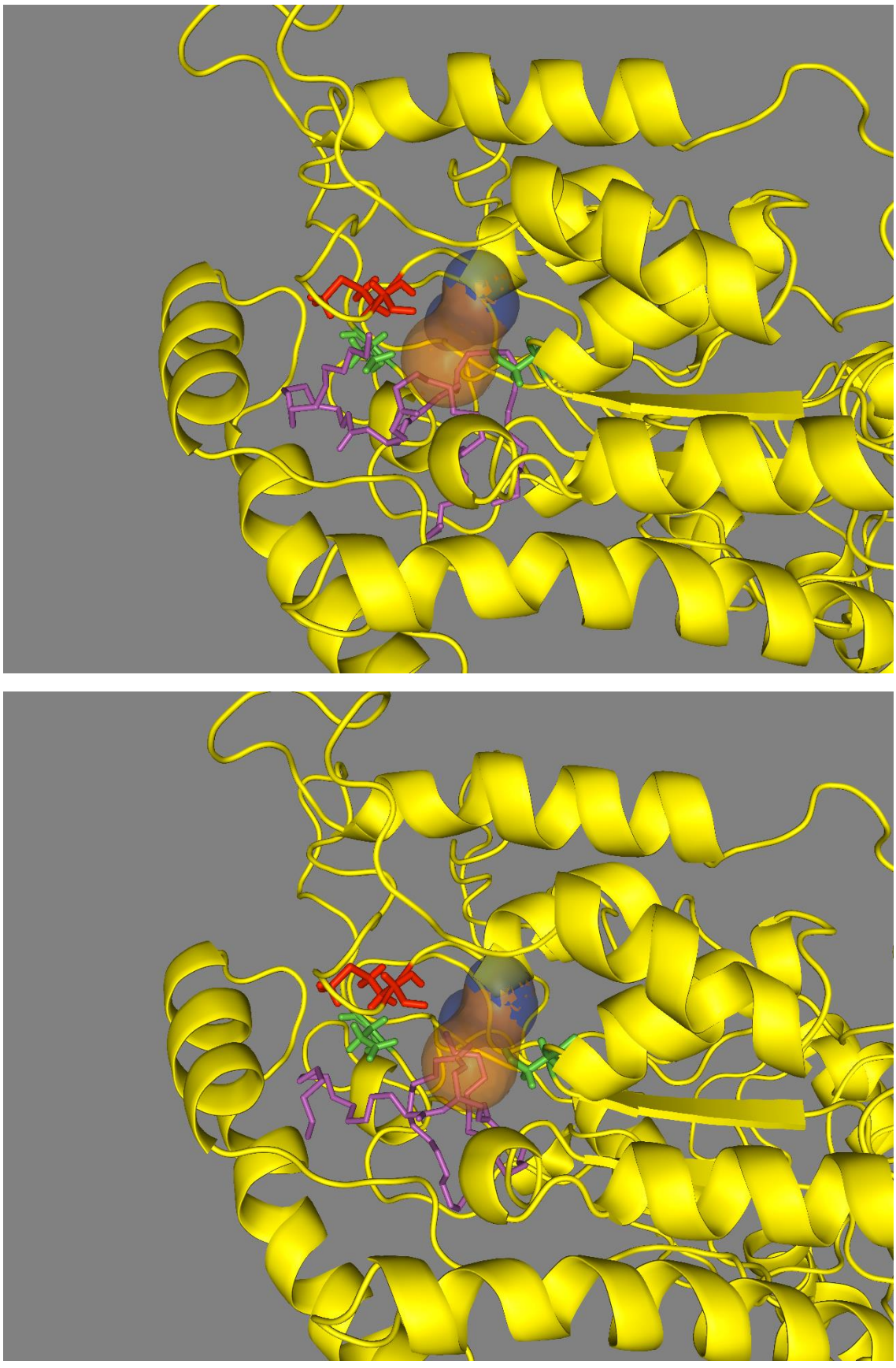


Figure 5.51 Primary docking mode of triolein and tripalmitin to I148M variant.
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; triolein.
Bottom panel; tripalmitin.

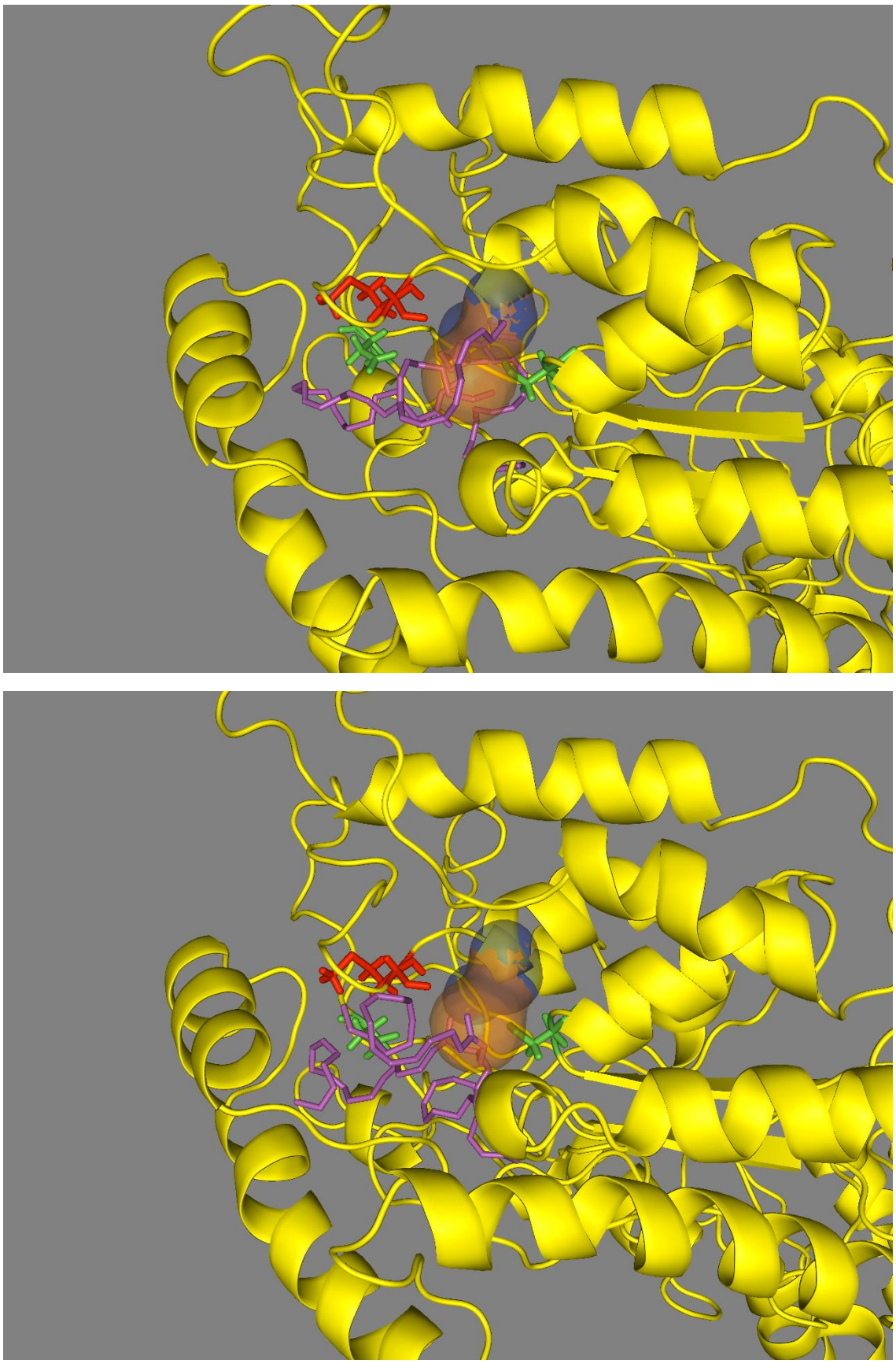


Figure 5.52 Primary docking mode of triarachidonin and trilinolein to I148M variant
Catalytic residues are highlighted in green, Isoleucine 148 in red, tunnel 1 in translucent blue and tunnel 2 in translucent orange. The docked ligands highlighted in magenta.
Top panel; triarachidonin.
Bottom panel; trilinolein.

5.5.5.5 Ligand binding strength

Although the calculations are not accurately quantitative, the binding strength for each ligand appears to be strong, based on low negative binding scores from AutoDock Vina (Tables 5.4 and 5.5).

In the wild type protein, the preference for ligand binding is retinol, retinoic acid, 1,2-diolein, palmitic acid, 1,3-diolein, 1,3-dilinolein, trimyristin, 1,2-dipalmitic acid, linoleic acid, oleic acid, triolein, tripalmitin, triarachidonin and trilinolein. This shows a particularly high affinity for retinol and retinoic acid.

In the variant protein, the order of affinity is altered and there is a clear change in ligand preference, however retinol and retinoic acid remain clearly the highest affinity ligands. Of particular note, triarachidonin, which was a very poor ligand in the wild type, is the third strongest ligand in the I148M variant.

The wild type enzyme appears to have stronger affinity for each of the ligands tested in the primary binding mode. The exceptions to this are triarachidonin and trilinolein which have higher affinity for the variant, and triolein which has nearly equal affinity. However, notably all of these exceptions do not have a primary binding mode in binding site A in the wild type enzyme.

Table 5.4 Ligand preference each protein

ligand	WT (kcal/mol)	Binding site	ligand	Variant (kcal/mol)	Binding site
retinol	-9.3	A	retinol	-8.1	B
retinoic_acid	-8.7	A	retinoic_acid	-7.6	A
1,2-diolein	-7	A	triarachidonin	-7.1	A
palmitic_acid	-7	A	1,3-dilinolein	-6.2	A
1,3-diolein	-6.6	A	1,2-diolein	-6.1	A
1,3-dilinolein	-6.5	A	palmitic_acid	-6	B
trimyristin	-6.4	A	triolein	-5.9	A
1,2-dipalmitin	-6.2	A	1,3-diolein	-5.7	A
linoleic_acid	-6.2	B	trilinolein	-5.7	A
oleic_acid	-6.1	A	linoleic_acid	-5.6	A
triolein	-6	B	1,2-dipalmitin	-5.5	A
tripalmitin	-6	A	tripalmitin	-5.5	A
triarachidonin	-5.5	C	trimyristin	-5.4	A
trilinolein	-5.5	C	oleic_acid	-5.3	A

Table 5.5 Overall affinity graphs

ligand	WT	Variant	Difference
1,2-diolein	-7	-6.1	0.9
1,2-dipalmitin	-6.2	-5.5	0.7
1,3-dilinolein	-6.5	-6.2	0.3
1,3-diolein	-6.6	-5.7	0.9
linoleic_acid	-6.2	-5.6	0.6
oleic_acid	-6.1	-5.3	0.8
palmitic_acid	-7	-6	1
retinoic_acid	-8.7	-7.6	1.1
retinol	-9.3	-8.1	1.2
triarachidonin	-5.5	-7.1	-1.6
trilinolein	-5.5	-5.7	-0.2
trimyristin	-6.4	-5.4	1
triolein	-6	-5.9	0.1
tripalmitin	-6	-5.5	0.5

5.5.6 Multiple one nanosecond repeats

All thirty of the one nanosecond simulations including the three PNPLA3 variants (148I, 148M and 148A), all remained stable throughout simulation. The previously observed conformational changes were not replicated in each of the runs, however there were overall trends which remained consistent with these findings.

The wild type protein variant obtained a similar conformation to 100ns simulation within 40% of the simulations; during which the catalytic residues remained in close contact (<6A), suggesting putative active conformations. However, this condition was not encountered in any variant repeats, in which all simulations positioned the catalytic residues more than 6A apart.

To test whether this apparent conformational change between variants is derived from the loss of isoleucine or gain of methionine, an additional 148A variant was simulated. Under simulated conditions, the 148A variant PNPLA3 mirrored the WT, with producing an active conformation across 50% of the simulations.

Pairwise Wilcoxon rank sum tests were applied on the average distances between residues to confirm that these observed changes were statistically significant.

All of the distances tested were significantly different between the I148M variant PNPLA3 protein and the I148A variant, while there were no statistical differences between the wild type and I148A variant (Table 5.5).

Between the wild type and the I148M variant, the distances between the catalytic residues (47 and 166) and the distance between residues 148 and 166, were both statistically different between the WT and the VAR (Table 5.6).

Each of these changes were represented by a directional shift within the catalytic pocket of the protein. In general, the I148M variant positioned the catalytic residues further apart, while the methionine residue was also closer to both catalytic residues, with a particular preference toward residue 166 (Figure 5.53).

Table 5.6 Wilcoxon rank sum tests comparing distances of PNPLA3 catalytic residues

Protein variants compared	Residue distance observed	P value
WT and VAR*	47, 166	0.02807
WT and VAR	47, 148	0.7969
WT and VAR*	148, 166	0.03998
WT and ALA	47, 166	0.4779
WT and ALA	47, 148	0.08795
WT and ALA	148, 166	0.3
VAR and ALA*	47, 166	0.02331
VAR and ALA*	47, 148	0.004102
VAR and ALA*	148, 166	0.004102

*statistically significant to 0.05 confidence limits

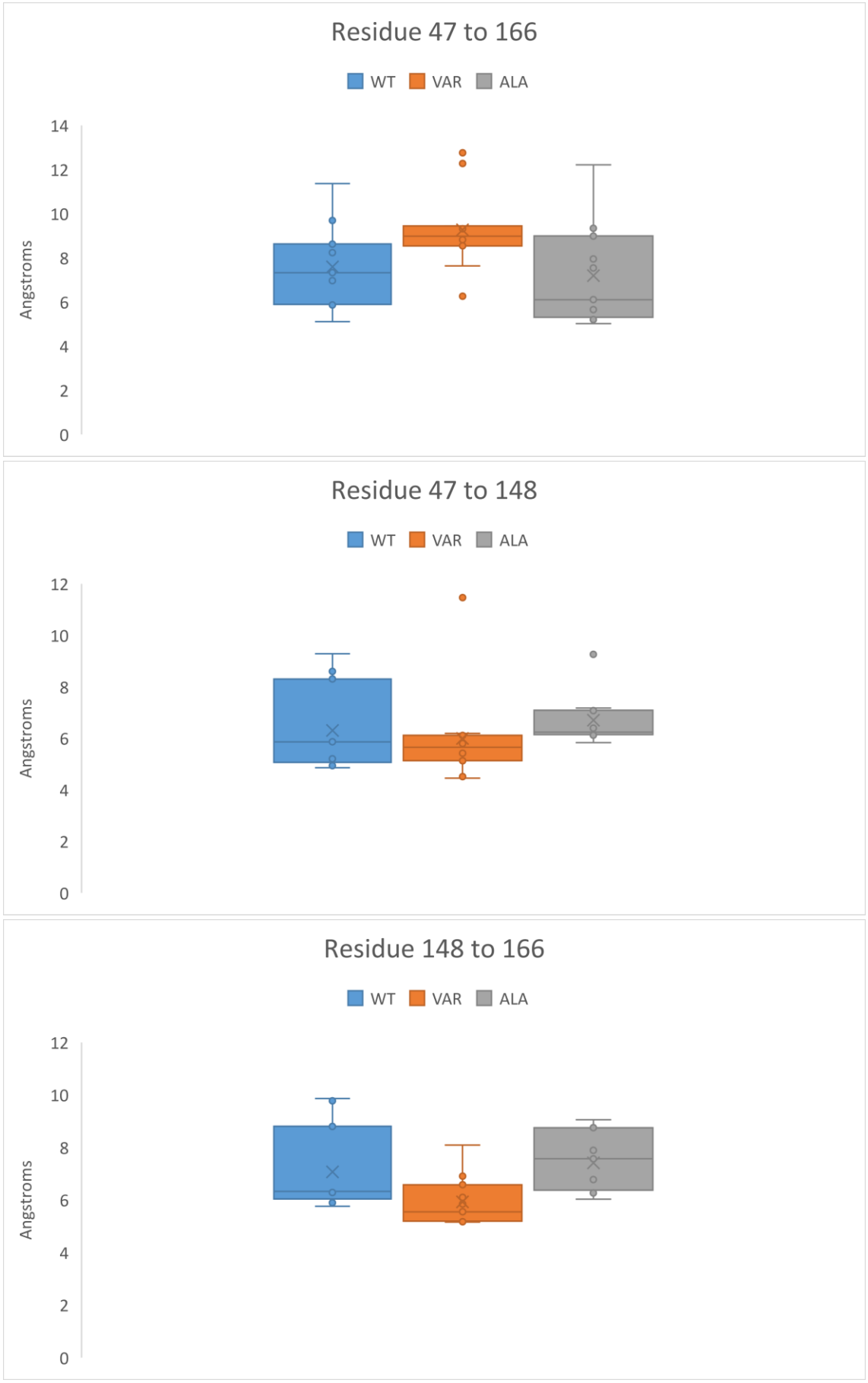


Figure 5.53 Boxplots representing the distances between the key residues across all 30 additional 1ns simulations (grouped by variant)

5.5.7 Ubiquitination sites

Ubiquitination sites were mapped to the surface of each protein variant (Figure 5.54).

Residue 179, the most likely ubiquitination site as predicted by iUbiq-Lys, is not exposed on the surface in either PNPLA3 variant. Residue 263 the most likely as predicted by UbPred is slightly more exposed in the wild type protein, however not largely exposed in either variant.

Residues 441 and 479, predicted to have moderate ubiquitination potential, while they are exposed surface residues in both variants, are positioned very differently because of the location of the C-terminal domain. In the wild type, these residues are exposed almost entirely into the solvent, while in the variant they are exposed, but positioned into the cleft between the left and right lobes of the protein.

Finally, the least probable sites of ubiquitination as predicted by UbPred were significantly changed in each variant and again more exposed in the wild type protein. In particular residue 434, was far more exposed in the wild type being positioned on a flexible loop in the start of the C-terminal domain, whereas it is shielded in the protein in the variant structure (Figure 5.55).

5.5.8 Surface hydrophobicity

The surface hydrophobicity of both protein variants shows no clear changes in patterns (Figure 5.56).

Notably both variants have hydrophobic residues lining the outside of the tunnels, which could facilitate interaction with lipids.

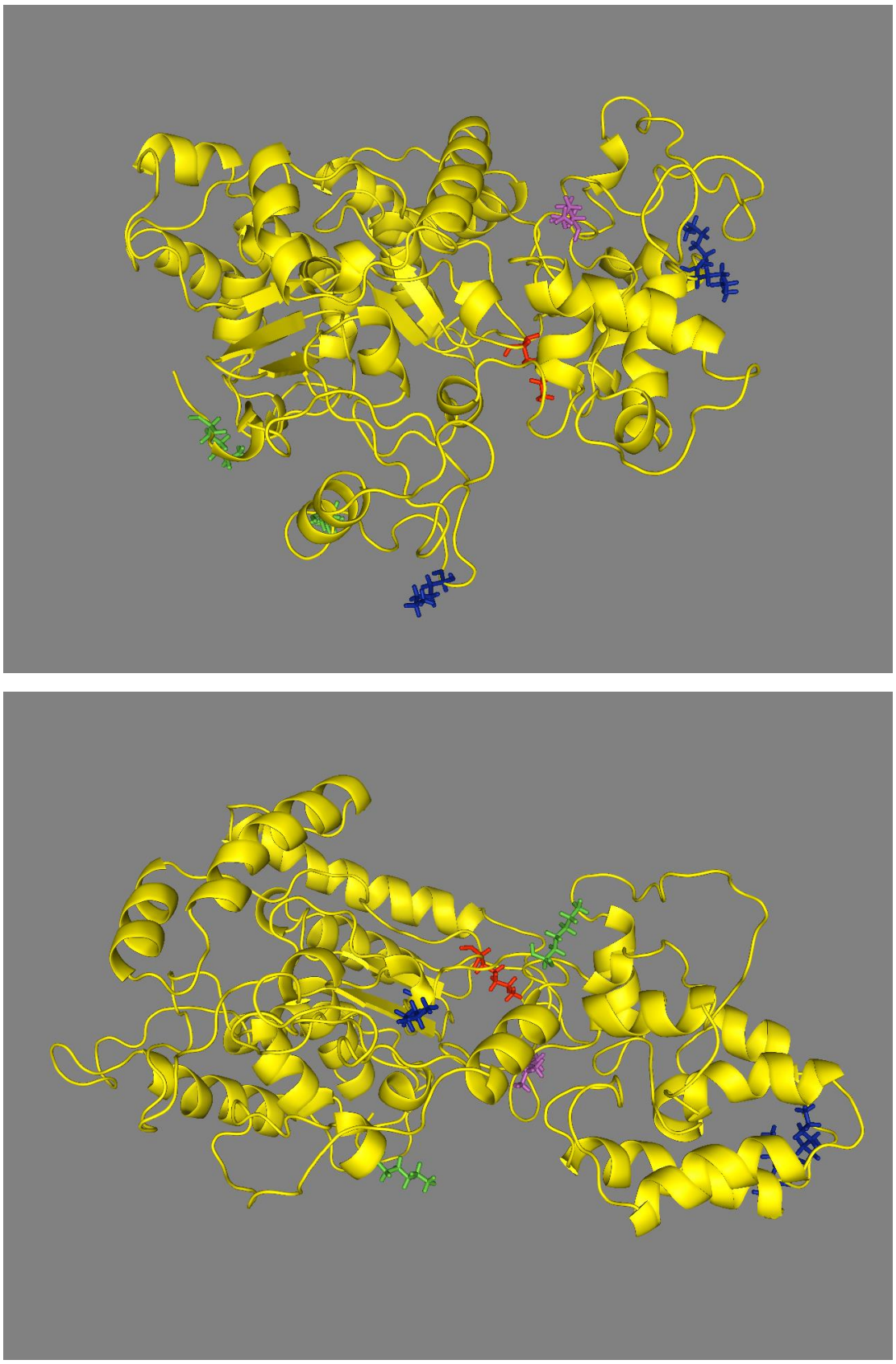


Figure 5.54 Putative ubiquitination sites within the PNPLA3 structure

Residue 179 predicted by iUbiq-lys is highlighted in red, residue 263 highly predicted by UbPred in magenta, residues 441 and 479 moderately predicted by UbPred in green and residues 333, 334 and 434 possibly predicted by UbPred in blue.

Top panel; wild type.

Bottom panel; I148M variant.

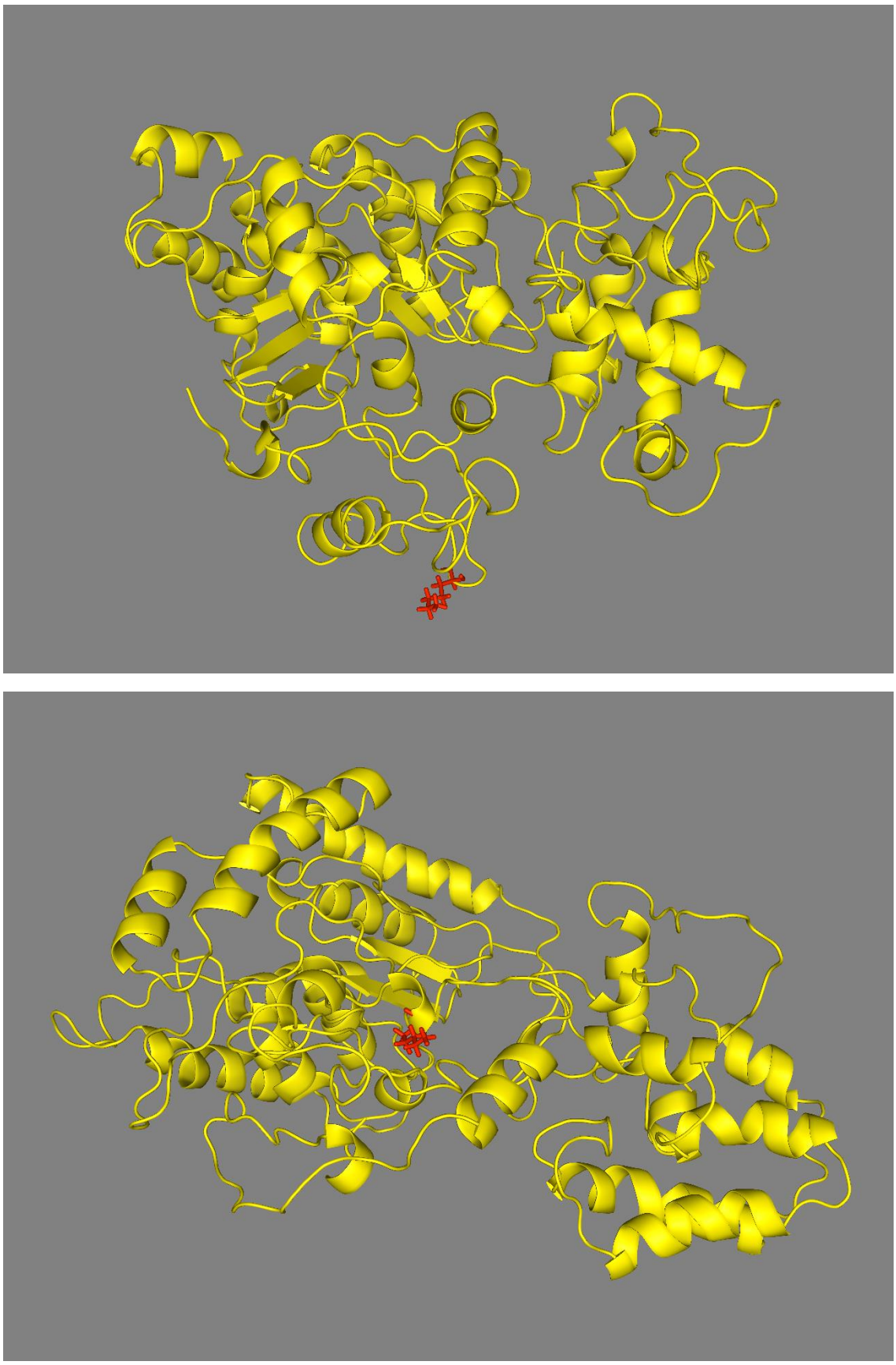


Figure 5.55 Position of Lysine 434 within the PNPLA3 structure

Residue 434 is highlighted in red.
Top panel; wild type.
Bottom panel; I148M variant.

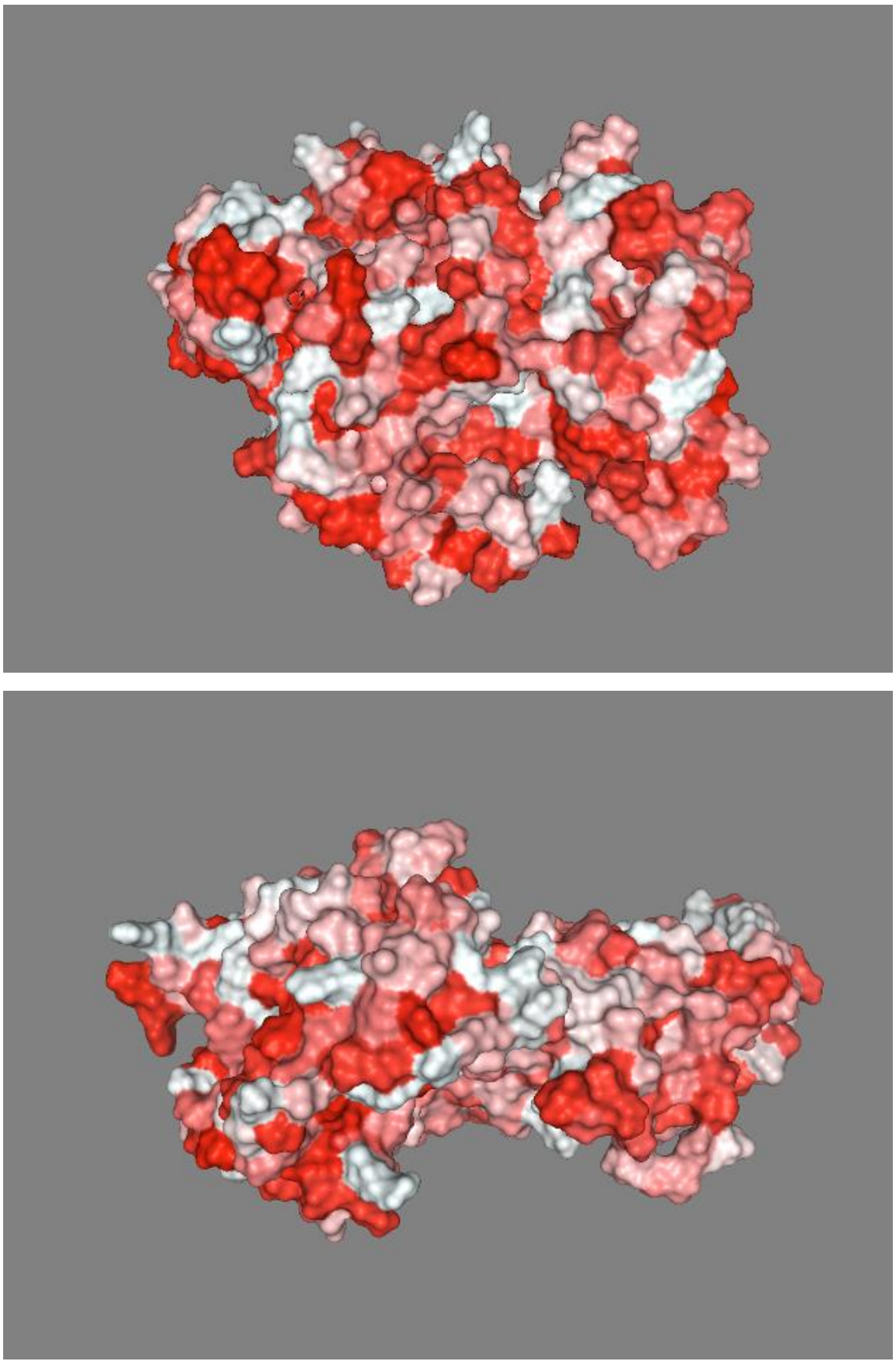


Figure 5.56 Hydrophobicity map of the PNPLA3 surface

Hydrophobic surfaces represented in red.
Top panel; wild type.
Bottom panel; I148M variant.

5.6 Discussion

5.6.1 Investigatory model simulations

The models which were generated were all based on low homology alignments, meaning there is a large potential for structural errors, particularly on the local scale. The fact that all the systems generated were able to be simulated without significant degradation or unfolding of the protein, lends support toward the basic structure of these models.

Overall, there was a small increase in RMSD during the 20ns timeframe in nearly all simulations, which shows that a true equilibrium had still not been reached despite the large equilibration step that was performed. The most likely explanation for this is simply the distance of the starting structure from true structural equilibria, as would be expected from models generated with such low homology.

5.6.1.1 Domain architecture

The idea that PNPLA3 consists of multiple domains has been suggested throughout the length of research on the protein. Namely the patatin domain appears to form a distinct domain of the protein based on homology alone.

When observing homology across a broad range of homologues, homology primarily extends between two distinct regions spanning residues 1-250 and more rarely, residues 300-400. This further supports two unique domains, with a linker region that is more amenable to mutation.

The consistencies in the observations across different models that were simulated, offer insight into the dynamic propensities of the PNPLA3 structure, and lends further credence to PNPLA3 being constructed of multiple domains.

Previously the patatin domain was determined as consisting of residues 1-179, based on the homology with patatin. The fact that the structure up to residue 239 also remains a highly stable substructure with the patatin domain under simulation conditions offers strong support for this region being included in the patatin domain. This is a reasonable assumption, particularly considering alignment with ExoU actually extends into this region, and it is the longest model which can be achieved using simple threading alone.

Truncated studies of protein expression, which have tried to express the patatin domain thus far have focused on either this smaller domain, or a larger domain extending up to 250 residues.

Domain boundary selection is always a complicated issue, and not having this stabilising helix may stop the protein functioning in a normal capacity and could even contribute to a shift from one predominant activity to another. This is an important consideration both for expression and for *ex vivo* overexpression studies. The fact that the RMSF increases rapidly around residue 239 implies this may be a more relevant domain boundary.

During these simulations the most stable portion of the models was the N-terminal 239 residues. In fact, in several simulations the C-terminal seems to be gradually losing its confirmation, suggesting that this is not a stable confirmation for this portion of the protein.

In general, the RMSF of the protein backbone showed an ordered structural protein, however there was notable increased fluctuation in the 200 to 350 residue range of the longer models. This corresponds with random coil predictions in the secondary structure, and therefore likely be more flexible. This would be an expected characteristic of a linker region and is likely a transition between domain fragments.

On visual inspection, we can readily observe in the model that the protein is divided into two separate lobes. The left lobe containing the patatin domain and the right lobe consisting of this second domain, we call the right lobe domain. However, the final C-terminal portion of the protein folds back to form part of the left lobe with the patatin domain.

Because of the separation with the right lobe domain, coupled with the loss of homology in this region, a third domain, the C-terminal domain is proposed. This domain is spatially oriented into the left lobe of the protein, and primarily interacts with the patatin domain thereby making it a unique structural element.

This provides us with a three-domain model of PNPLA3, which can be used to accurately describe the unique domain shifts we observe between the wild type protein and the I148M variant under simulation.

5.6.1.2 Distances between S47, D166 and I148

Across the different models there was a broad range of distances between the catalytic residues, ranging from 5A to 22A. When the distance between these residues was over 8A, this was coupled with inherently large variations across the simulation. So, it is likely that this represents instability within this region of the protein, rather than an alternate stable conformation.

Models 1, 2 5, 6 and 9 had the most consistent distance between the three residues of interest, and in all cases the Isoleucine 148 was further from the catalytic serine than the Aspartate.

5.6.2 Selecting the model for further investigation

Selecting the best performing model based on simulation results presents several challenges, as caution must be taken to avoid over analysing the initial results, which may bias the findings in the long timescale simulations. This meant there was an active decision not to simulate a variant, nor to visually inspect this group of simulations. The aim was to maximise the stability of the simulation, with particular weight given to the regions around the catalytic residues and patatin domain, which was predicted with higher confidence.

Simply selecting models which performed best under these simulation conditions, does lead to an innate bias toward stable conformations, which may not truly represent the system. It could be argued that a more confident model should have been selected above the best performing model, because of issues differentiating between fluctuation caused by low quality and inherent flexibility a stable conformation was more suitable for further work without an experimental structure to guide the study.

Interestingly it was not the models predicted to have highest structural quality that remained most stable under simulation as may be predicted. Although model 3 had the most confidence and the best predicted model quality, under simulation this model was unstable.

This may be due to the fact that the fragment length does not sufficiently represent a full domain to maintain the structure and could explain why previous investigations using a model of this length were only run for 10ns. This shouldn't be used as evidence against the initial structure defined, merely that it is not able to adequately represent the whole protein under longer simulations.

The best candidates to take forward were deemed models 2, 4 and 5, as these models had the greatest stability in the patatin domain. Model 5 was selected as the best performing simulation and was used for the basis of molecular dynamic investigation into the I148M variant for several key reasons:

Firstly, model 5 was predicted to have the best local stability within the active site. While model 4 showed very stable conformations in the patatin domain, it exhibited an increasing distance between I148 and D166, while the distances between the catalytic residues seemed more

chaotic throughout the simulation. Similarly, the distances between I148 and D166 in model 2 fluctuated wildly. This raised concerns over the local stability of the active site in these other models.

Secondly, despite a seemingly large increase in RMSD over the simulation in model 5, the active site residues and I148M loop remained highly stable. The majority of the RMSD shift appeared to be caused by inherent movement between domains of the protein, as this was the only full-length candidate. The RMSD in this model further influenced by flexible linker regions, which could be removed to improve the visual appearance of the graph. This was highlighted by the more stable RMSD of the patatin domain alone.

Finally, while selecting a protein truncation length for simulation, there is the inherent risk that a truncation that is too short will be missing vital elements needed to observe the natural behaviour of the system, as implicated by the simulation result of model 3. Experimental expression trials of shorter clone fragments also performed poorly within chapter 3, when coupled with the fact patatin is often found in a denatured state *in vivo*, posed potential problems with these shorter domain fragments.

Since the full-length protein of model 5 performed adequately under simulation, the longer variant was deemed more interesting to investigate, and it was felt the risk of missing key interactions without this domain outweighed the risk of it negatively impacting the simulation.

5.6.3 Full length simulations

To be as consistent as possible between variant structures, a simple substitution was implemented to the starting structure of model 5. In the absence of a computational approach capable of adequately modelling structural changes which may occur from a single residue change, as the subtle sequence variation would often be overridden in threading portions of the algorithm, this was deemed the fairest approach.

5.6.3.1 Quality of simulations

The final 100ns simulations while producing simulations of adequate performance were still non-ideal, with notable observed RMSD changes in the structure over time. Despite this, the simulations showed minimal fluctuation of less than 0.5A at any given time point, showing there was inherent stability in the local environment.

Upon closer observations, large amounts of the RMSD increase was due to significant domain movements in the right lobe domain and C-terminal domain. The resulting graphs could be made to appear artificially more stable by removing flexible loops from the analysis.

Most importantly for this study, the patatin domains of both variants were stable throughout simulation, after the first 20ns, in which the system further equilibrated. This is crucial, as the patatin domain is the region of highest interest in the structure.

It is notable that the wild type protein appeared to achieve a stable structure in the last 20ns of simulation, where no further domain movements were observed. For this reason, the last frames of each simulation were chosen as the most reliable structures of each variant for further investigation.

5.6.4 Notable structural changes

5.6.4.1 Global structural changes in PNPLA3

When observing snapshots of the protein variants throughout simulation, the I148M variant again was clearly more stable. The RMSD fluctuated less than the wild type protein, and little movement from any key regions can be observed.

Indeed, when comparing the secondary structures of the final models with the predicted secondary structures, we can see that the patatin domain in the isoleucine variant maintains two key predicted β-sheets in the patatin domain and is likely a more accurate representation of the functional protein unit (Figure 4.57).

In contrast to this, the wild type protein displayed much larger changes in overall structure. The right lobe of the protein was observed to compress into the left lobe, and the C-terminal domain transition across the face of the protein. This caused an overall compression of the protein which led to a much more globular appearance, and less ordered right lobe domain.

The observed shift is coupled with an observed bend of 45^{o} in helix I, which allows the right lobe to be positioned across the face of strands 1 and 2, as compared to extending parallel beyond the base of strands 1 and 2.

The question remains as to the validity of this C-terminal conformation, as it appears clearly less stable than the right lobe domain in the I148M variant simulation. However, if this is to be

believed, then it may present the largest conformational shift caused by this variant which could have massive ramifications on the overall interactions within this domain of the protein.

This large structural and dynamic variation between the protein variants lends strong support to the large structural impact this change may have on the PNPLA3 and in turn deleterious pathogenic effects.

5.6.4.2 Differences between stabilities

As expected, based on the experimental data, the variants had largely differing molecular dynamic profiles.

In general, the I148M variant was much more stable throughout the simulation with very little change in the full length of the protein. The wild type appeared to have more interdomain flexibility, with large RMSD changes in the C-terminal half of the protein. It is unclear to what extent the variant is responsible for the changes in overall flexibility, however it is likely it plays a role in this feature.

Notably there is a large degree of disorder observed in the C-terminal portion of PNPLA3, which is much greater in the wild type protein.

This may be caused by random chance and poor modelling of a region which shares low homology; however, it cannot be ruled out that this is in fact a functional level of disorder facilitating the proper functioning of PNPLA3.

There has been a long-held assumption that folded protein structure is necessary for the protein function. However it has become increasingly appreciated that not all proteins are folded into a globular state to achieve their function, and in fact disordered regions of the protein may be crucial to a range of function including promiscuous binding to a range of partners or ligands.[464]

Under these assumptions, the I148M variant could negatively impact binding with a binding partner in this region because of the reduced flexibility.

5.6.4.3 Active site pocket changes

Of course, residue 148 is located in close proximity to the active site, which coupled with the conferred loss of function proposed by experimental work, has led to vast interest in the potential impact the I148M variant has on the active site.

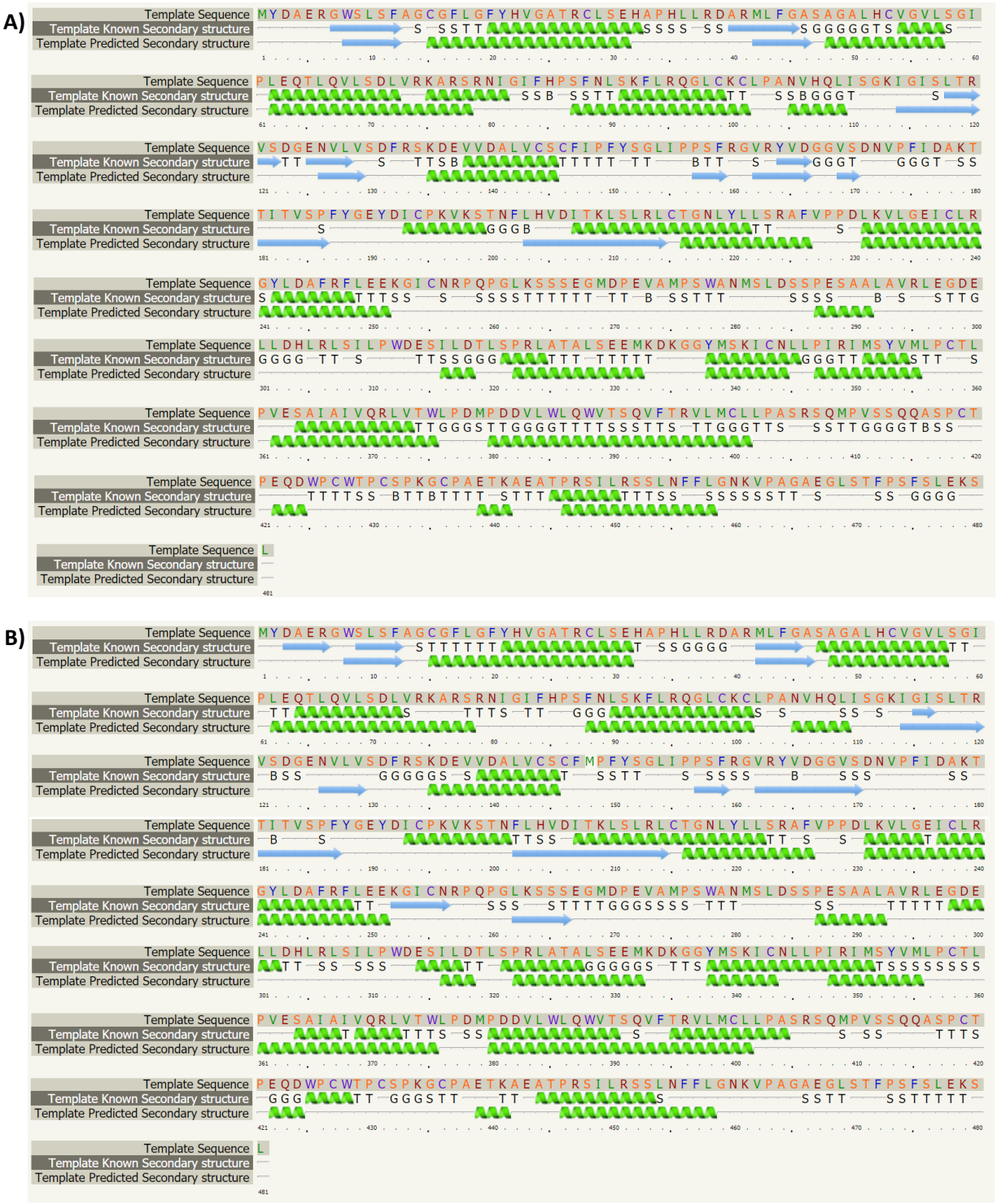


Figure 5.57 Secondary structure against predicted secondary structure of final models

A) Wild-type PNPLA3 **B)** I148M variant PNPLA3. For each model the sequence is represented along the top row with single letter abbreviations representing each amino acid in the protein chain. The "known" template in the figure refers to the respective molecular model, with the predicted secondary structures of the region shown below; the β-sheets represented as blue arrows; the α-helices are represented as green helices. Additional assigned secondary structures by the program DSSP are represented by single letters as follows:

G: 3-turn helix (310 helix);
I: 5-turn helix (π helix);
T: hydrogen bonded turn;
B: residue in isolated β-bridge;
S: bend.

Previous investigation into the dynamics of PNPLA3 were used to suggest the I148M mutant has little impact on the catalytic pocket as a whole, but rather, it is discussed that the 148M variant displays a less open conformation, with less access to the active site.

In the wild-type enzyme, the distance between Cα of Ile148 and the midpoint of line segment connecting Cαs of catalytic dyad was used in these studies to determine access to the active site.

Our findings shed new light on previous investigations into the dynamic structure of PNPLA3, as once the additional residues of the enzyme are included in simulation, there are additional helices, which form a "lid" across the active site. This makes the previous measurement of the active site opening misleading. Taking this into account, it seems unlikely the small residue change significantly protrudes into the active site or has the potential to actively influence the size of the active site by steric hindrance.

In our simulation, the 148M variant is observed to largely disrupt the active site pocket and potential impact even further structural changes beyond the active site.

A key observation here is that in the wild type the catalytic residues S47 and D166 are located in close proximity throughout at under 5A between their centre of masses. At this distance there is very strong catalytic potential, and the active site appears to remain stable in this position.

The proposed catalytic site of PNPLA3 is comprised of the catalytic dyad Ser47 and Asp166. The most probable catalytic mechanism for this catalytic dyad would be consistent with the general hydrolases, whereby the adjacent Aspartate would act as a general base, whose Oγ2 would activate the Oγ of Serine47 by abstracting its proton. This would turn Serine 47 into a potent nucleophile, which could then attack the substrate (Figure 5.58).[401,408,465,466]

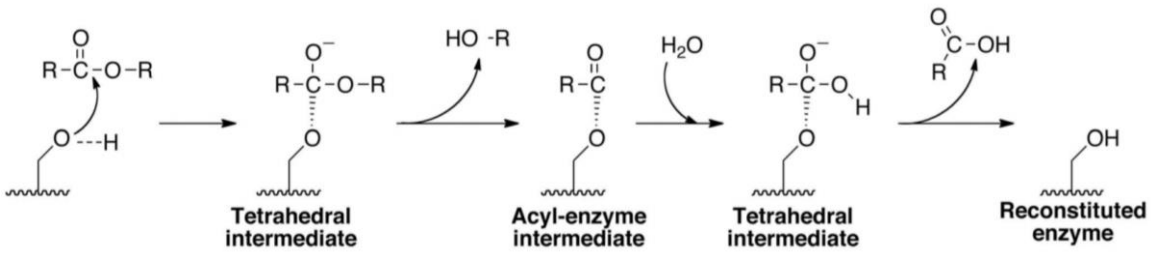


Figure 5.58 Proposed catalytic mechanism for the active site of PNPLA3

Asp166 (not shown) acts as a general base and activates the Ser47 nucleophile by abstracting its terminal hydrogen. The activated Ser47 (shown as –OH) attacks the acyl carbon of the substrate forming a tetrahedral intermediate whose negative charge is shielded by the oxyanion hole of PNPLA3 (not shown). Loss of R-OH yields an acyl-enzyme intermediate that is hydrolysed rapidly (passing through another tetrahedral intermediate) to release the acyl moiety and regenerate the enzyme (Adapted from Wijeyesakere *et. al* 2014).[408]

Based on this mechanism, the simple effect of disrupting the stable catalytic form within the active site, would be enough to inhibit all catalytic activity, as there is no longer a readily available base to activate the catalytic Serine.

The I148M variant has a vastly different architecture, where the catalytic residues are now separated by over 10A. This is coupled by a new stable interaction between methionine 148 and aspartate 166. This structural change provides a novel putative mechanism by which the I148M variant may cause a loss of lipase activity.

The shift in the active site conformation observed, occurs early on in the simulation, beginning in the heating stages. This does not appear to occur due to any overtly large Gmax forces, but as a natural shift on the protein architecture.

In fact, the wild type protein does not start in a position of a perfectly formed active site. However, upon simulation the catalytic residues move together, whereas in the I148M variant this is disrupted by the methionine and the catalytic residues move apart. In this instance the I148M may simply destabilise the catalytic conformation no longer allowing the protein to easily transition into a closed conformation.

Notably, the variant would not need to completely inhibit this conformation, but by destabilising one conformation and stabilising another, simply shift the equilibria in the direction of non-formed active sites, instantly modifying the activation energy and rate of catalysis.

Interestingly, previous simulations of PNPLA3 by Xin *et al.* also showed an increased distance between the catalytic residues in the variant protein, and a shortening of the distance between D166 and I148, which agrees with the conformational change observed in our data. However, the time scale of the transition was much slower, and the simulations too short to observe whether a similar stable active site substructure would have been formed (Figure 5.59).

In addition, we observe the I148M variant having more specific impacts on the local secondary structure of the protein. Notably on the β-strand formation of strand 3, 4 and 5, all of which are absent in the variant simulation.

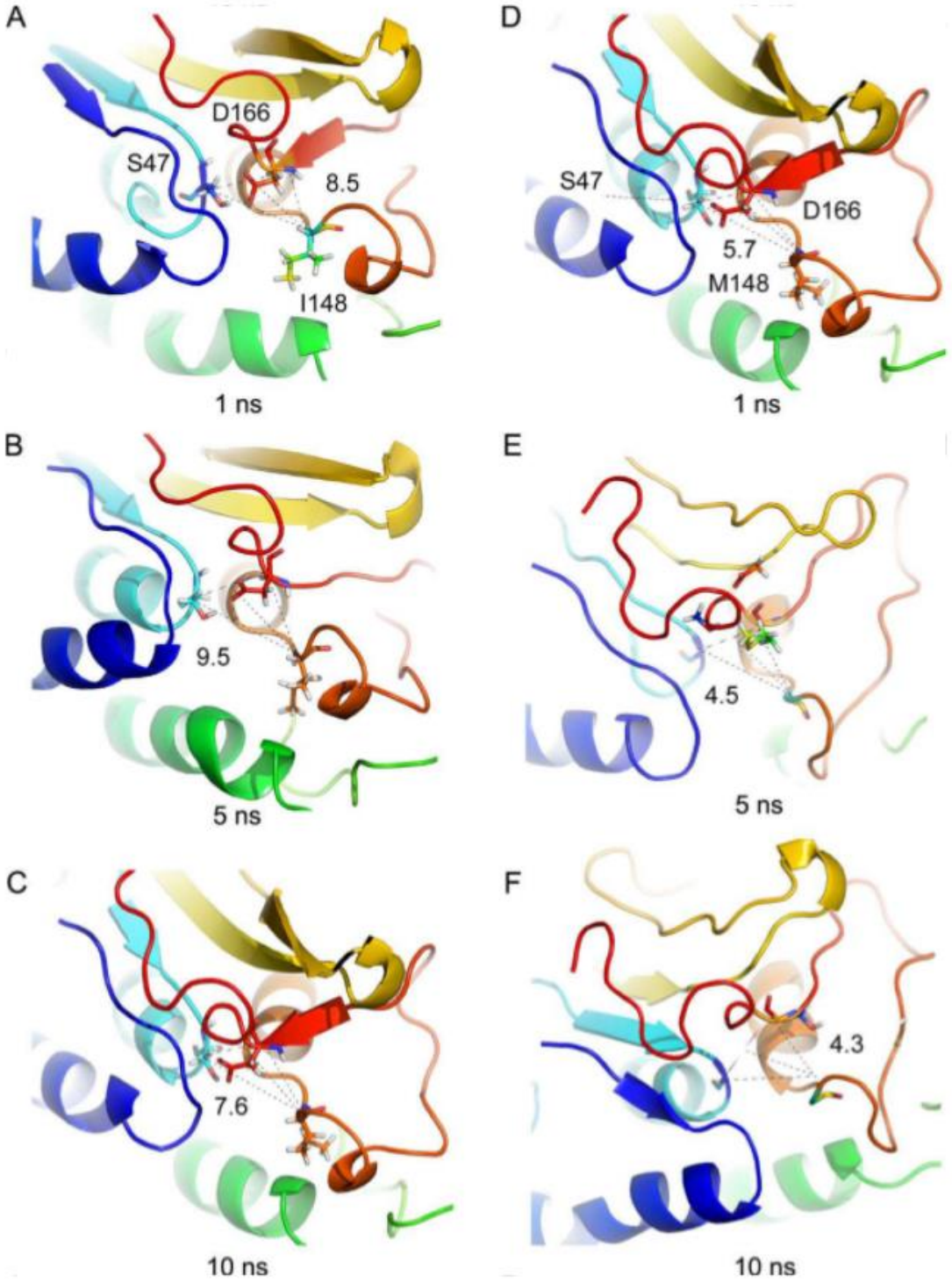


Figure 5.59 Structure snapshots of wild-type and I148M mutant enzyme in substrate-free systems

(A–C) present conformations of the wild-type protein at 1, 5, 10 ns, respectively, while (D–F) present the I148M mutant (Adapted from Xin et. al 2013). [412]

Again, this loss of β character was observed in simulations performed by Xin *et. al*, under different conditions with a different starting model, lending support that this may be a biologically accurate representation of the impact on the local structure (Figure 5.58). However, this change was not discussed.

Remarkably, the I148M variant triggers a loss of β character from β-strand 5 almost instantaneously on heating the system. This is then coupled with loss of alignment with strands 3 and 4, and further loss of β character of the strands. While the exact causative sequence of this cascade is unclear, this appears to be a vital step in the loss of the constrained active site structure which facilitates the formation of the contacting catalytic residues.

We attempted mutating the final variant residue 148 back to Isoleucine, to see whether we would recover the active site, however the protein remained stable throughout in the variant conformation. This could suggest that this is an alternate conformation which the protein could adapt *in vivo*.

5.6.4.4 Lipogenic catalytic functions

The lipogenic catalytic ability of PNPLA3 has remained under debate across the literature, with some studies finding gain of lipogenic function in the I148M variant. Although the catalytic position, is greatly altered in the variant and confirmation of the enzyme impacted, the overall stability of the active site is maintained, which does suggest potential for catalytic activity remains.

The predominant lipogenic functions require Aspartate and or Histidines in close proximity to a catalytic serine for the most common mechanisms of action, however we found no close Asp or His residues with catalytic potential within the active site (results not shown). This would suggest that a lipogenic function of the protein in this model is unlikely. However, the active site could facilitate natural lysis simply by forming a stable binding site for relevant molecules.

5.6.4.5 Oxyanion hole

The putative oxyanion hole is so broadly conserved across homologous proteins, that it is highly likely this plays a vital role within the enzyme function.

In both variants of PNPLA3, the oxyanion hole was positioned close to the active site, suggesting there was the potential to act as stabilising entity during catalysis. However, in the final

structural positions, the residues which constitute the oxyanion hole in the I148M variant are positioned with the positive amides away from the active site itself. This could be a side effect resulting from this variant existing in the APO form, however, could be an additional factor reducing the potential for catalysis within this variant.

5.6.5.6 Active site tunnels

Ligand access to the active site pocket is often driven by the geometry and properties of tunnels leading to the binding pocket within the protein. It is vital to assess the impact an amino acid variation may have on the shape and properties of these key access channels.

In particular, it has been suggested on multiple occasions that the I148M variant, in fact protrudes into the binding site of PNPLA3 and would therefore inhibit the putative lipolytic function. To investigate this hypothesis tunnels to the catalytic residues were actively mapped within the structure.

Coupled with the change in the overall architecture of the active site is a drastic change in the access tunnels to the catalytic serine. When observing tunnels to the catalytic serine, there is a slight reduction in the radius of the tunnel in the variant of 0.5 A, however this is a specific tunnel to the catalytic site and does not account for the much larger active site cavity. Additionally, the variant tunnels are shorter and would therefore be easier to access by a substrate.

This shows that the I148M variant does not cause reduced access to the active site.

Notably, the wild type protein shows 4 potential channels to the active site which could theoretically play a role in catalysis, whereas the variant only has 2 large tunnels. Additionally, the wild type tunnels carry a net negative charge, while the variant tunnels carry a net positive charge which could affect ligand binding.

An alternate simulation that was performed on a wild type model also showed a similar tunnel architecture reinforcing the current findings (Figure 5.60).

It is not possible to tell which tunnel is the correct ligand binding tunnel, however in the wild type protein, tunnel 1 and 2 do form a long channel which is connected to the conserved oxyanion hole. This would suggest that these tunnels are the position we would expect to observe ligand binding.

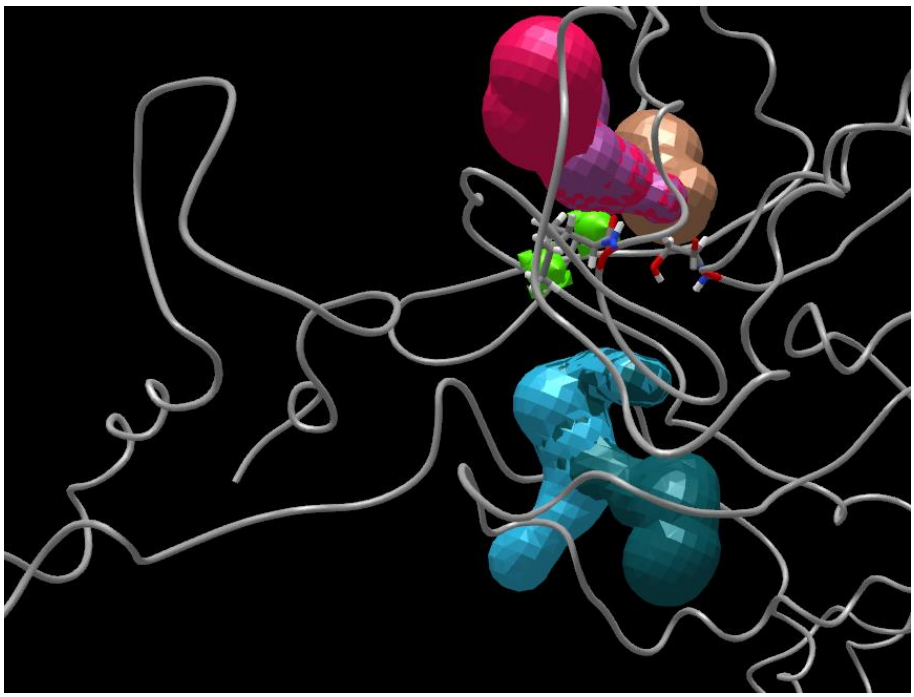


Figure 5.60 Tunnels to catalytic serine in alternative wild type simulation

5.6.5 Docking

Docking was performed without accounting for structural changes in the enzyme that could occur during an induced fit model of binding. This was done to assess the initial binding of the ligands within the current models and not accurately position fully bound ligands, because the binding is likely to be strongly impacted by the lipid droplet surface on the protein.

The WT clearly has much more complex binding patterns than the variant and seems to facilitate more interaction with lipids in other surfaces of the protein. It is clear that in a lipid rich environment, this could have a large impact on the stability of the protein structure and stability.

Within the wild type protein, tunnels 1 and 2 seem to act as a protein channel, allowing ligands to dock through the protein. This allows the larger fatty acids to present to the active site and would appear to be the most important binding mode for putative catalysis. This leads us to predict the tunnels 1 and 2 are imperative for catalytic function. This theory is supported further by the fact that this is where the conserved oxyanion hole is positioned.

Tunnel 4 is the only tunnel with I148 lining the tunnel. No ligands were shown to bind in this cavity at all. This implies even further that methionine in this position does not reduce access to the active site in the wild type protein.

While binding energies of each ligand have not been measured experimentally, PNPLA3 has been shown to have the highest activity against TAGs, and in particular with a preference for Oleic fatty acid groups.[165] Our binding did not show a clear link between these ligands which have higher activities and strength of binding in the docking study. This could mean the binding affinity is not involved in a rate limiting step or be a side effect of not having a more advanced flexible docking approach. Although this is the only ligand which appeared with strongest binding modes in tunnel 3 of the protein.

The best triolein binding mode exhibited a preference for a terminal acid to enter the tunnel which could support preference of cleavage at SN-1 and sn-2 positions. Of course, affinity is only one factor to be considered, and the actual binding of the ligand and catalytic intermediate energies will play a much larger role in understanding this pattern.

The strongest ligand binding was observed with retinol and retinoic acid in both the wild type and variant protein. These ligands were orders of magnitude stronger binding agents, and supports a role for retinol-esterase activity of PNPLA3.[143]

Xin *et. al* performed docking of palmitic acid, and found the binding energy was significantly reduced in the I148M variant (Figure 5.61). Interestingly, palmitic acid was the only ligand in our docking which also showed greater affinity to the wild type variant of PNPLA3 (-5.9, -5.8 Kcal/mol respectively). All other ligands bound with greater affinity to the variant protein.

We also observed similar predicted binding sites within our wild type model, however our docking was performed on a final simulated structure which had already undergone conformational shifts.

During simulations with the substrate they found that the mutant enzyme is more stable, while the wild type enzyme showed high variation. Based on the findings of this study, this type of ligand-based domain shift could be due to less intrinsic stability in the APO conformation, so readily adapted by the variant protein.

We chose not to undertake further simulation of ligand complexes at this point in time, as it is unlikely that the docked conformations are of adequate accuracy to facilitate a reasonable simulation.

Notably, in the variant protein, ligands dock into the full active site cavity within the protein. During this binding, the protein is essentially locked in a conformational space by which the catalytic residues cannot move together spatially. This would lead to the protein being locked into an inactive conformation (Figure 5.62).

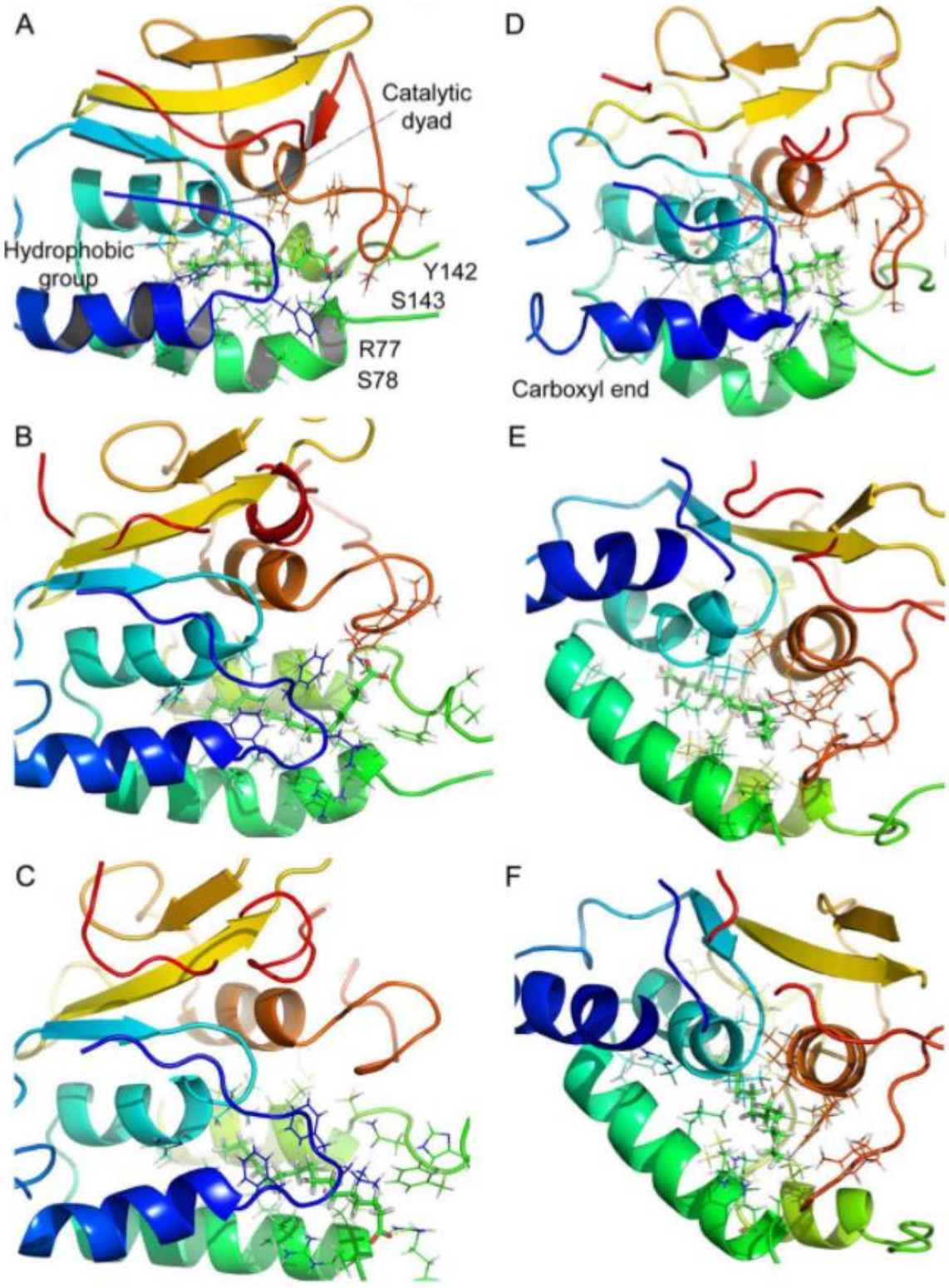


Figure 5.61 Structure snapshots of wild-type and I148M mutant enzyme in substrate-bound systems

Subplots (A–C) present conformations of the wild-type protein at 1, 5, 10 ns, respectively, while (D–F) present the I148M mutant. The residues in substrate-binding pocket are displayed in the stick form (Adapted from Xin *et al.* 2013).[412]

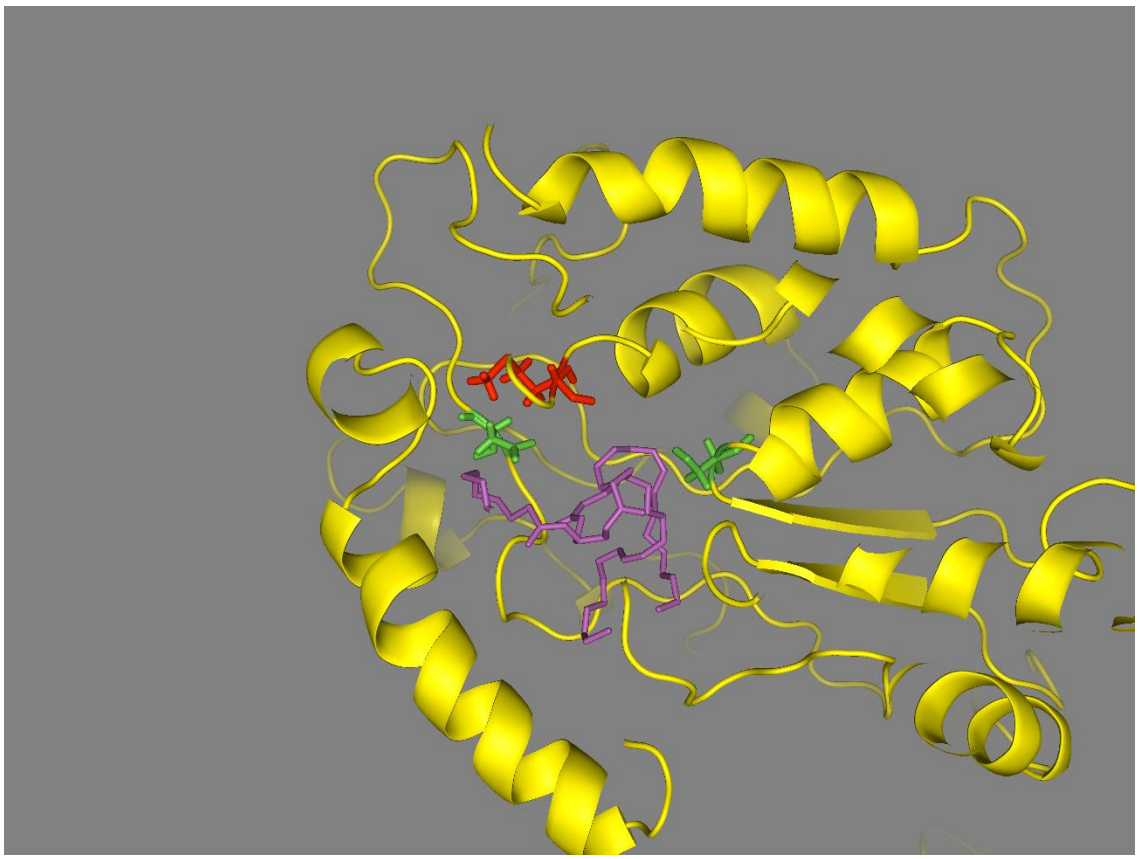


Figure 5.62 Highlighting docking of inactive enzymatic form in PNPLA3 I148M variant

5.6.6 Multiple one nanosecond repeats

Short repeats of the simulated systems were performed in order to test the reliability of the observed conformational shifts. This was used in particular to investigate the local conformation of the active site, because stable conformations in these regions were reached almost instantly in each simulation.

Across these simulations, we observed that the conformations which had initially been seen in the previous simulations were only present 40% of the time. The other 60% of the time, the structure appears similar to that of the variant protein. Interestingly however, the active conformation was never observed within the variant protein simulations.

These results only lend further support that the conformations observed by PNPLA3 are not irreversible conformations and may be representative of dynamic equilibria which could putatively be observed *in vivo*. The potential for the wild type protein to more readily adopt the active conformation would then lend a large catalytic advantage, even though the protein is not active 100% of the time. These equilibria would likely be further altered by the interactions of any binding partners or lipid droplets within the cell system.

To further explore the nature of the conformational differences between the wild type and variant PNPLA3 variants a 148A, alanine variant, was also created and simulated. Alanine was selected as it maintains the same hydrophobic nature, without having an observable long chain impact on the local environment.

Simulations of the alanine variant, were consistent with the findings of the wild type protein, showing that the impact on the local conformation within the active site is indeed dependent on the gain of a methionine, rather than a loss of the isoleucine in position 148. Although this does not rule out changes caused by any other amino acid variations which may be present.

Statistical comparison of these systems showed that the I148M variant protein, indeed exhibits a change in the active site conformation in which the methionine residue is positioned closer to the active site residues and is actually positioned between the residues, causing an increase in the distance between the catalytic residues themselves. This supports the observations shown in the 100ns simulations, and the hypothesis that the I148M variant protein effects catalytic activity due to a disruption of the active conformation of PNPLA3.

5.6.7 Ubiquitination

5.6.7.1 Loss of ubiquitination as a pathogenic mechanism

It has recently come to light, that the I148M variant of PNPLA3 has a vastly different ubiquitination profile when compared with the wild type protein. Namely, the I148M variant undergoes significantly less ubiquitination.[467] This raises a key question relating to the importance of this process in the pathogenic role of PNPLA3 I148M.

Ubiquitination is mainly used within the cell to mark a protein for degradation and therefore the obvious impact of a reduction in ubiquitination would be the reduced turnover of PNPLA3 protein; This could lead to increased quantities of PNPLA3 in the cell and on the surface of lipid droplets. This is supported by recent evidence showing accumulation of I148M on the surface of the droplets.[169]

We know that the ubiquitination degradation pathway is extremely important when it comes to lipid droplet homeostasis and the overall lipid metabolism within the cell. In particular, PNPLA2 (ATGL) a homologue of PNPLA3 and the rate limiting lipolytic enzyme within the cell, is suggested to undergo tight regulation through carefully controlled ubiquitination.[468]

It stands to reason, as a homologue of PNPLA2, PNPLA3 may also undergo similar regulatory mechanisms.

Another distant homologue of PNPLA3, lipid droplet associated hydrolase (LDAH) has recently been shown to have many similar behavioural patterns to PNPLA3. Despite being upregulated in response to SREBP, its primary function is as lipase. However, LDAH does cause a lipogenic response within the cell.

Overexpression of LDAH also facilitates an increase in lipid droplet size and accumulation of TAG, however is not been implicated as a problematic function *in vivo*. Additionally, like PNPLA3 removal of the protein *in vivo* has no physiological effects.[468]

It has been noted that expression of this protein actually increased the ubiquitination of ATGL, thereby reducing the functionality and stability of the protein, as well as masking increasing ATGL upon increasing intracellular TAG stores. While the true mechanism of this protein also remains unclear, the parallels between what is observed between PNPLA3 are remarkable.

Unfortunately, ATGL ubiquitination was not measured in the study of PNPLA3, and ubiquitination of LDAH was also not tracked, although levels of ATGL did not imply this was impacted by either PNPLA3 variant. It is feasible to expect, that the PNPLA3 variant may impact another lipase similarly to LDAH, and therefore impact the overall metabolism within the cell. For example, the inability to bind ubiquitin, may leave more available ubiquitin to bind ATGL, providing a mechanism leading to increased TAG storage.

Further to this ATGL has been shown to be regulated by a large host of factors within the cell, including Fsp27 and CIDEA.[469,470] Despite no current experimental explanation on the impact on ATGL, there are many ways by which PNPLA3 could perform regulatory functions with ATGL.

The similar patatin domain architecture may mean that PNPLA3 is able to interact with any number of ATGL binding partners, and thereby modulate ATGL activity indirectly. We observed a dramatic shift in architecture based on the I148M variation in our simulations, beyond effecting the active site, this could easily be thought to change affinity for other proteins modulating this effect.

We have seen that binding of G0/G1 Switch Gene 2 (G0S2) protein to ATGL not only inhibits ATGL but also stabilises the protein from ubiquitination, so the loss of ubiquitination could be a side effect of additional binding partners in the lipidome.[471]

In an early study of PNPLA3, it was seen that the I148M variant was observed with increased levels of CGI-58 on the surface of droplets, only lending more support to this concept, although this may be simple correlation.[174]

5.6.7.2 Mechanism underlying loss of ubiquitination

Despite observing a correlation between decreased ubiquitination of the I148M variant, no mechanism for this change was presented within the research.

After simulation, the most confident models of both wild type and I148M variant PNPLA3 displayed drastically different positions of the potential sites of ubiquitination on our models. Notably, in the I148M variant potential lysines were far more often facing into the core of the protein. Additionally, the C-terminal domain in the I148M variant was more ordered, offering less random coil which is amenable to protein binding. Taken together this suggests that the alternate conformation alone may be enough to cause a reduction in ubiquitination of the I148M variant.

In addition to this, one of the potential sites of ubiquitination is also a commonly observed SNP at residue 434. This SNP, E434K, has also been associated with liver injury in a recent study by Donati *et al.*[472]

The study found a more frequent co-inheritance of the E variant with the M148 variant. This could imply that the previous findings showing increased ubiquitination is partly due to the co-inheritance of this variant. It is imperative that the link between these associations is further explored.

K434 again lies on the surface of the wild type protein, and is more shielded in the variant, suggesting that even without variation may be less amenable to ubiquitination. In our simulations, only the K variant was simulated which is proposed to be the non-risk allele.

Since this study observed a simultaneous decrease in mRNA for the protein, it may be that this variant is in linkage disequilibrium with another non-coding variant in a promotor region, which was not investigated in the study and is the root cause in changing protein levels. While another study did not find the connection between the disease, however this has yet to be extensively observed across populations.[473]

5.6.7.3 Ubiquitin as a protein activator

As discussed in the previous chapter, the bacterial homologue ExoU is activated upon binding of mammalian ubiquitin.[415] It is possible that ubiquitin acts as a binding partner and activator or PNPLA3 *in vivo*.

This provides a hypothetical link between the reduction of observed ubiquitination and a loss of catalytic activity. Because there is still some degree of ubiquitination, this would not have to be a deleterious loss of catalytic activity but would be a permissive mutation which was linked to the dramatic decrease in activity.

A thorough investigation into the binding partners of PNPLA3 is a definite key step moving forward into the understanding of this crucial enzyme.

5.6.8 Limitations

While the results of these simulations are extremely promising and provide novel insight into the putative dynamics of PNPLA3 and structural implications of the I148M variant, the results must be interpreted with caution, as with any study of protein dynamics there are several limitations to this study.

The starting point for simulation is a hypothetical structural model, generated based on low levels of homology. This means that while the backbone architecture may be roughly correct, there is an intrinsically high probability the protein structure is not accurate at an atomistic level.

The active site was simulated with a molecular mechanics approach, which while accurate in general, will overlook bond polarisation within the active site residues. This simplification could lead to results which would not be seen *in vivo*. A QM/MM simulation should be performed with the quantum region defined around the active site to gain a better understanding of the unique interaction of these key residues.

The protein has been simulated as a single molecule in isolation. This ignores the potential impact of multimerization, which has been suggested in several studies, and of course the impact of interaction with lipid droplets.

The simulation may still be too short, we witness stabilisation across time frame of the simulation, which in particular modelled the shift in C-terminal domain positioning. Toward the

final 20ns the simulation the two domains appeared to have equilibrated and interdomain drift minimised.

As expected, we do observe low levels of energy drift across the timeframe of the simulation. While energy drift is to some extent inevitable during constant energy simulations due to round off errors in the energy at each integration step of the run, this could cause the simulations to lose accuracy over time and appear superficially stable toward the end of the simulation. This could be minimised further by increasing the electrostatic cut-offs “cut” and “rgbmax” to higher settings. The current selection is based off commonly recommended values for systems of this size and is a balance between functional running time and accuracy. Despite the moderate energy drift, we do not see problematic temperature shift in the simulation, which suggests that the impact on the energy of the system was minimal, and acceptable for interpretation.

In addition to this, due to time constraints we had limited potential for longer time scale repeats. While several repeats of the system on a shorter 1ns timescale were performed, they are limited in their power due to the fact we cannot be sure whether the system would remain stable under longer simulation.

Finally, we do not have any direct experimental evidence to support the model, although it is consistent with the functional experimental data to date.

5.7 Conclusion

Key findings:

Support for previous findings:

- Loss of lipase enzyme activity caused by the I148M substitution was supported by the simulation result.
- The beginning of the simulation saw a transition of increasing distances between the catalytic residues.

Novel findings:

- Stable simulations of full length PNPLA3 were performed for over 100ns.
- The patatin domain up to residue 239 moved as a single domain.
- The previous hypothesis of enzyme inhibition from steric hindrance to the active site from 148M was not supported.
- A novel mechanism of inhibition of enzyme activity was predicted in which the I148M variant impacts the conformation of the active site, leading to increased distance between the catalytic residues.
- A range of additional changes to the conformation of the whole protein were observed, suggesting possible APO and HOLO conformations.
- Conformational changes caused by I148M variation revealed a potential impact to the availability of residues for ubiquitination.

In summary, the first simulations of a full length PNPLA3 have been successfully performed. These simulations support the multiple domain hypothesis of PNPLA3, in which the patatin domain predominantly has lipase activity, which is perturbed by the I148M substitution.

The previous hypothesis that the I148M variant caused a loss of lipase activity through steric hindrance, reducing access to the active site was not supported by the current findings. In fact, the I148M variant was observed to have a larger active site binding pocket and show docking with greater affinity for nearly all known PNPLA3 substrates.

The I148M variant was observed to cause destabilisation of the active conformation of PNPLA3, in turn causing a conformational shift to an alternate conformation. This shift is predicted to cause a loss of lipase activity by separating the spatial localisation of the two catalytic residues. It is hypothesised that this shift occurs due to a loss of β-strand character of the strand preceding residue 148, which causes a reduced chemical interaction with strands 3 and 4 in the structure.

These findings shed new light on the impact of the I148M variant on the structure and function of PNPLA3 and may serve to assist in the assessment of PNPLA3 as a therapeutic target.

Chapter 6

General discussion

"Endings and beginnings are merely paired facets of an imagined stone curtain, behind which a plethora of opportunities await."

Ged Thompson

6.1 Context

Liver disease is currently of great international concern, not only because of the huge burden of disease and disability, but also because of increasing death rates attributable to the disease.

The two most significant risk factors known to influence the development of liver disease are alcohol and obesity. With trends of alcohol consumption[474] and the prevalence of obesity worldwide[475] rising significantly over the last 40 years, it is likely incidence of liver disease will continue to increase further (Figures 6.1 to 6.3).

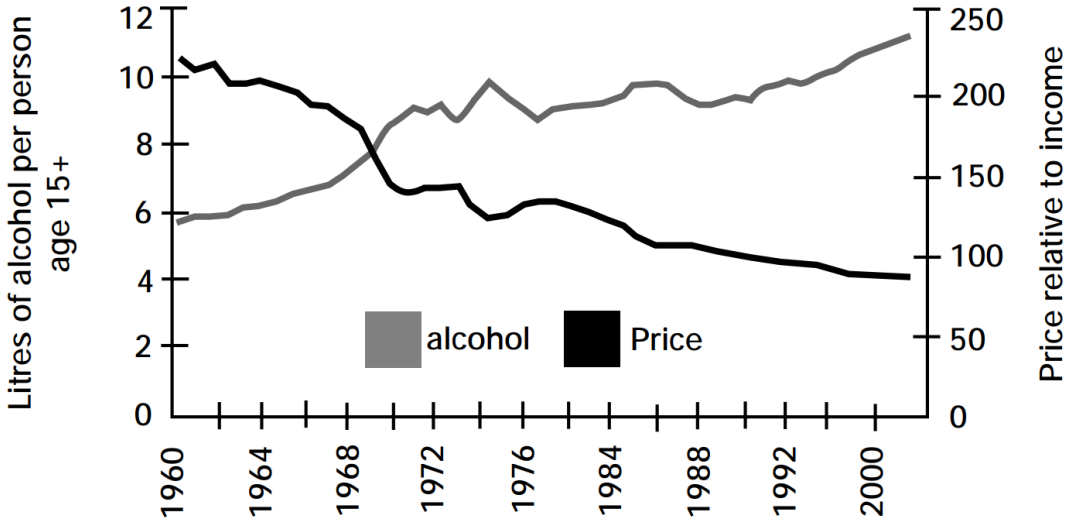


Figure 6.1 Consumption of alcohol in the UK from 1960-2002

(Adapted from Sheron N 2004)[474]

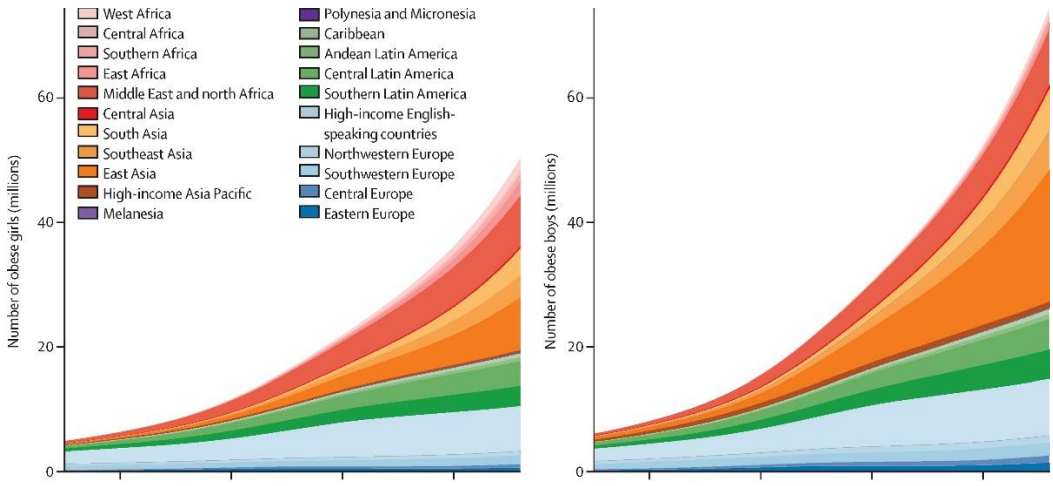


Figure 6.2 Number of children and adolescents (aged 5–19 years) with obesity by region

(Adapted from the NCD Risk Factor Collaboration 2017).[475]

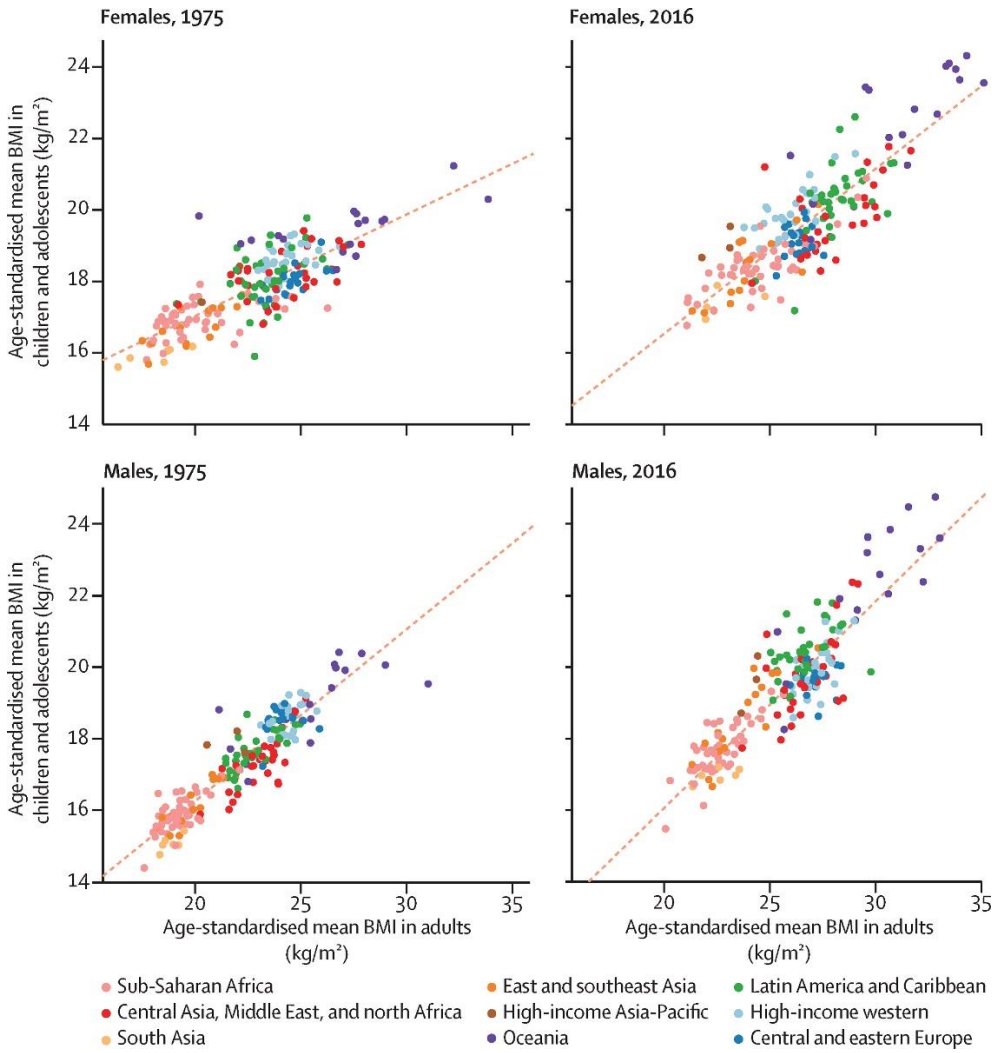


Figure 6.3 Comparison of age-standardised mean body-mass index (BMI) in children and adolescents (5–19 years) and in adults

Each point shows one country, the dotted line shows the linear association between the two outcomes (Adapted from the NCD Risk Factor Collaboration 2017).[475]

The rs738409 variant of PNPLA3, encoding an isoleucine to methionine variation at residue 148, has been associated independently with higher risk of severity of all stages of NAFLD and ALD, liver injury in Wilsons disease, chronic hepatitis C and hepatocellular carcinoma.[3]

The strength of associations and prevalence of the gene in society have led to proposals to PNPLA3-associated steatohepatitis (PASH) be used as a unique gene-based liver disease entity.[476]

Despite clinical interest in the PNPLA3 protein, the activity of the protein and the effect of the I148M variation, remains unknown. However, several hypotheses have been presented, which attempt to link the I148M variant with the pathogenesis of liver disease:

(i) PNPLA3 is lipolytic in its wild type form and the I148M variant is associated with lose of that activity resulting in fat accumulation in hepatocytes.[169]

(ii) PNPLA3 is lipogenic in its wild type form and the I148M variant is associated with an increase in function leading again to fat accumulation in hepatocytes.[166]

(iii) A functional change *per se* is not the cause of steatosis, but rather the accumulation of inactive PNPLA3 on lipid droplets causes accumulation of TG by restricting access to the lipid droplet or sequestering a factor required for hydrolysis.[174]

(iv) The I148M causes an accumulation on lipid droplets due to reduced ubiquitination and leads to increased net activity. This impacts the homeostatic lipid balance has complex downstream effects leading to a pathogenic response.[173]

Each of these hypothesis presents a putative disease mechanism linking the I148M variant with live disease, however, further characterisation of PNPLA3 was needed to elucidate the true nature of the protein and evaluate the protein as a drug target.

6.2 Summary of findings

This thesis concerns itself with trying to improve the understanding of the protein PNPLA3 and the role that the I148M variant may play in the pathogenesis of liver disease. In doing so, a range of techniques have been applied to the investigation and a summary of novel findings is presented below to facilitate broader discussion.

6.2.1 Chapter 2

In this chapter, an *in silico* investigation was undertaken using the primary sequence of PNPLA3., A range of distant protein homologues were identified using a conserved domain-based search and PNPLA3 was better contextualised, both within the background of related PNPLAs, and the broader context of the Cpla2_and_patatin superfamily.

Notably, PNPLA5 was determined to be the closest structural homologue to PNPLA3, rather than PNPLA2. The *E. coli* protein ExoU was identified as the closest homologue with a known structure, which was leveraged in model creation in chapter 4.

While it is interesting to note that both hydrolase and acyltransferase activities were predicted for PNPLA3, unfortunately this does not assist in better understanding the protein and diverse experimental findings.

A key finding of this chapter was to see that PNPLA3 is not predicted to be amenable to *in vitro* expression, purification or crystallisation. While this is going to be biased by the current database of known structures, the fact so few proteins within this superfamily have known structures, supports the fact this may be a challenge.

By assessing regions of hydrophobicity, putative domain boundaries and other protein properties, a range of potential truncated clones were predicted to facilitate further experimental investigation into the protein.

6.2.2 Chapter 3

A range of both human and murine PNPLA3 recombinant protein homologues were expressed and purified from *E. coli*. Both human and murine homologues were shown to behave similarly *in vitro* and form high molecular weight multimers of around 670kDa which could not be separated with non-destructive techniques.

While protein expression was achieved, problems were encountered with degradation and insolubility, which remain unresolved, as predicted in the previous chapter. A lipase activity assay was performed which showed the human protein had stronger lipase activity than the murine, supporting previous experimental findings.[166]

A significant contaminant, ArnA, was identified even after Ni-affinity chromatography, which co-purified with PNPLA3. Multiple characteristic similarities with PNPLA3 made this a difficult contaminant to remove and identify, showing extreme caution must be taken in purification from an *E. coli* expression host.

6.2.3 Chapter 4

PNPLA3 has low levels of homology with any proteins of known structure. This low homology means that generating confident models is extremely challenging. In this chapter, nine models of PNPLA3 were generated using a range of templates and both homology modelling and

threading. Similarities in the structural cores of the patatin domain between these models dramatically increases confidence of the structure produced.

The models support the fact that the classic hydrolase fold is maintained as was predicted previously.[151] Using a range of approaches allowed the prediction of models with higher confidence than the previously published model, and the generation of the first full length model of PNPLA3.

Crucially, within this model it became clear that two additional helices form an integral part of the patatin domain and leads us to redefine the patatin domain from residues 1-239.

Across the models both I148 and D166 lie on the same flexible loop region near the active site pocket. Multiple conformations adopted by different models suggest a potential APO and HOLO conformational shift, by which the predominant shift is performed by this flexible loop.

6.2.4 Chapter 5

In this chapter, the first full length models of both the wt and I148M variant of PNPLA3 underwent 100ns simulation and provides the seminal findings of this thesis. The simulations were able to further refine the structural models of PNPLA3 and help to characterise the protein as a three-domain protein consisting of the patatin domain, right lobe domain and C-terminal domain.

The additional residues previously left out of modelling studies were shown to be an integral and stable part of the catalytic pocket, and the first active site pocket was defined, whereby two joining tunnels to the catalytic residues facilitate binding of a broad range of long lipid moieties through the protein.

Remarkably, despite isoleucine and methionine sharing similar amino acid properties, the I148M variant was shown to have a significant impact on the local and global structure of the protein.

Namely, there was a local loss of β character in the vicinity of residue 148, which caused a loss of interaction between β strands 3, 4 and 5 which are all absent in the variant simulation. This in turn is followed by a compression of the right lobe domain of the protein, and a transitioning across the face of the patatin domain.

The most important part of this change was the destabilisation of the conformation with an active site whereby the catalytic residues S47 and D166 now move above 10A apart.

This also drastically altered the active site pocket, by opening a much wider pocket, which had an extremely altered propensity for ligand binding. Indeed, simple docking appeared to show stronger affinity for binding in the I148M variant.

The observed structural change provides evidence for a novel mechanism by which the I148M variant may cause a loss of activity, through destabilising the adjacent catalytic residues, causing a rotation effectively disabling the oxyanion hole and altering the right lobe domain to an extent likely to affect the binding of any partner proteins.

6.3 The implications of this research

These findings have increased our understanding of the role of the I148M PNPLA3 variant in liver injury, by providing a novel mechanism by which the I148M variant may affect the function and structure of the protein.

It now seems highly likely that the overall impact of the I148M variant is due to a loss of activity of the protein, which narrows down the theories of pathogenesis of disease.

Whether it is the observed loss of activity, or the structural change which is the primary issue causing increased intracellular lipids is not yet clear. However, this study suggests that there may be potential to restore native functioning of PNPLA3 by lowering the energy of the active conformation with supplementary ligands.

Furthermore, improved understanding of domain boundaries can aid in future experimental design. Additionally, the similarities between homologous proteins PNPLA5, LPAAT and ExoU have been highlighted. The human proteins in particular share many characteristics of PNPLA3, and future studies into these proteins may provide insight into PNPLA3, whether its putative functionality, or improved purification protocols.

On another level, PNPLA3 has proved to be a challenging protein to work with experimentally. By using a range of molecular modelling techniques and carefully applying molecular dynamic simulations this thesis has not only been able to identify a possible molecular mechanism for the influence of the I148M variation, but also provide an example how *in silico* investigations can provide a mechanistic alternative for structural investigations even in cases where homology is relatively low. This study could act as a framework for similar studies into key proteins which are not able to be purified within *in vitro* conditions.

6.4 Future work

6.4.1 Understanding the function of PNPLA3

As you can see, despite the large quantity of work which has been undertaken on PNPLA3, the role of the I148M variant in the pathogenesis of the disease remains unclear. There is still much to be learned before we could use this in a clinical setting above and beyond increased screening for patients deemed at risk.

Increased understanding of the precise pathogenic mechanism of PNPLA3 will need to be assessed to facilitate downstream targeting of other pathways to mediate the pathogenic effect. This requires a deeper knowledge of the underlying mechanisms of disease.

It is increasingly likely that PNPLA3 is a broadly acting hydrolase, that can perform a wide range of activities *in vivo*. A combination of the multimeric appearance upon *in vitro* purification and the typical characteristics of homologous proteins such as ATGL and ExoU, imply that PNPLA3 is highly likely to operate in conjunction with other protein binding partners.

It seems feasible therefore, that the diverse results observed to date are due to the influence of binding partners on the functional capacity of PNPLA3. It has already been observed that the I148M variation effects ubiquitination (potentially due to conformational changes), so it stands to reason may impact other putative binding partners.

At this stage, where extensive phenotypic characterisation of the protein in a range of cell types has already been performed, the next logical step would be to determine protein-protein interaction partners. This could be investigated using a quantitative affinity purification mass spectrometry to quantitively assess native binding partners between the two variants.[477]

Understanding potential binding partners could provide a vital link in improving purification of the protein, comparative *in vitro* functional assays, facilitate experimental structural determination and further improve dynamic simulation of the protein behaviour. In particular, investigation into the impact of ubiquitin on the function of PNPLA3 should be investigated.

6.4.2 Recent advances

Indeed, since the completion of this work, PNPLA3 has now been shown to interact with CGI-58 as a binding partner *in vivo*.[478] This is a powerful regulator of PNPLA2, and could play an important role on the solubility and stability of PNPLA3. Co-expression with CGI-58, may provide

soluble protein expression which could facilitate structural investigation and more accurate enzyme kinetics.

The fact PNPLA3 is interacting with CGI-58 is extremely interesting and initial evidence suggests that CGI-58 does indeed promote PNPLA3 activity in the WT. Additionally, PNPLA3 is able to interact and inhibit the ATGL-CGI-58 complex, while not reducing the quantity of ATGL-CGI-58 on the surface of lipid droplets. This suggests a complex role, perhaps whereby incomplete binding with the complex causes non-specific aggregation and inhibition; which would be supported by the tendency to aggregate demonstrated by patatin like proteins.

A further study also suggests that the I148M variant has a reduced ability to be ubiquitinated, which reduced the speed of clearance of PNPLA3 and causes a PNPLA3 to accumulate on the surface of lipid droplets. It is hypothesised that it is through the inhibition of ATGL, that PNPLA3 I148M is able to increase the size of lipid droplets, and an accumulation of the I148M variant is the pathogenic factor.[479] The findings produced in this thesis fully support the hypotheses presented in these papers that PNPLA3 I148M would exhibit a loss of lipase activity and experience reduced ubiquitination. As a catalytically inactive PNPLA3 also results in the same loss, this does not rule out the potential that ubiquitin is an activator and catalysis is required for release from the surface of the lipid droplet in order to be cleared from the cell.

Based on this additional information, I propose a hypothesis that PNPLA3 is a lipogenic protein, which operates both to rapidly inhibit PNPLA2 under carbohydrate stress, and perform a more complex role in maintaining lipid homeostasis within the cell.

6.4.3 Developing treatments for liver disease

While there are currently a range of drugs which are being repurposed to attempt to interfere with the pathogenic processes in the disease, these are non-specific, of moderate effect and come with a range of side effects. This means there is an urgent need for the development of new therapies which could be used to target the development and progression of liver disease.

6.4.3.1 PNPLA3 as a drug target

Since the initial association studies between PNPLA3 I148M and liver injury, PNPLA3 has been an obvious candidate for drug therapies aiming to either reduce the gene related risk inferred by the methionine genotype and potentially even treat non-risk carriers.

In order to facilitate targeted drug discovery, full characterisation of the protein is required; in particular, the function of the protein and impact of the I148M variant on activity must be known.

Developing drugs which increase enzyme activity is a much more challenging task, compared to those which inhibit activity. As evidence has grown that PNPLA3 risk variant appears most likely to provide a loss of function, this makes targeting PNPLA3 a more challenging prospect, as inhibition of the enzyme will not be able to mediate the pathogenic effects.

The authors of recent studies suggest we have sufficient evidence to consider knockdown of PNPLA3 as a therapeutic approach. While this is a valid option, supported by a lack of symptoms in mouse knockout models, this comes with an additional long-term risk of not having the PNPLA3 protein; if it does indeed play a significant role.

I believe we have several alternate options available to us, based on the information we have to date. Firstly, small molecules have already been observed to inhibit the interaction between CGI-58 and another inhibitor Plin1 within the cell, this suggests a similar compound may be able to have the same therapeutic effect without removing PNPLA3.[480] It is worth noting, that since the wt PNPLA3 also has some inhibitory effects on PNPLA2, this treatment may still prove effective in patients without the risk variant.

Secondly, while the I148M variant does have reduced function, there is experimental evidence that some activity remains. Simulations of both variants also suggest that the protein may have increased stability in an alternate conformation but should still be able to form an energetically less stable, active conformation.

Under these unique circumstances, it remains theoretically possible to design ligands which may stabilise the catalytically active structure, causing a conformational shift and thereby restore the native activity of the wild type enzyme.

Because of difficulties purifying large quantities of PNPLA3, investigation and development of this type of therapeutic would likely require extensive research by trial and error using *ex vivo* cell lines. The three dimensional models presented in this thesis could be used to assist in initial pharmacophore prediction to reduce both the cost and time span of this research.[481]

Of course, other therapeutic strategies should also be considered, which may reduce binding of PNPLA3 to lipid droplets, or reduce overall quantities available in the cell through gene expression.[482]

6.4.3.2 Alternate drug targets

An alternative approach one may take is to look at other proteins within the regulatory pathway. Since PNPLA3 is causing a loss of function, the question remains as to how this loss of function impacts disease. Understanding this would provide a basis to plan a strategy to modify other appropriate pathways to adjust for PNPLA3 loss of function as discussed above.

For example, increasing levels of a PNPLA3 end product may be possible via inhibition of other downstream enzymes. While this method requires a far greater understanding of the specific role of PNPLA3 within the pathogenesis of the disease, as well as additional supplementary enzymes within the lipid metabolism, it opens the door to a potentially broad array of additional drug targets.

If the inhibition of PNPLA2 is indeed the dominant cause of the PNPLA3 I148M induced steatosis, interfering with another PNPLA2 inhibitor, for example Plin1 or Plin5 may be enough to combat the impact of the I148M variant.

Overall, the last few years have provided much greater insight into the elusive PNPLA3 protein, and we are gradually getting closer to understanding the role the I148M variant plays in liver disease. Hopefully this will lead to the creation of effective therapeutics for the disease in the near future.

References

1 Williams R, Aspinall R, Bellis M *et al.* Addressing liver disease in the UK: a blueprint for attaining excellence in health care and reducing premature mortality from lifestyle issues of excess consumption of alcohol, obesity, and viral hepatitis. *Lancet* 2014; **384**: 1953–97.

2 Spengler EK, Loomba R. Recommendations for diagnosis, referral for liver biopsy, and treatment of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Mayo Clin Proc* 2015; **90**: 1233–46.

3 Bruschi FV, Tardelli M, Claudel T, Trauner M. PNPLA3 expression and its impact on the liver: current perspectives. *Hepat Med* 2017; **9**: 55–66.

4 Xu R, Tao A, Zhang S, Deng Y, Chen G. Association between patatin-like phospholipase domain containing 3 gene (PNPLA3) polymorphisms and nonalcoholic fatty liver disease: a HuGE review and meta-analysis. *Sci Rep* 2015; **5**: 9284.

5 Salameh H, Raff E, Erwin A *et al.* PNPLA3 gene polymorphism is associated with predisposition to and severity of alcoholic liver disease. *Am J Gastroenterol* 2015; **110**: 846–56.

6 Huang Y, He S, Li JZ *et al.* A feed-forward loop amplifies nutritional regulation of PNPLA3. *Proc Natl Acad Sci U S A* 2010; **107**: 7892–7.

7 Liu Y-M, Moldes M, Bastard J-P *et al.* Adiponutrin: A new gene regulated by energy balance in human adipose tissue. *J Clin Endocrinol Metab* 2004; **89**: 2684–9.

8 Grisham JW. Organizational principles of the liver. In: *The Liver*. John Wiley & Sons, Ltd: Chichester, UK, 2009, pp 1–15.

9 Lefkowitch JH. Anatomy and Function. In: *Sherlock's Diseases of the Liver and Biliary System*. Wiley-Blackwell: Oxford, UK, 2011, pp 1–19.

10 Arias I, Wolkoff A, Boyer J, Fausto N, Cohen D. *The Liver : Biology and Pathobiology.* Wiley, 2011.

11 Burgess SC, He T, Yan Z *et al.* Cytosolic Phosphoenolpyruvate Carboxykinase Does Not Solely Control the Rate of Hepatic Gluconeogenesis in the Intact Mouse Liver. *Cell Metab* 2007; **5**: 313–320.

12 Cherrington AD. Banting Lecture 1997. Control of glucose uptake and release by the liver in vivo. *Diabetes* 1999; **48**: 1198–214.

13 Lewis GF, Vranic M, Harley P, Giacca A. Fatty acids mediate the acute extrahepatic effects of insulin on hepatic glucose production in humans. *Diabetes* 1997; **46**: 1111–9.

14 Lin HV, Accili D. Hormonal Regulation of Hepatic Glucose Production in Health and Disease. *Cell Metab* 2011; **14**: 9–19.

15 Obici S, Rossetti L. Minireview: Nutrient Sensing and the Regulation of Insulin Action and Energy Balance. *Endocrinology* 2003; **144**: 5172–5178.

16 Petersen KF, Laurent D, Rothman DL, Cline GW, Shulman GI. Mechanism by which glucose and insulin inhibit net hepatic glycogenolysis in humans. *J Clin Invest* 1998; **101**: 1203–1209.

17 Ramnanan CJ, Edgerton DS, Rivera N *et al.* Molecular Characterization of Insulin-Mediated Suppression of Hepatic Glucose Production In Vivo. *Diabetes* 2010; **59**: 1302–1311.

18 Roach PJ. Glycogen and its metabolism. *Curr Mol Med* 2002; **2**: 101–20.

19 Schwer B, Verdin E. Conserved Metabolic Regulatory Functions of Sirtuins. *Cell Metab* 2008; **7**: 104–112.

20 Zhang BB, Zhou G, Li C. AMPK: An Emerging Drug Target for Diabetes and the Metabolic Syndrome. *Cell Metab* 2009; **9**: 407–416.

21 Metabolic Syndrome ePoster. http://www.nature.com/nm/e-poster/Liver-Organ-Normal.html (accessed 25 Aug2017).

22 Reisner H. *Pathology: A Modern Case Study*. McGraw-Hill Education, 2014https://books.google.co.uk/books/about/Pathology_A_Modern_Case_Study.html (accessed 25 Aug2017).

23 Associate Degree Nursing Physiology Review. http://www.austincc.edu/apreview/PhysText/Digestive.html (accessed 25 Aug2017).

24 Seo HJ, Nam SH, Im H-J *et al.* Rapid Hepatobiliary Excretion of Micelle-Encapsulated/Radiolabeled Upconverting Nanoparticles as an Integrated Form. *Sci Rep* 2015; **5**: 15685.

25 Blouin A, Bolender RP, Weibel ER. Distribution of organelles and membranes between hepatocytes and nonhepatocytes in the rat liver parenchyma. A stereological study. *J Cell Biol* 1977; **72**: 441–55.

26 Weibel ER, Stäubli W, Gnägi HR, Hess FA. Correlated morphometric and biochemical studies on the liver cell. I. Morphometric model, stereologic methods, and normal morphometric data for rat liver. *J Cell Biol* 1969; **42**: 68–91.

27 Wisse E, De Zanger RB, Charels K, Van Der Smissen P, McCuskey RS. The liver sieve: considerations concerning the structure and function of endothelial fenestrae, the sinusoidal wall and the space of Disse. *Hepatology*; **5**: 683–92.

28 Wack K, Ross MA, Zegarra V, Sysko LR, Watkins SC, Stolz DB. Sinusoidal ultrastructure evaluated during the revascularization of regenerating rat liver. *Hepatology* 2001; **33**: 363–378.

29 Arthur MJ, Mann DA, Iredale JP. Tissue inhibitors of metalloproteinases, hepatic stellate cells and liver fibrosis. *J Gastroenterol Hepatol* 1998; **13 Suppl**: S33-8.

30 Parker GA, Picut CA. Liver Immunobiology. *Toxicol Pathol* 2005; **33**: 52–62.

31 Smedsrød B, Le Couteur D, Ikejima K *et al.* Hepatic sinusoidal cells in health and disease: update from the 14th International Symposium. *Liver Int* 2009; **29**: 490–501.

32 Selden C, Khalil M, Hodgson HJ. What keeps hepatocytes on the straight and narrow? Maintaining differentiated function in the liver. *Gut* 1999; **44**: 443–6.

33 Benedetti A, Bassotti C, Rapino K, Marucci L, Jezequel AM. A morphometric study of the epithelium lining the rat intrahepatic biliary tree. *J Hepatol* 1996; **24**: 335–42.

34 Sayiner M, Koenig A, Henry L, Younossi ZM. Epidemiology of Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis in the United States and the Rest of the World. *Clin Liver Dis* 2016; **20**: 205–214.

35 Kochanek KD, Murphy SL, Xu J, Tejada-Vera B. National Vital Statisitcs Reports, Volume 65, Number 4, (06/30/2016). 2014.https://www.cdc.gov/nchs/data/nvsr/nvsr65/nvsr65_04.pdf (accessed 26 Mar2018).

36 U.S. Department of Health and Human Services C for DC and P. Summary Health Statistics: National Health Interview Survey. 2015.https://ftp.cdc.gov/pub/Health_Statistics/NCHS/NHIS/SHS/2015_SHS_Table_A-4.pdf (accessed 26 Mar2018).

37 Schuppan D, Schattenberg JM. Non-alcoholic steatohepatitis: pathogenesis and novel therapeutic approaches. *J Gastroenterol Hepatol* 2013; **28 Suppl 1**: 68–76.

38 Pless G. Artificial and bioartificial liver support. *Organogenesis* 2007; **3**: 20–4.

39 Liu JP, Gluud LL, Als-Nielsen B, Gluud C. Artificial and bioartificial support systems for liver failure. In: Liu JP (ed). *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd: Chichester, UK, 2004, p CD003628.

40 Kjaergard LL, Liu J, Als-Nielsen B, Gluud C. Artificial and bioartificial support systems for acute and acute-on-chronic liver failure: a systematic review. *JAMA* 2003; **289**: 217–22.

41 Sakhuja P. Pathology of alcoholic liver disease, can it be differentiated from nonalcoholic steatohepatitis? *World J Gastroenterol* 2014; **20**: 16474–9.

42 Crabb DW. Pathogenesis of alcoholic liver disease: newer mechanisms of injury. *Keio J Med* 1999; **48**: 184–8.

43 Edmondson HA, Peters RL, Frankel HH, Borowsky S. The early stage of liver injury in the alcoholic. *Medicine (Baltimore)* 1967; **46**: 119–29.

44 Bellentani S, Tiribelli C. The spectrum of liver disease in the general population: lesson from the Dionysos study. *J Hepatol* 2001; **35**: 531–7.

45 Sørensen TI, Orholm M, Bentsen KD, Høybye G, Eghøje K, Christoffersen P. Prospective evaluation of alcohol abuse and alcoholic liver injury in men as predictors of development of cirrhosis. *Lancet* 1984; **2**: 241–4.

46 Mathurin P, Hadengue A, Bataller R *et al.* EASL clinical practical guidelines: management of alcoholic liver disease. *J Hepatol* 2012; **57**: 399–420.

47 Kondo F. Histological features of early hepatocellular carcinomas and their developmental process: for daily practical clinical application : Hepatocellular carcinoma. *Hepatol Int* 2009; **3**: 283–93.

48 Uchida T, Kao H, Quispe-Sjogren M, Peters RL. Alcoholic foamy degeneration--a pattern of acute alcoholic injury of the liver. *Gastroenterology* 1983; **84**: 683–92.

49 Burt AD, Portmann B, Ferrell LD, MacSween RNM. *MacSween's pathology of the liver.* Churchill Livingstone/Elsevier, 2012.

50 Goodman ZD, Ishak KG. Occlusive venous lesions in alcoholic liver disease. A study of 200 cases. *Gastroenterology* 1982; **83**: 786–96.

51 Sleisenger MH, Fordtran JS, Feldman M, Friedman LS (Lawrence S, Brandt LJ. *Sleisenger and Fordtran's gastrointestinal and liver disease : pathophysiology, diagnosis, management. Vol. 1*. Saunders/Elsevier, 2010.

52 Heidelbaugh JJ, Sherbondy M. Cirrhosis and Chronic Liver Failure: Part II. Complications and Treatment. 2006; **74**.www.aafp.org/afp (accessed 26 Mar2018).

53 Johnston DE. Special considerations in interpreting liver function tests. *Am Fam Physician* 1999; **59**: 2223–30.

54 Kim WR. Clinical implications of short-term variability in liver function test results. *Gastroenterology* 2008; **135**: 1010-1-2.

55 Child C, Murray-lyon I, Dawson J. The liver and portal hypertension: ChildCL, ed. *Surg Portal Hypertens* 1964; : 50–8.

56 Kamath PS, Wiesner RH, Malinchoc M *et al.* A model to predict survival in patients with end-stage liver disease. *Hepatology* 2001; **33**: 464–70.

57 van Hoek B. Non-alcoholic fatty liver disease: a brief review. *Scand J Gastroenterol Suppl* 2004; : 56–9.

58 Tilg H, Moschen AR. Evolution of inflammation in nonalcoholic fatty liver disease: The multiple parallel hits hypothesis. *Hepatology* 2010; **52**: 1836–1846.

59 Yu J, Marsh S, Hu J, Feng W, Wu C. The Pathogenesis of Nonalcoholic Fatty Liver Disease: Interplay between Diet, Gut Microbiota, and Genetic Background. *Gastroenterol Res Pract* 2016; **2016**: 1–13.

60 Lecoultre V, Egli L, Carrel G *et al.* Effects of fructose and glucose overfeeding on hepatic insulin sensitivity and intrahepatic lipids in healthy humans. *Obesity* 2013; **21**: 782–785.

61 Ouyang X, Cirillo P, Sautin Y, … SM-J of, 2008 undefined. Fructose consumption as a risk factor for non-alcoholic fatty liver disease. *journal-of-hepatology.eu*http://www.journal-of-hepatology.eu/article/S0168-8278(08)00164-5/abstract (accessed 6 Apr2018).

62 Obstfeld AE, Sugaru E, Thearle M *et al.* C-C Chemokine Receptor 2 (CCR2) Regulates the Hepatic Recruitment of Myeloid Cells That Promote Obesity-Induced Hepatic Steatosis. *Diabetes* 2010; **59**: 916–925.

63 Dumas M, Barton R, … AT-P of the, 2006 undefined. Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. *Natl Acad Sci*http://www.pnas.org/content/103/33/12511.short (accessed 6 Apr2018).

64 Kern PA, Saghizadeh M, Ong JM, Bosch RJ, Deem R, Simsolo RB. The expression of tumor necrosis factor in human adipose tissue. Regulation by obesity, weight loss, and relationship to lipoprotein lipase. *J Clin Invest* 1995; **95**: 2111–2119.

65 Marí M, Caballero F, Colell A *et al.* Mitochondrial free cholesterol loading sensitizes to TNF- and Fas-mediated steatohepatitis. *Cell Metab* 2006; **4**: 185–198.

66 Amar J, Burcelin R, Ruidavets JB *et al.* Energy intake is associated with endotoxemia in apparently healthy men. *Am J Clin Nutr* 2008; **87**: 1219–1223.

67 Brown AJ, Goldsworthy SM, Barnes AA *et al.* The Orphan G Protein-coupled Receptors GPR41 and GPR43 Are Activated by Propionate and Other Short Chain Carboxylic Acids. *J Biol Chem* 2003; **278**: 11312–11319.

68 Maslowski KM, Vieira AT, Ng A *et al.* Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. *Nature* 2009; **461**: 1282–1286.

69 Barbuio R, Milanski M, … MB-J of, 2007 undefined. Infliximab reverses steatosis and improves insulin signal transduction in liver of rats fed a high-fat diet. *Soc Endocrinol*http://joe.endocrinology-journals.org/content/194/3/539.short (accessed 6 Apr2018).

70 Lee A-H, Scapa EF, Cohen DE, Glimcher LH. Regulation of Hepatic Lipogenesis by the Transcription Factor XBP1. *Science (80-)* 2008; **320**: 1492–1496.

71 Stevens EA, Mezrich JD, Bradfield CA. The aryl hydrocarbon receptor: a perspective on potential roles in the immune system. *Immunology* 2009; **127**: 299–311.

72 Alkhouri N, Dixon LJ, Feldstein AE. Lipotoxicity in nonalcoholic fatty liver disease: not all lipids are created equal. *Expert Rev Gastroenterol Hepatol* 2009; **3**: 445–451.

73 Day C. Pathogenesis of steatohepatitis. *Best Pract Res Clin Gastroenterol* 2002; **16**: 663–678.

74 Neuschwander-Tetri BA. Hepatic lipotoxicity and the pathogenesis of nonalcoholic steatohepatitis: the central role of nontriglyceride fatty acid metabolites. *Hepatology* 2010; **52**: 774–88.

75 Sharifnia T, Antoun J, Verriere TGC *et al.* Hepatic TLR4 signaling in obese NAFLD. *Am J Physiol Liver Physiol* 2015; **309**: G270–G278.

76 Elamin EE, Masclee AA, Dekker J, Jonkers DM. Ethanol metabolism and its effects on the intestinal epithelial barrier. *Nutr Rev* 2013; **71**: 483–499.

77 Feldstein AE, Canbay A, Guicciardi ME, Higuchi H, Bronk SF, Gores GJ. Diet associated hepatic steatosis sensitizes to Fas mediated liver injury in mice. *J Hepatol* 2003; **39**: 978–83.

78 Arguello G, Balboa E, Arrese M, Biophysica SZ-B et, 2015 undefined. Recent insights on the role of cholesterol in non-alcoholic fatty liver disease. *Elsevier*https://www.sciencedirect.com/science/article/pii/S0925443915001647 (accessed 6 Apr2018).

79 Nair S, Cope K, Terence R, of AD-TA journal, 2001 undefined. Obesity and female gender increase breath ethanol concentration: potential implications for the pathogenesis of nonalcoholic steatohepatitis. *Elsevier*https://www.sciencedirect.com/science/article/pii/S0002927001022651 (accessed 27 Mar2018).

80 Cope K, Risby T, Gastroenterology AD-, 2000 undefined. Increased gastrointestinal ethanol production in obese mice: implications for fatty liver disease pathogenesis. *gastrojournal.org*http://www.gastrojournal.org/article/S0016-5085(00)20235-7/abstract (accessed 27 Mar2018).

81 Engstler A, Aumiller T, Degen C, Dürr M, Gut EW-, 2016 undefined. Insulin resistance alters hepatic ethanol metabolism: studies in mice and children with non-alcoholic fatty liver disease. *gut.bmj.com*http://gut.bmj.com/content/65/9/1564.abstract (accessed 27 Mar2018).

82 Medeiros I de, hypotheses J de L-M, 2015 undefined. Is nonalcoholic fatty liver disease an endogenous alcoholic fatty liver disease?–A mechanistic hypothesis. *medical-hypotheses.com*http://www.medical-hypotheses.com/article/S0306-9877(15)00164-4/abstract (accessed 27 Mar2018).

83 Hrubec Z, Omenn GS. Evidence of genetic predisposition to alcoholic cirrhosis and psychosis: twin concordances for alcoholism and its biological end points by zygosity among male veterans. *Alcohol Clin Exp Res* 1981; **5**: 207–15.

84 Jablon S, Neel J V, Gershowitz H, Atkinson GF. The NAS-NRC twin panel: methods of construction of the panel, zygosity diagnosis, and proposed use. *Am J Hum Genet* 1967; **19**: 133–61.

85 Krawczyk M, Rau M, Schattenberg JM *et al.* Combined effects of the PNPLA3 rs738409, TM6SF2 rs58542926, and MBOAT7 rs641738 variants on NAFLD severity: a multicenter biopsy-based study. *J Lipid Res* 2017; **58**: 247–255.

86 Metabolic Syndrome ePoster. http://www.nature.com/nm/e-poster/Liver-Organ-

Disease.html (accessed 25 Aug2017).

87 Cohen S, Ahn J. Review article: the diagnosis and management of alcoholic hepatitis. *Aliment Pharmacol Ther* 2009; **30**: 3–13.

88 Dunn W, Shah VH. Pathogenesis of Alcoholic Liver Disease. *Clin Liver Dis* 2016; **20**: 445–56.

89 Purohit V, Gao B, Song B-J. Molecular mechanisms of alcoholic fatty liver. *Alcohol Clin Exp Res* 2009; **33**: 191–205.

90 Setshedi M, Wands JR, de la Monte SM. Acetaldehyde Adducts in Alcoholic Liver Disease. *Oxid Med Cell Longev* 2010; **3**: 178–185.

91 Ji C, Chan C, Kaplowitz N. Predominant role of sterol response element binding proteins (SREBP) lipogenic pathways in hepatic steatosis in the murine intragastric ethanol feeding model. *J Hepatol* 2006; **45**: 717–724.

92 Parlesak A, Schäfer C, Schütz T, Bode JC, Bode C. Increased intestinal permeability to macromolecules and endotoxemia in patients with chronic alcohol abuse in different stages of alcohol-induced liver disease. *J Hepatol* 2000; **32**: 742–7.

93 Uesugi T, Froh M, Arteel GE, Bradford BU, Thurman RG. Toll-like receptor 4 is involved in the mechanism of early alcohol-induced liver injury in mice. *Hepatology* 2001; **34**: 101–108.

94 Baulande S, Lasnier F, Lucas M, Pairault J. Adiponutrin, a transmembrane protein corresponding to a novel dietary- and obesity-linked mRNA specifically expressed in the adipose lineage. *J Biol Chem* 2001; **276**: 33336–44.

95 Polson DA, Thompson MP. Adiponutrin mRNA expression in white adipose tissue is rapidly induced by meal-feeding a high-sucrose diet. *Biochem Biophys Res Commun* 2003; **301**: 261–6.

96 Polson D, Thompson M. Adiponutrin gene expression in 3T3-L1 adipocytes is downregulated by troglitazone. *Horm Metab Res* 2003; **35**: 508–10.

97 Moldes M, Beauregard G, Faraj M *et al.* Adiponutrin gene is regulated by insulin and glucose in human adipose tissue. *Eur J Endocrinol* 2006; **155**: 461–8.

98 Romeo S, Kozlitina J, Xing C *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 2008; **40**: 1461–5.

99 Valenti L, Al-Serri A, Daly AK *et al.* Homozygosity for the patatin-like phospholipase-3/adiponutrin I148M polymorphism influences liver fibrosis in patients with nonalcoholic fatty liver disease. *Hepatology* 2010; **51**: 1209–1217.

100 Hotta K, Yoneda M, Hyogo H *et al.* Association of the rs738409 polymorphism in PNPLA3 with liver damage and the development of nonalcoholic fatty liver disease. *BMC Med Genet* 2010; **11**: 172.

101 Tian C, Stokowski RP, Kershenobich D, Ballinger DG, Hinds DA. Variant in PNPLA3 is associated with alcoholic liver disease. *Nat Genet* 2010; **42**: 21–3.

102 Stickel F, Buch S, Lau K *et al.* Genetic variation in the PNPLA3 gene is associated with alcoholic liver injury in caucasians. *Hepatology* 2011; **53**: 86–95.

103 Rotman Y, Koh C, Zmuda JM, Kleiner DE, Liang TJ, NASH CRN. The association of genetic variability in patatin-like phospholipase domain-containing protein 3 (PNPLA3) with histological severity of nonalcoholic fatty liver disease. *Hepatology* 2010; **52**: 894–903.

104 Sookoian S, Pirola CJ. Meta-analysis of the influence of I148M variant of patatin-like phospholipase domain containing 3 gene (PNPLA3) on the susceptibility and histological severity of nonalcoholic fatty liver disease. *Hepatology* 2011; **53**: 1883–1894.

105 Valenti L, Rumi M, Galmozzi E *et al.* Patatin-Like phospholipase domain-containing 3 I148M polymorphism, steatosis, and liver damage in chronic hepatitis C. *Hepatology* 2011; **53**: 791–799.

106 Trépo E, Pradat P, Potthoff A *et al.* Impact of patatin-like phospholipase-3 (rs738409 C>G) polymorphism on fibrosis progression and steatosis in chronic hepatitis C. *Hepatology* 2011; **54**: 60–69.

107 Zain SM, Mohamed R, Mahadeva S *et al.* A multi-ethnic study of a PNPLA3 gene variant and its association with disease severity in non-alcoholic fatty liver disease. *Hum Genet* 2012; **131**: 1145–1152.

108 Burza MA, Pirazzi C, Maglio C *et al.* PNPLA3 I148M (rs738409) genetic variant is associated with hepatocellular carcinoma in obese individuals. *Dig Liver Dis* 2012; **44**: 1037–1041.

109 Viganò M, Valenti L, Lampertico P *et al.* Patatin-like phospholipase domain-containing 3 I148M affects liver steatosis in patients with chronic hepatitis B. *Hepatology* 2013; **58**: 1245–1252.

110 Liu Y-L, Patman GL, Leathart JBS *et al.* Carriage of the PNPLA3 rs738409 C >G polymorphism confers an increased risk of non-alcoholic fatty liver disease associated hepatocellular carcinoma. *J Hepatol* 2014; **61**: 75–81.

111 Trépo E, Nahon P, Bontempi G *et al.* Association between the *PNPLA3* (rs738409 C>G) variant and hepatocellular carcinoma: Evidence from a meta-analysis of individual participant data. *Hepatology* 2014; **59**: 2170–2177.

112 Singal AG, Manjunath H, Yopp AC *et al.* The Effect of PNPLA3 on Fibrosis Progression and Development of Hepatocellular Carcinoma: A Meta-analysis. *Am J Gastroenterol* 2014; **109**: 325–334.

113 De Nicola S, Dongiovanni P, Aghemo A *et al.* Interaction between PNPLA3 I148M Variant and Age at Infection in Determining Fibrosis Progression in Chronic Hepatitis C. *PLoS One* 2014; **9**: e106022.

114 Brouwer WP, van der Meer AJ, Boonstra A *et al.* The impact of *PNPLA3* (*rs738409* C>G) polymorphisms on liver histology and long-term clinical outcome in chronic hepatitis B patients. *Liver Int* 2015; **35**: 438–447.

115 Stättermayer AF, Traussnigg S, Dienes H-P *et al.* Hepatic steatosis in Wilson disease – Role of copper and PNPLA3 mutations. *J Hepatol* 2015; **63**: 156–163.

116 Krawczyk M, Stokes CS, Romeo S *et al.* HCC and liver disease risks in homozygous PNPLA3 p.I148M carriers approach monogenic inheritance. *J Hepatol* 2015; **62**: 980–981.

117 Luukkonen PK, Zhou Y, Sädevirta S *et al.* Hepatic ceramides dissociate steatosis and insulin resistance in patients with non-alcoholic fatty liver disease. *J Hepatol* 2016; **64**: 1167–1175.

118 Mancina RM, Spagnuolo R, Milano M *et al.* PNPLA3 148M Carriers with Inflammatory Bowel Diseases Have Higher Susceptibility to Hepatic Steatosis and Higher Liver Enzymes. *Inflamm Bowel Dis* 2016; **22**: 134–140.

119 Atkinson SR, Way MJ, McQuillin A, Morgan MY, Thursz MR. Homozygosity for rs738409:G in PNPLA3 is associated with increased mortality following an episode of severe alcoholic hepatitis. *J Hepatol* 2017; **67**: 120–127.

120 Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2013; **41**: D8–D20.

121 The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2014; **43**: D204-212.

122 Flicek P, Amode MR, Barrell D *et al.* Ensembl 2014. *Nucleic Acids Res* 2014; **42**: D749–D755.

123 Auton A, Abecasis GR, Altshuler DM *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.

124 Walter K, Min JL, Huang J *et al.* The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**: 82–90.

125 Baulande S, Fève B. Identification de nouveaux gènes associés à l'adipogenèse. médecine Sci. 2003; : 151–154.

126 Hoekstra M, Li Z, Kruijt JK, Van Eck M, Van Berkel TJC, Kuiper J. The expression level of non-alcoholic fatty liver disease-related gene PNPLA3 in hepatocytes is highly influenced by hepatic lipid status. *J Hepatol* 2010; **52**: 244–51.

127 Miyaaki H, Miuma S, Taura N *et al. PNPLA3* as a liver steatosis risk factor following living-donor liver transplantation for hepatitis C. *Hepatol Res* 2018; **48**: E335–E339.

128 Wiesner G, Morash BA, Ur E, Wilkinson M. Food restriction regulates adipose-specific cytokines in pituitary gland but not in hypothalamus. *J Endocrinol* 2004; **180**: R1-6.

129 Polson DA, Thompson MP. Macronutrient composition of the diet differentially affects leptin and adiponutrin mRNA expression in response to meal feeding. *J Nutr Biochem* 2004; **15**: 242–6.

130 Faraj M, Beauregard G, Loizon E *et al.* Insulin regulation of gene expression and concentrations of white adipose tissue-derived proteins in vivo in healthy men: relation to adiponutrin. *J Endocrinol* 2006; **191**: 427–35.

131 Kollerits B, Coassin S, Beckmann ND *et al.* Genetic evidence for a role of adiponutrin in the metabolism of apolipoprotein B-containing lipoproteins. *Hum Mol Genet* 2009; **18**: 4669–76.

132 Liang H, Xu J, Xu F *et al.* The SRE motif in the human PNPLA3 promoter (-97 to -88 bp) mediates transactivational effects of SREBP-1c. *J Cell Physiol* 2015; **230**: 2224–2232.

133 Hua X, Nohturfft A, Goldstein JL, Brown MS. Sterol resistance in CHO cells traced to point mutation in SREBP cleavage-activating protein. *Cell* 1996; **87**: 415–26.

134 Dubuquoy C, Robichon C, Lasnier F *et al.* Distinct regulation of adiponutrin/PNPLA3 gene expression by the transcription factors ChREBP and SREBP1c in mouse and human hepatocytes. *J Hepatol* 2011; **55**: 145–53.

135 Perttila J, Huaman-Samanez C, Caron S *et al.* PNPLA3 is regulated by glucose in human hepatocytes, and its I148M mutant slows down triglyceride hydrolysis. *AJP Endocrinol Metab* 2012; **302**: E1063–E1069.

136 Qiao A, Liang J, Ke Y *et al.* Mouse patatin-like phospholipase domain-containing 3 influences systemic lipid and glucose homeostasis. *Hepatology* 2011; **54**: 509–521.

137 Eberlé D, Hegarty B, Bossard P, Ferré P, Foufelle F. SREBP transcription factors: master regulators of lipid homeostasis. *Biochimie* 2004; **86**: 839–48.

138 Kershaw EE, Schupp M, Guan H-P, Gardner NP, Lazar MA, Flier JS. PPARgamma regulates adipose triglyceride lipase in adipocytes in vitro and in vivo. *Am J Physiol Endocrinol Metab* 2007; **293**: E1736-45.

139 Xu F, Li Z, Zheng X *et al.* SIRT1 mediates the effect of GLP-1 receptor agonist exenatide on ameliorating hepatic steatosis. *Diabetes* 2014; **63**: 3637–46.

140 Oliver P, Caimari A, Díaz-Rúa R, Palou A. Cold exposure down-regulates adiponutrin/PNPLA3 mRNA expression and affects its nutritional regulation in adipose tissues of lean and obese Zucker rats. *Br J Nutr* 2012; **107**: 1283–95.

141 Calvo RM, Obregon MJ. Tri-iodothyronine upregulates adiponutrin mRNA expression in rat and human adipocytes. *Mol Cell Endocrinol* 2009; **311**: 39–46.

142 Maslov LN, Vychuzhanova EA, Gorbunov AS, Tsybul'nikov SI, Khaliulin IG, Chauski E. Role of thyroid system in adaptation to cold. *Ross Fiziol zhurnal Im IM Sechenova / Ross Akad Nauk* 2014; **100**: 670–83.

143 Pirazzi C, Valenti L, Motta BM *et al.* PNPLA3 has retinyl-palmitate lipase activity in human hepatic stellate cells. *Hum Mol Genet* 2014; **23**: 4077–4085.

144 Shukla SD, Restrepo R, Fish P, Lim RW, Ibdah JA. Different mechanisms for histone acetylation by ethanol and its metabolite acetate in rat primary hepatocytes. *J Pharmacol Exp Ther* 2015; **354**: 18–23.

145 Bateman A, Coin L, Durbin R *et al.* The Pfam protein families database. *Nucleic Acids Res* 2004; **32**: 138D–141.

146 Galliard T. The enzymic deacylation of phospholipids and galactolipids in plants. Purification and properties of a lipolytic acyl-hydrolase from potato tubers. *Biochem J* 1971; **121**: 379–90.

147 Schrag JD, Li Y, Wu S, Cygler M. Ser-His-Glu triad forms the catalytic site of the lipase from Geotrichum candidum. *Nature* 1991; **351**: 761–764.

148 Paiva E, Lister RM, Park WD. Induction and accumulation of major tuber proteins of potato in stems and petioles. *Plant Physiol* 1983; **71**: 161–8.

149 Wilson P a, Gardner SD, Lambie NM, Commans S a, Crowther DJ. Characterization of the human patatin-like phospholipase family. *J Lipid Res* 2006; **47**: 1940–1949.

150 Kienesberger PC, Oberer M, Lass A, Zechner R. Mammalian patatin domain containing proteins: a family with diverse lipolytic activities involved in multiple biological functions. *J Lipid Res* 2009; **50 Suppl**: S63-8.

151 Lake AC, Sun Y, Li J-L *et al.* Expression, regulation, and triglyceride hydrolase activity of Adiponutrin family members. *J Lipid Res* 2005; **46**: 2477–87.

152 Zechner R, Kienesberger PC, Haemmerle G, Zimmermann R, Lass A. Adipose triglyceride lipase and the lipolytic catabolism of cellular fat stores. *J Lipid Res* 2009; **50**: 3–21.

153 Haemmerle G, Lass A, Zimmermann R *et al.* Defective Lipolysis and Altered Energy Metabolism in Mice Lacking Adipose Triglyceride Lipase. *Science (80-)* 2006; **312**: 734–737.

154 Igal RA, Rhoads JM, Coleman RA. Neutral lipid storage disease with fatty liver and cholestasis. *J Pediatr Gastroenterol Nutr* 1997; **25**: 541–7.

155 Fischer J, Lefèvre C, Morava E *et al.* The gene encoding adipose triglyceride lipase (PNPLA2) is mutated in neutral lipid storage disease with myopathy. *Nat Genet* 2007; **39**: 28–30.

156 Chen Z, Gao X, Lei T *et al.* Molecular characterization, expression and chromosomal localization of porcine PNPLA3 and PNPLA4. *Biotechnol Lett* 2011; **33**: 1327–1337.

157 Amino Acid. https://en.wikipedia.org/wiki/Amino_acid (accessed 4 Aug2017).

158 He S, McPhaul C, Li JZ *et al.* A sequence variation (I148M) in PNPLA3 associated with nonalcoholic fatty liver disease disrupts triglyceride hydrolysis. *J Biol Chem* 2010; **285**: 6706–15.

159 Ding Y, Zhang S, Yang L *et al.* Isolating lipid droplets from multiple species. *Nat Protoc* 2013; **8**: 43–51.

160 Murugesan S, Goldberg EB, Dou E, Brown WJ. Identification of diverse lipid droplet targeting motifs in the PNPLA family of triglyceride lipases. *PLoS One* 2013; **8**: e64950.

161 Winberg ME, Khalaj Motlagh M, Stenkula KG, Holm C, Jones H a. Adiponutrin: A multimeric plasma protein. *Biochem Biophys Res Commun* 2014; **446**: 1114–1119.

162 Hotta K, Funahashi T, Arita Y *et al.* Plasma concentrations of a novel, adipose-specific protein, adiponectin, in type 2 diabetic patients. *Arterioscler Thromb Vasc Biol* 2000; **20**: 1595–1599.

163 Jenkins CM, Mancuso DJ, Yan W, Sims HF, Gibson B, Gross RW. Identification, cloning, expression, and purification of three novel human calcium-independent phospholipase A2 family members possessing triacylglycerol lipase and acylglycerol transacylase activities. *J Biol Chem* 2004; **279**: 48968–75.

164 Pingitore P, Pirazzi C, Mancina RM *et al.* Recombinant PNPLA3 protein shows triglyceride hydrolase activity and its I148M mutation results in loss of function. *Biochim Biophys Acta* 2014; **1841**: 574–80.

165 Huang Y, Cohen JC, Hobbs HH. Expression and characterization of a PNPLA3 protein isoform (I148M) associated with nonalcoholic fatty liver disease. *J Biol Chem* 2011; **286**: 37085–93.

166 Kumari M, Schoiswohl G, Chitraju C *et al.* Adiponutrin functions as a nutritionally regulated lysophosphatidic acid acyltransferase. *Cell Metab* 2012; **15**: 691–702.

167 Kershaw EE, Hamm JK, Verhagen LAW, Peroni O, Katic M, Flier JS. Adipose triglyceride lipase: function, regulation by insulin, and comparison with adiponutrin. *Diabetes* 2006; **55**: 148–57.

168 Pirazzi C, Adiels M, Burza MA *et al.* Patatin-like phospholipase domain-containing 3 (PNPLA3) I148M (rs738409) affects hepatic VLDL secretion in humans and in vitro. *J Hepatol* 2012; **57**: 1276–82.

169 Ruhanen H, Perttilä J, Hölttä-Vuori M *et al.* PNPLA3 mediates hepatocyte triacylglycerol remodeling. *J Lipid Res* 2014; **55**: 739–46.

170 Li JZ, Huang Y, Karaman R *et al.* Chronic overexpression of PNPLA3I148M in mouse liver causes hepatic steatosis. *J Clin Invest* 2012; **122**: 4130–44.

171 Chen W, Chang B, Li L, Chan L. Patatin-like phospholipase domain-containing 3/adiponutrin deficiency in mice is not associated with fatty liver disease. *Hepatology* 2010; **52**: 1134–42.

172 Basantani MK, Sitnick MT, Cai L *et al.* Pnpla3/Adiponutrin deficiency in mice does not contribute to fatty liver disease or metabolic syndrome. *J Lipid Res* 2011; **52**: 318–29.

173 Mitsche MA, Hobbs HH, Cohen JC. Patatin-like phospholipase domain-containing protein 3 promotes transfers of essential fatty acids from triglycerides to phospholipids in hepatic lipid droplets. *J Biol Chem* 2018; **293**: 6958–6968.

174 Smagris E, BasuRay S, Li J *et al.* Pnpla3I148M knockin mice accumulate PNPLA3 on lipid droplets and develop hepatic steatosis. *Hepatology* 2015; **61**: 108–18.

175 Ducharme NA, Bickel PE. Lipid droplets in lipogenesis and lipolysis. *Endocrinology* 2008; **149**: 942–9.

176 Bruschi FV, Claudel T, Tardelli M *et al.* The PNPLA3 I148M variant modulates the fibrogenic phenotype of human hepatic stellate cells. *Hepatology* 2017; **65**: 1875–1890.

177 Furnham N, Garavelli JS, Apweiler R, Thornton JM. Missing in action: enzyme functional annotations in biological databases. *Nat Chem Biol* 2009; **5**: 521–525.

178 Chemistry of amino acids and protein structure (article) | Khan Academy. https://www.khanacademy.org/test-prep/mcat/biomolecules/amino-acids-and-proteins1/a/chemistry-of-amino-acids-and-protein-structure (accessed 18 Jul2017).

179 McCammon JA, Harvey SC. *Dynamics of proteins and nucleic acids*. Cambridge University Press: Cambridge, 1987 doi:10.1017/CBO9781139167864.

180 Fang Y. Thermodynamic Principle Revisited: Theory of Protein Folding. *Adv Biosci Biotechnol* 2015; **6**: 37–48.

181 Feher J. *Quantitative Human Physiology (Second Edition)*. 2017https://www.sciencedirect.com/topics/engineering/alpha-helix (accessed 21 Feb2019).

182 Weiss MA, Ellenberger T, Wobbe CR, Lee JP, Harrison SC, Struhl K. Folding transition in the DMA-binding domain of GCN4 on specific binding to DNA. *Nature* 1990; **347**: 575–578.

183 Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. Intrinsic disorder and protein function. *Biochemistry* 2002; **41**: 6573–82.

184 Onuchic JN, Wolynes PG. Theory of protein folding. *Curr Opin Struct Biol* 2004; **14**: 70–75.

185 Rogerdodd. File:Ubiquiting_cartoon-2. https://en.wikipedia.org/wiki/Ubiquitin#/media/File:Ubiquitin_cartoon-2-.png.

186 Dill KA. Dominant forces in protein folding. *Biochemistry* 1990; **29**: 7133–7155.

187 File:COX-2 inhibited by Aspirin.png - Wikimedia Commons. https://commons.wikimedia.org/wiki/File:COX-2_inhibited_by_Aspirin.png (accessed 22 Feb2019).

188 Hubbard TJP, Blundell TL. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *"Protein Eng Des Sel* 1987; **1**: 159–171.

189 Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992; **89**: 10915–9.

190 Holm L, Sander C. Decision support system for the evolutionary classification of protein structures. *Proceedings Int Conf Intell Syst Mol Biol* 1997; **5**: 140–6.

191 Grishin N V. Fold Change in Evolution of Protein Structures. *J Struct Biol* 2001; **134**: 167–185.

192 Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. *Genome Res* 2002; **12**: 1619–23.

193 Wilkins MR, Gasteiger E, Bairoch A *et al.* Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol* 1999; **112**: 531–52.

194 Matthew JB, Friend SH, Botelho LH, Lehman LD, Hanania GI, Gurd FR. Discrete charge calculations of potentiometric titrations for globular proteins: sperm whale myoglobin, hemoglobin alpha chain, cytochrome c. *Biochem Biophys Res Commun* 1978; **81**: 416–21.

195 Pace CN, Vajdos F, Fee L, Grimsley G, Gray T. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci* 1995; **4**: 2411–2423.

196 Wingfield PT. Overview of the purification of recombinant proteins. *Curr Protoc protein Sci* 2015; **80**: 6.1.1-35.

197 Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982; **157**: 105–132.

198 Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004; **337**: 635–45.

199 Melamud E, Moult J. Evaluation of disorder predictions in CASP5. *Proteins Struct Funct Genet* 2003; **53**: 561–565.

200 Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 2007; **35**: W460–W464.

201 Myers JK, Oas TG. Preorganized secondary structure as an important determinant of fast protein folding. *Nat Struct Biol* 2001; **8**: 552–558.

202 Källberg M, Wang H, Wang S *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* 2012; **7**: 1511–1522.

203 Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 2008; **9**: 40.

204 Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol* 1997; **268**: 209–225.

205 Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999; **292**: 195–202.

206 Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res* 2013; **41**: W349-57.

207 Wang S, Peng J, Ma J, Xu J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep* 2016; **6**: 18962.

208 Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 2008; **24**: 1662–8.

209 Viklund H, Bernsel A, Skwark M, Elofsson A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 2008; **24**: 2928–9.

210 Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 1994; **33**: 3038–49.

211 Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 2007; **23**: 538–44.

212 Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001; **305**: 567–80.

213 Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* 2007; **35**: W429-32.

214 Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics* 2009; **10**: 159.

215 Reeb J, Kloppmann E, Bernhofer M, Rost B. Evaluation of transmembrane helix predictions in 2014. *Proteins Struct Funct Bioinforma* 2015; **83**: 473–484.

216 Bryson K, Cozzetto D, Jones DT. Computer-assisted protein domain boundary prediction using the DomPred server. *Curr Protein Pept Sci* 2007; **8**: 181–8.

217 Marsden RL, McGuffin LJ, Jones DT. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci* 2002; **11**: 2814–24.

218 Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* 2013; **29**: i247-56.

219 Khoury GA, Baliban RC, Floudas CA. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep* 2011; **1**: 90.

220 Bendtsen JD, Jensen LJ, Blom N, von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel* 2004; **17**: 349–356.

221 Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. *BMC Microbiol* 2005; **5**. doi:10.1186/1471-2180-5-58.

222 Xue Y, Zhou F, Fu C, Xu Y, Yao X. SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res* 2006; **34**: W254–W257.

223 Minneci F, Piovesan D, Cozzetto D, Jones DT. FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences. *PLoS One* 2013; **8**: e63754.

224 Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 2004; **4**: 1633–1649.

225 Qiu W-R, Xiao X, Lin W-Z, Chou K-C. iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J Biomol Struct Dyn* 2015; **33**: 1731–42.

226 Radivojac P, Vacic V, Haynes C *et al.* Identification, analysis, and prediction of protein ubiquitination sites. *Proteins Struct Funct Bioinforma* 2010; **78**: 365–380.

227 Deller MC, Kong L, Rupp B. Protein stability: a crystallographer's perspective. *Acta Crystallogr Sect F Struct Biol Commun* 2016; **72**: 72–95.

228 Slabinski L, Jaroszewski L, Rychlewski L, Wilson IA, Lesley SA, Godzik A. XtalPred: a web

server for prediction of protein crystallizability. *Bioinformatics* 2007; **23**: 3403–5.

229 Wang H, Feng L, Zhang Z, Webb GI, Lin D, Song J. Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci Rep* 2016; **6**: 21383.

230 Ashburner M, Ball C, Blake J, Botstein D, … HB-N, 2000 undefined. Gene ontology: tool for the unification of biology. *nature.com*https://www.nature.com/articles/ng0500_25 (accessed 26 Feb2019).

231 Ashburner M, Ball C, Blake J, Botstein D, … HB-N, 2000 undefined. Gene ontology: tool for the unification of biology. *nature.com*https://www.nature.com/articles/ng0500_25 (accessed 22 Feb2019).

232 Martin D, … MB-B, 2004 undefined. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *bmcbioinformatics.biomedcentral …*https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-178 (accessed 26 Feb2019).

233 Chitale M, Hawkins T, Park C, Bioinformatics DK-, 2009 undefined. ESG: extended similarity group method for automated protein function prediction. *academic.oup.com*https://academic.oup.com/bioinformatics/article-abstract/25/14/1739/224617 (accessed 26 Feb2019).

234 Wilkins A, Bachman B, … SE-C opinion in, 2012 undefined. The use of evolutionary patterns in protein annotation. *Elsevier*https://www.sciencedirect.com/science/article/pii/S0959440X12000759 (accessed 22 Feb2019).

235 biology BR-J of molecular, 2002 undefined. Enzyme function less conserved than anticipated. *Elsevier*https://www.sciencedirect.com/science/article/pii/S0022283602000165 (accessed 26 Feb2019).

236 Letunic I, Goodstadt L, Dickens NJ *et al.* Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res* 2002; **30**: 242–4.

237 Bateman A, Birney E, Cerruti L *et al.* The Pfam protein families database. *Nucleic Acids Res* 2002; **30**: 276–80.

238 Marchler-Bauer A, Derbyshire MK, Gonzales NR *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Res* 2015; **43**: D222–D226.

239 Marchler-Bauer A, Bo Y, Han L *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* 2017; **45**: D200–D203.

240 Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* 2002; **30**: 281–3.

241 Marchler-Bauer A, Anderson JB, Derbyshire MK *et al.* CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 2007; **35**: D237–D240.

242 Tjong H, Qin S, Zhou H-X. PI2PE: protein interface/interior prediction engine. *Nucleic Acids Res* 2007; **35**: W357-62.

243 Chou K-C, Shen H-B. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Commun* 2007; **360**: 339–45.

244 Pellegrini-Calace M, Maiwald T, Thornton JM. PoreWalker: A Novel Tool for the Identification and Characterization of Channels in Transmembrane Proteins from Their Three-Dimensional Structure. *PLoS Comput Biol* 2009; **5**: e1000440.

245 Meruelo AD, Samish I, Bowie JU. TMKink: a method to predict transmembrane helix kinks. *Protein Sci* 2011; **20**: 1256–64.

246 Fontana P, Bindewald E, Toppo S, Velasco R, Valle G, Tosatto SCE. The SSEA server for protein secondary structure alignment. *Bioinforma Appl NOTE* 2005; **21**: 393–395.

247 Dessen A. Phospholipase A2 enzymes: structural diversity in lipid messenger metabolism. *Structure* 2000; **8**: R15–R22.

248 Gao JG, Simon M. A comparative study of human GS2, its paralogues, and its rat orthologue. *Biochem Biophys Res Commun* 2007; **360**: 501–506.

249 Dupont N, Chauhan S, Arko-Mensah J *et al.* Neutral Lipid Stores and Lipase PNPLA5 Contribute to Autophagosome Biogenesis. *Curr Biol* 2014; **24**: 609–620.

250 Lazniewski M, Steczkiewicz K, Knizewski L, Wawer I, Ginalski K. Novel transmembrane lipases of alpha/beta hydrolase fold. *FEBS Lett* 2011; **585**: 870–874.

251 Bowie JU. Stabilizing membrane proteins. *Curr Opin Struct Biol* 2001; **11**: 397–402.

252 Smyth DR, Mrozkiewicz MK, McGrath WJ, Listwan P, Kobe B. Crystal structures of fusion proteins with large-affinity tags. *Protein Sci* 2003; **12**: 1313–22.

253 Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 2005; **21**: 3369–3376.

254 Structural Genomics Consortium SG, China Structural Genomics Consortium A et F des M, Northeast Structural Genomics Consortium BSG *et al.* Protein production and purification. *Nat Methods* 2008; **5**: 135–46.

255 Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* 2001; **40**: 346–58.

256 Bernal JD, Crowfoot D. X-Ray Photographs of Crystalline Pepsin : Abstract : Nature. *Nature* 1934; **133**: 794–795.

257 Egli M. Diffraction techniques in structural biology. *Curr Protoc nucleic acid Chem* 2010; **Chapter 7**: Unit 7.13.

258 Zheng H, Handing KB, Zimmerman MD, Shabalin IG, Almo SC, Minor W. X-ray crystallography over the past decade for novel drug discovery – where are we heading next? *Expert Opin Drug Discov* 2015; **10**: 975–989.

259 Sayers Z, Avşar B, Cholak E, Karmous I. Application of advanced X-ray methods in life sciences. *Biochim Biophys Acta - Gen Subj* 2017; **1861**: 3671–3685.

260 Gräslund S, Nordlund P, Weigelt J *et al.* Protein production and purification. *Nat Methods* 2008; **5**: 135–146.

261 Vector Database. https://www.addgene.org/vector-database/ (accessed 2 Aug2017).

262 Tayar S, Kleinberger-Doron N. vector's anatomy. 2014.http://wolfson.huji.ac.il/expression/vector/vec-anat.html (accessed 2 Aug2017).

263 Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. *Trends Biotechnol* 2004; **22**: 346–353.

264 Hughes RA, Miklos AE, Ellington AD. Gene Synthesis. In: *Methods in enzymology*. 2011, pp 277–309.

265 Chin JX, Chung BK-S, Lee D-Y. Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics* 2014; **30**: 2210–2212.

266 Welch M, Govindarajan S, Ness JE *et al.* Design Parameters to Control Synthetic Gene Expression in Escherichia coli. *PLoS One* 2009; **4**: e7002.

267 Cai H, Li Y, Zhang H, Feng F. [Effects of gene design on recombinant protein expression: a review]. *Sheng Wu Gong Cheng Xue Bao* 2013; **29**: 1201–13.

268 Klock HE, Koesema EJ, Knuth MW, Lesley SA. Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins Struct Funct Bioinforma* 2008; **71**: 982–994.

269 Hammarström M, Hellgren N, van den Berg S, Berglund H, Härd T. Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. *Protein Sci* 2009; **11**: 313–321.

270 Waugh DS. Making the most of affinity tags. *Trends Biotechnol* 2005; **23**: 316–320.

271 Carson M, Johnson DH, McDonald H, Brouillette C, DeLucas LJ. His-tag impact on structure. *Acta Crystallogr Sect D Biol Crystallogr* 2007; **63**: 295–301.

272 Waugh DS. An overview of enzymatic reagents for the removal of affinity tags. *Protein Expr Purif* 2011; **80**: 283–93.

273 Nettleship JE, Assenberg R, Diprose JM, Rahman-Huq N, Owens RJ. Recent advances in the production of proteins in insect and mammalian cells for structural biology. *J Struct Biol* 2010; **172**: 55–65.

274 Peti W, Page R. Strategies to maximize heterologous protein expression in Escherichia coli with minimal cost. *Protein Expr Purif* 2007; **51**: 1–10.

275 Studier FW, Moffatt BA. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol* 1986; **189**: 113–30.

276 Studier FW, Daegelen P, Lenski RE, Maslov S, Kim JF. Understanding the differences between genome sequences of Escherichia coli B strains REL606 and BL21(DE3) and comparison of the E. coli B and K-12 genomes. *J Mol Biol* 2009; **394**: 653–80.

277 Tegel H, Tourle S, Ottosson J, Persson A. Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). *Protein Expr Purif* 2010; **69**: 159–167.

278 Dumon-Seignovert L, Cariot G, Vuillard L. The toxicity of recombinant proteins in Escherichia coli: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3). *Protein Expr Purif* 2004; **37**: 203–206.

279 Hage DS, Matsuda R. Affinity Chromatography: A Historical Perspective. Humana Press, New York, NY, 2015, pp 1–19.

280 Porath J, Carlsson J, Olsson I, Belfrage G. Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature* 1975; **258**: 598–9.

281 Arnold FH. Metal-affinity separations: a new dimension. *Nat Biotechnol* 1991; **9**: 151–156.

282 Cheung RCF, Wong JH, Ng TB. Immobilized metal ion affinity chromatography: a review on its applications. *Appl Microbiol Biotechnol* 2012; **96**: 1411–1420.

283 Pasquinelli RS, Shepherd RE, Koepsel RR, Zhao A, Ataai MM. Design of Affinity Tags for One-Step Protein Purification from Immobilized Zinc Columns. *Biotechnol Prog* 2000; **16**: 86–91.

284 SYNGE RLM. Fractionation of hydrolysis products of amylose by electrokinetic ultrafiltration in an agaragar jelly. *Biochem J* 1950; **2**: 41–2.

285 WHEATON RM, BAUMAN WC. Nonionic separations with ion exchange resins. *Ann N Y Acad Sci* 1953; **57**: 159–76.

286 Huo Q. Chapter 16 – Synthetic Chemistry of the Inorganic Ordered Porous Materials. In: *Modern Inorganic Synthetic Chemistry*. 2011, pp 339–373.

287 Flodin PGM. Process for preparign hydrophilic copolymerization and product obtained thereby. 1965.file:///C:/Users/Billy/Downloads/US3208994.pdf (accessed 1 Aug2017).

288 Striegel AM, Yau WW, Kirkland JJ, Bly DD. *Modern Size-Exclusion Liquid Chromatography*. John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009 doi:10.1002/9780470442876.

289 Heyden Y Vander, Popovici ST, Schoenmaker PJ. Evaluation of size-exclusion chromatography and size-exclusion electrochromatography calibration curves. *J Chromatogr A* 2002; **957**: 127–37.

290 Kostanski LK, Keller DM, Hamielec AE. Size-exclusion chromatography—a review of calibration methodologies. *J Biochem Biophys Methods* 2004; **58**: 159–186.

291 Hong P, Koza S, Bouvier ESP. Size-Exclusion Chromatography for the Analysis of Protein Biotherapeutics and their Aggregates. *J Liq Chromatogr Relat Technol* 2012; **35**: 2923–2950.

292 Using a Gel Filtration Chromatogram to Estimate Molecular Weight - Bitesize Bio. https://bitesizebio.com/29685/determine-molecular-weight-gel-filtration-chromatogram/ (accessed 23 Mar2018).

293 Hames BD. *Gel Electrophoresis of Proteins*. 3rd edition. Oxford University Press, 1998http://www.aun.edu.eg/molecular_biology/Protein workshop/Hames_Gel Electrophoresis of Proteins-A Practical Approach 3rd ed.pdf (accessed 8 Aug2017).

294 Vesterberg O. Staining of protein zones after isoelectric focusing in polyacrylamide gels. *Biochim Biophys Acta - Protein Struct* 1971; **243**: 345–348.

295 Guan Y, Zhu Q, Huang D, Zhao S, Lo LJ, Peng J. An equation to estimate the difference between theoretically predicted and SDS PAGE-displayed molecular weights for an acidic peptide. *Nat Publ Gr* 2015. doi:10.1038/srep13370.

296 Introduction to PAGE | Sigma-Aldrich. http://www.sigmaaldrich.com/technical-documents/articles/biology/sds-page.html (accessed 8 Aug2017).

297 Egger D, Bienz K. Protein (Western) Blotting. https://link.springer.com/content/pdf/10.1007%2FBF02921696.pdf (accessed 22 Aug2017).

298 General Western Blot Protocol - Leinco. https://www.leinco.com/general_wb (accessed 8 Aug2017).

299 Giegé R, Sauter C. Biocrystallography: past, present, future. *HFSP J* 2010; **4**: 109–21.

300 Lima-de-Faria J. *Historical Atlas of Crystallography : J. Lima-de-Faria : 9780792306498*. 1990th ed. Kluwer Academic Publishers: Dordrecht, 1990https://www.bookdepository.com/Historical-Atlas-Crystallography-J-Lima-de-Faria/9780792306498 (accessed 3 Aug2017).

301 Drenth J, Haas C. Protein crystals and their stability. *J Cryst Growth* 1992; **122**: 107–109.

302 McPherson A, Malkin A, Kuznetsov Y. The science of macromolecular crystallization. *Structure* 1995; **3**: 759–768.

303 Blow DM, Chayen NE, Lloyd LF, Saridakis E. Control of nucleation of protein crystals. *Protein Sci* 1994; **3**: 1638–1643.

304 Sato K, Fukuba Y, Mitsuda T, Hirai K, Moriya K. Observation of lattice defects in orthorhombic hen-egg white lysozyme crystals with laser scattering tomography. *J Cryst Growth* 1992; **122**: 87–94.

305 McPherson A, Gavira JA. Introduction to protein crystallization. *Acta Crystallogr Sect F, Struct Biol Commun* 2014; **70**: 2–20.

306 Gavira JA. Current trends in protein crystallization. *Arch Biochem Biophys* 2016; **602**: 3–11.

307 Chirgadze D. Protein crystallisation in action. 2001.http://www.xray.bioc.cam.ac.uk/xray_resources/whitepapers/xtal-in-action/xtal-in-action-html.html (accessed 3 Aug2017).

308 Asherie N. Protein crystallization and phase diagrams. *Methods* 2004; **34**: 266–272.

309 Sauter C, Lorber B, Kern D, Cavarelli J, Moras D, Giegé R. Crystallogenesis studies on yeast aspartyl-tRNA synthetase: use of phase diagram to improve crystal quality. *Acta Crystallogr Sect D Biol Crystallogr* 1999; **55**: 149–156.

310 Bard J. Automated systems for protein crystallization. *Methods* 2004; **34**: 329–347.

311 Fazio VJ, Peat TS, Newman J. A drunken search in crystallization space. *Acta Crystallogr Sect F, Struct Biol Commun* 2014; **70**: 1303–11.

312 Chayen NE. Turning protein crystallisation from an art into a science. *Curr Opin Struct Biol* 2004; **14**: 577–583.

313 Way MJ. *The genetics of alcohol related liver disease*. 2016.

314 Bird LE, Rada H, Flanagan J, Diprose JM, Gilbert RJC, Owens RJ. Application of In-Fusion™ Cloning for the Parallel Construction of E. coli Expression Vectors. In: *Methods in molecular biology (Clifton, N.J.)*. 2014, pp 209–234.

315 Addgene. https://www.addgene.org/ (accessed 4 Aug2017).

316 Takara. pCold TF DNA product manual. 2015.file:///C:/Users/billy/Downloads/3365_e.v1509Da.pdf (accessed 4 Aug2017).

317 Kaiser CM, Chang H-C, Agashe VR *et al.* Real-time observation of trigger factor function on translating ribosomes. *Nature* 2006; **444**: 455–60.

318 Hussain R, Benning K, Myatt D *et al.* CDApps: integrated software for experimental planning and data processing at beamline B23, Diamond Light Source. Corrigendum. *J Synchrotron Radiat* 2015; **22**: 862.

319 Shevchenko A, Tomas H, Havli\[sbreve] J, Olsen J V, Mann M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* 2007; **1**: 2856–

2860.

320 Graca R, Messick J, McCullough S, Barger A, Hoffmann W. Validation and diagnostic efficacy of a lipase assay using the substrate 1,2-o-dilauryl-rac-glycero glutaric acid-(6'methyl resorufin)-ester for the diagnosis of acute pancreatitis in dogs. *Vet Clin Pathol* 2005; **34**: 39–43.

321 Rath A, Glibowicka M, Nadeau VG, Chen G, Deber CM. Detergent binding explains anomalous SDS-PAGE migration of membrane proteins. *Proc Natl Acad Sci U S A* 2009; **106**: 1760–5.

322 Rosano GL, Ceccarelli EA. Recombinant protein expression in Escherichia coli: advances and challenges. *Front Microbiol* 2014; **5**: 172.

323 Shewry PR. Tuber storage proteins. *Ann Bot* 2003; **91**: 755–69.

324 Andersen KR, Leksa NC, Schwartz TU. Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification. *Proteins Struct Funct Bioinforma* 2013; **81**: 1857–1861.

325 Sugiki T, Kobayashi N, Fujiwara T. Modern Technologies of Solution Nuclear Magnetic Resonance Spectroscopy for Three-dimensional Structure Determination of Proteins Open Avenues for Life Scientists. *Comput Struct Biotechnol J* 2017; **15**: 328–339.

326 Ilari A, Savino C. Protein Structure Determination by X-Ray Crystallography. In: *Methods in molecular biology (Clifton, N.J.)*. 2008, pp 63–87.

327 W?thrich K. The way to NMR structures of proteins. *Nat Struct Biol* 2001; **8**: 923–925.

328 Zhou ZH. Atomic resolution cryo electron microscopy of macromolecular complexes. In: *Advances in protein chemistry and structural biology*. 2011, pp 1–35.

329 Jensen MR, Markwick PRL, Meier S *et al.* Quantitative Determination of the Conformational Properties of Partially Folded and Intrinsically Disordered Proteins Using NMR Dipolar Couplings. *Structure* 2009; **17**: 1169–1185.

330 Salmon L, Jensen MR, Bernadó P, Blackledge M. Measurement and Analysis of NMR Residual Dipolar Couplings for the Study of Intrinsically Disordered Proteins. In: *Methods in molecular biology (Clifton, N.J.)*. 2012, pp 115–125.

331 Gil S, Ho?ek T, Solyom Z *et al.* NMR Spectroscopic Studies of Intrinsically Disordered Proteins at Near-Physiological Conditions. *Angew Chemie Int Ed* 2013; **52**: 11808–11812.

332 Jensen MR, Ruigrok RW, Blackledge M. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr Opin Struct Biol* 2013; **23**: 426–435.

333 Frueh DP, Goodrich AC, Mishra SH, Nichols SR. NMR methods for structural studies of large monomeric and multimeric proteins. *Curr Opin Struct Biol* 2013; **23**: 734–9.

334 Vonck J, Mills DJ. Advances in high-resolution cryo-EM of oligomeric enzymes. *Curr Opin Struct Biol* 2017; **46**: 48–54.

335 Bartesaghi A, Merk A, Banerjee S *et al.* 2.2 A resolution cryo-EM structure of ?-galactosidase in complex with a cell-permeant inhibitor. *Science (80-)* 2015; **348**: 1147–1151.

336 Zheng H, Handing KB, Zimmerman MD, Shabalin IG, Almo SC, Minor W. X-ray crystallography over the past decade for novel drug discovery – where are we heading next? *Expert Opin Drug Discov* 2015; **10**: 975–989.

337 Khan FI, Wei D-Q, Gu K-R, Hassan MI, Tabrez S. Current updates on computer aided protein modeling and designing. *Int J Biol Macromol* 2016; **85**: 48–62.

338 Vyas VK, Ukawala RD, Ghate M, Chintha C. Homology modeling a fast tool for drug discovery: current perspectives. *Indian J Pharm Sci* 2012; **74**: 1–17.

339 Lesk A, Chothia C. Design, construction and properties of novel protein molecules - The response of protein structures to amino-acid sequence changes. *Philos Trans R Soc London A Math Phys Eng Sci* 1986; **317**.http://rsta.royalsocietypublishing.org/content/317/1540/345 (accessed 22 Aug2017).

340 Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. *Proteins Struct Funct Genet* 2003; **53**: 352–368.

341 Holm L, Sander C. Mapping the protein universe. *Science* 1996; **273**: 595–603.

342 Flöckner H, Braxenthaler M, Lackner P, Jaritz M, Ortner M, Sippl MJ. Progress in fold recognition. *Proteins Struct Funct Genet* 1995; **23**: 376–386.

343 Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. *Proteins Struct Funct Bioinforma* 2011; **79**: 37–58.

344 Centeno NB, Planas-Iglesias J, Oliva B. Comparative modelling of protein structure and its impact on microbial cell factories. *Microb Cell Fact* 2005; **4**: 20.

345 Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. *Structure* 1997; **5**: 1093–108.

346 Holm L, Rosenstrï¿½m P. Dali server: conservation mapping in 3D. *Nucleic Acids Res* 2010; **38**: W545–W549.

347 Berman HM, Westbrook J, Feng Z *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000; **28**: 235–42.

348 Hubbard TJ, Murzin AG, Brenner SE, Chothia C. SCOP: a structural classification of proteins database. *Nucleic Acids Res* 1997; **25**: 236–9.

349 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–410.

350 Altschul SF, Madden TL, Schäffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; **25**: 3389–402.

351 Park J, Karplus K, Barrett C *et al.* Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* 1998; **284**: 1201–1210.

352 Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998; **14**: 755–63.

353 Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov Models in Computational Biology. *J Mol Biol* 1994; **235**: 1501–1531.

354 Jones DT, Taylort WR, Thornton JM. A new approach to protein fold recognition. *Nature* 1992; **358**: 86–89.

355 Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* 1998; **23**: 403–5.

356 Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of

progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994; **22**: 4673–80.

357 Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 1998; **14**: 846–56.

358 Edgar RC, Olander KS. SATCHMO: sequence alignment and tree construction using hidden Markov models. *BIOINFORMATICS* 2003; **19**: 1404–1411.

359 Simossis VA, Heringa J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res* 2005; **33**: W289-94.

360 Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 1987; **326**: 347–352.

361 Aszódi A, Taylor WR. Homology modelling by distance geometry. *Fold Des* 1996; **1**: 325–334.

362 Totrov M. Loop Simulations. In: *Methods in molecular biology (Clifton, N.J.)*. 2011, pp 207–229.

363 Krieger E, Nabuurs SB, Vriend G. Homology modeling. *Methods Biochem Anal* 2003; **44**: 509–23.

364 Mendes J, Baptista AM, Carrondo MA, Soares CM. Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins* 1999; **37**: 530–43.

365 Berjanskii M, Liang Y, Zhou J *et al.* PROSESS: a protein structure evaluation suite and server. *Nucleic Acids Res* 2010; **38**: W633–W640.

366 Kryshtafovych A, Barbato A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A. Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins* 2015. doi:10.1002/prot.24919.

367 Han R, Leo-Macias A, Zerbino D, Bastolla U, Contreras-Moreira B, Ortiz AR. An efficient conformational sampling method for homology modeling. *Proteins Struct Funct Bioinforma* 2008; **71**: 175–188.

368 Zhu J, Fan H, Periole X, Honig B, Mark AE. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins Struct Funct Bioinforma* 2008; **72**: 1171–1188.

369 van Gelder CWG, Leusen FJJ, Leunissen JAM, Noordik JH. A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. *Proteins Struct Funct Genet* 1994; **18**: 174–185.

370 Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 2005; **33**: W244–W248.

371 Rykunov D, Steinberger E, Madrid-Aliste CJ, Fiser A. Improved scoring function for comparative modeling using the M4T method. *J Struct Funct Genomics* 2009; **10**: 95–99.

372 Webb B, Sali A. Protein Structure Modeling with MODELLER. In: *Methods in molecular biology (Clifton, N.J.)*. 2014, pp 1–15.

373 Eramian D, Madhusudhan MS, Marti-Renom MA, Shen M-Y, Sali A. ModWeb version r189. https://modbase.compbio.ucsf.edu/modweb/ (accessed 24 Aug2017).

374 Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 2015; **10**: 845–858.

375 Biasini M, Bienert S, Waterhouse A *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* 2014; **42**: W252-8.

376 Das R, Baker D. Macromolecular Modeling with Rosetta. *Annu Rev Biochem* 2008; **77**: 363–382.

377 Yang J, Zhang Y. Protein Structure and Function Prediction Using I-TASSER. *Curr Protoc Bioinforma* 2015; **52**: 5.8.1-15.

378 Russel D, Lasker K, Webb B *et al.* Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLoS Biol* 2012; **10**: e1001244.

379 Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 2004; **32**: W526-31.

380 Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins Struct Funct Bioinforma* 2014; **82**: 1–6.

381 Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 2006; **22**: 195–201.

382 Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* 2009; **30**: S162–S173.

383 Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* 2009; **37**: D387-92.

384 Bienert S, Waterhouse A, de Beer TAP *et al.* The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res* 2017; **45**: D313–D319.

385 Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *Electrophoresis* 1997; **18**: 2714–2723.

386 Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 2011; **27**: 343–350.

387 Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins Struct Funct Bioinforma* 2007; **69**: 108–117.

388 Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 2007; **5**: 17.

389 Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 2010; **5**: 725–738.

390 Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct Funct Bioinforma* 2004; **57**: 702–710.

391 Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci* 2004; **101**: 7594–7599.

392 Yang J, Wang Y, Zhang Y. ResQ: An Approach to Unified Estimation of B -Factor and Residue-Specific Error in Protein Structure Prediction. *J Mol Biol* 2016; **428**: 693–701.

393 Sali A. Modeling mutations and homologous proteins. *Curr Opin Biotechnol* 1995; **6**: 437–51.

394 Fiser A, Do RKG, Šali A. Modeling of loops in protein structures. *Protein Sci* 2000; **9**: 1753–1773.

395 Vásquez M. Modeling side-chain conformation. *Curr Opin Struct Biol* 1996; **6**: 217–21.

396 Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative Protein Structure Modeling of Genes and Genomes. *Annu Rev Biophys Biomol Struct* 2000; **29**: 291–325.

397 Schmidt T, Bergner A, Schwede T. Modelling three-dimensional protein structures for applications in drug design. *Drug Discov Today* 2014; **19**: 890–897.

398 Sievers F, Higgins DG. Clustal Omega. *Curr Protoc Bioinforma* 2014; **48**: 3.13.1-16.

399 Dong R, Peng Z, Zhang Y, Yang J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* 2018; **34**: 1719–1725.

400 Biasini M, Bienert S, Waterhouse A *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* 2014; **42**: W252-8.

401 Rydel TJ, Williams JM, Krieger E *et al.* The crystal structure, mutagenesis, and activity studies reveal that patatin is a lipid acyl hydrolase with a Ser-Asp catalytic dyad. *Biochemistry* 2003; **42**: 6696–708.

402 Ku B, Lee K-H, Park WS *et al.* VipD of Legionella pneumophila targets activated Rab5 and Rab22 to interfere with endosomal trafficking in macrophages. *PLoS Pathog* 2012; **8**: e1003082.

403 Halavaty AS, Borek D, Tyson GH *et al.* Structure of the type III secretion effector protein ExoU in complex with its chaperone SpcU. *PLoS One* 2012; **7**: e49388.

404 Lucas M, Gaspar AH, Pallara C *et al.* Structural basis for the recruitment and activation of the Legionella phospholipase VipD by the host GTPase Rab5. *Proc Natl Acad Sci U S A* 2014; **111**: E3514-23.

405 Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 2014; **12**: 7–8.

406 Gendrin C, Contreras-Martel C, Bouillot S *et al.* Structural basis of cytotoxicity mediated by the type III secretion toxin ExoU from Pseudomonas aeruginosa. *PLoS Pathog* 2012; **8**: e1002637.

407 da Mata Madeira PV, Zouhir S, Basso P *et al.* Structural Basis of Lipid Targeting and Destruction by the Type V Secretion System of Pseudomonas aeruginosa. *J Mol Biol* 2016; **428**: 1790–803.

408 Wijeyesakere SJ, Richardson RJ, Stuckey JA. Crystal structure of patatin-17 in complex with aged and non-aged organophosphorus compounds. *PLoS One* 2014; **9**: e108245.

409 Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *J Mol Graph* 1996; **14**: 33–38.

410 Stivala A, Wybrow M, Wirth A, Whisstock JC, Stuckey PJ. Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics* 2011; **27**: 3315–3316.

411 Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* 2004; **32**: W500–W502.

412 Xin Y-N, Zhao Y, Lin Z-H, Jiang X, Xuan S-Y, Huang J. Molecular dynamics simulation of PNPLA3 I148M polymorphism reveals reduced substrate access to the catalytic cavity. *Proteins* 2013; **81**: 406–14.

413 Burke JE, Dennis EA. Phospholipase A2 structure/function, mechanism, and signaling. *J Lipid Res* 2009; **50 Suppl**: S237-42.

414 Gaspar AH, Machner MP. VipD is a Rab5-activated phospholipase A1 that protects Legionella pneumophila from endosomal fusion. *Proc Natl Acad Sci* 2014; **111**: 4560–4565.

415 Anderson DM, Schmalzer KM, Sato H *et al.* Ubiquitin and ubiquitin-modified proteins activate the Pseudomonas aeruginosa T3SS cytotoxin, ExoU. *Mol Microbiol* 2011; **82**: 1454–1467.

416 Hofer P, Boeszoermenyi A, Jaeger D *et al.* Fatty Acid-binding Proteins Interact with Comparative Gene Identification-58 Linking Lipolysis with Lipid Ligand Shuttling. *J Biol Chem* 2015; **290**: 18438–53.

417 Finck-Barbançon V, Yahr TL, Frank DW. Identification and characterization of SpcU, a chaperone required for efficient secretion of the ExoU cytotoxin. *J Bacteriol* 1998; **180**: 6224–31.

418 Valley CC, Cembran A, Perlmutter JD *et al.* The methionine-aromatic motif plays a unique role in stabilizing protein structure. *J Biol Chem* 2012; **287**: 34979–91.

419 Mcauley M, Timson DJ. Modulating Mobility: a Paradigm for Protein Engineering? *Appl Biochem Biotechnol* 2010. doi:10.1007/s12010-016-2200-y.

420 Boehr DD, McElheny D, Dyson HJ, Wright PE. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science (80-)* 2006; **313**: 1638–1642.

421 Loria JP, Berlow RB, Watt ED. Characterization of Enzyme Motions by Solution NMR Relaxation Dispersion. *Acc Chem Res* 2008; **41**: 214–221.

422 Watt ED, Shimada H, Kovrigin EL, Loria JP. The mechanism of rate-limiting motions in enzyme function. *Proc Natl Acad Sci* 2007; **104**: 11981–11986.

423 Henzler-Wildman K, Kern D. Dynamic personalities of proteins. *Nature* 2007; **450**: 964–972.

424 Goodey NM, Benkovic SJ. Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol* 2008; **4**: 474–482.

425 Ricketson D, Hostick U, Fang L, Yamamoto KR, Darimont BD. A Conformational Switch in the Ligand-binding Domain Regulates the Dependence of the Glucocorticoid Receptor on Hsp90. *J Mol Biol* 2007; **368**: 729–741.

426 and RC, Loria* JP. Evidence for Flexibility in the Function of Ribonuclease A†. 2002. doi:10.1021/BI025655M.

427 Karplus M, Kuriyan J. Molecular dynamics and protein function. *Proc Natl Acad Sci* 2005; **102**: 6679–6685.

428 Khan FI, Bisetty K, Singh S, Permaul K, Hassan MI. Chitinase from Thermomyces lanuginosus SSBP and its biotechnological applications. *Extremophiles* 2015; **19**: 1055–1066.

429 Khan FI, Shahbaaz M, Bisetty K *et al.* Large scale analysis of the mutational landscape in ?-glucuronidase: A major player of mucopolysaccharidosis type VII. *Gene* 2016; **576**:

36–44.

430 Gramany V, Khan FI, Govender A, Bisetty K, Singh S, Permaul K. Cloning, expression, and molecular dynamics simulations of a xylosidase obtained from *Thermomyces lanuginosus*. *J Biomol Struct Dyn* 2016; **34**: 1681–1692.

431 Jones DT. Protein structure prediction in the postgenomic era. *Curr Opin Struct Biol* 2000; **10**: 371–379.

432 Lupala CS, Rasaeifar B, Gomez-Gutierrez P, Perez JJ. Using molecular dynamics for the refinement of atomistic models of GPCRs by homology modeling. *J Biomol Struct Dyn* 2017; : 1–13.

433 Karplus M, Petsko GA. Molecular dynamics simulations in biology. *Nature* 1990; **347**: 631–639.

434 From the Schrödinger Equation to Molecular Dynamics. In: *Numerical Simulation in Molecular Dynamics*. Springer Berlin Heidelberg: Berlin, Heidelberg, 2007, pp 17–36.

435 Liu H, Elstner M, Kaxiras E, Frauenheim T, Hermans J, Yang W. Quantum mechanics simulation of protein dynamics on long timescale. *Proteins Struct Funct Genet* 2001; **44**: 484–489.

436 Groenhof G. Introduction to QM/MM Simulations. Humana Press, Totowa, NJ, 2013, pp 43–66.

437 Warshel A, Levitt M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol* 1976; **103**: 227–249.

438 Paquet E, Viktor HL. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review. *Biomed Res Int* 2015; **2015**: 1–18.

439 Lippert RA, Bowers KJ, Dror RO *et al.* A common, avoidable source of error in molecular dynamics integrators. *J Chem Phys* 2007; **126**: 46101.

440 Murdock SE, Tai K, Ng H *et al.* Quality assurance for biomolecular simulations. *J Chem Theory Comput* 2006; **2**: 1477–1481.

441 Vanommeslaeghe K, Guvench O, MacKerell AD, Jr. Molecular mechanics. *Curr Pharm Des* 2014; **20**: 3281–92.

442 Wang J, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 2006; **25**: 247–260.

443 Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. *J Comput Chem* 2004; **25**: 1157–1174.

444 Vanommeslaeghe K, Raman EP, MacKerell AD. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J Chem Inf Model* 2012; **52**: 3155–3168.

445 Zoete V, Cuendet MA, Grosdidier A, Michielin O. SwissParam: A fast force field generation tool for small organic molecules. *J Comput Chem* 2011; **32**: 2359–2368.

446 Case DA, Cerutti DS, Cheatham TE *et al.* AMBER 2017. 2017.

447 Chasman DI. *Protein structure : determination, analysis and applications for drug discovery*. Marcel Dekker, 2003.

448 Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: A web-based graphical user interface for

CHARMM. *J Comput Chem* 2008; **29**: 1859–1865.

449 Phillips JC, Braun R, Wang W *et al.* Scalable molecular dynamics with NAMD. *J Comput Chem* 2005; **26**: 1781–1802.

450 Pronk S, Páll S, Schulz R *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 2013; **29**: 845–854.

451 Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* 2005; **33**: W368–W371.

452 Myers J, Grothaus G, Narayanan S, Onufriev A. A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins Struct Funct Bioinforma* 2006; **63**: 928–938.

453 Anandakrishnan R, Aguilar B, Onufriev A V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res* 2012; **40**: W537-41.

454 Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 2015; **11**: 3696–3713.

455 Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* 2013; **9**: 3878–3888.

456 Le Grand S, Götz AW, Walker RC. SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput Phys Commun* 2013; **184**: 374–380.

457 Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* 2013; **9**: 3084–3095.

458 Xie Y (Mathematician). *Dynamic documents with R and knitr*. .

459 Wickham H. *Ggplot2 : elegant graphics for data analysis*. Springer, 2009.

460 Team RC. R: A language and environment for statistical computing. 2017.https://www.r-project.org.

461 Team Rs. RStudio: Integrated development for R. 2015.http://www.rstudio.com.

462 Sehnal D, Svobodová Vařeková R, Berka K *et al.* MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *J Cheminform* 2013; **5**: 39.

463 Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010; **31**: 455–61.

464 Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 1999; **293**: 321–331.

465 Wijeyesakere SJ, Richardson RJ, Stuckey JA. Modeling the tertiary structure of the patatin domain of neuropathy target esterase. *Protein J* 2007; **26**: 165–72.

466 Dessen A, Tang J, Schmidt H *et al.* Crystal structure of human cytosolic phospholipase A2 reveals a novel topology and catalytic mechanism. *Cell* 1999; **97**: 349–60.

467 BasuRay S, Smagris E, Cohen JC, Hobbs HH. The PNPLA3 variant associated with fatty

liver disease (I148M) accumulates on lipid droplets by evading ubiquitylation. *Hepatology* 2017; **66**: 1111–1124.

468 Goo Y-H, Son S-H, Paul A. Lipid Droplet-Associated Hydrolase Promotes Lipid Droplet Fusion and Enhances ATGL Degradation and Triglyceride Accumulation. *Sci Rep* 2017; **7**: 2743.

469 Bersuker K, Olzmann JA. Establishing the lipid droplet proteome: Mechanisms of lipid droplet protein targeting and degradation. *Biochim Biophys Acta - Mol Cell Biol Lipids* 2017; **1862**: 1166–1177.

470 Yang X, Heckmann BL, Zhang X, Smas CM, Liu J. Distinct Mechanisms Regulate ATGL-Mediated Adipocyte Lipolysis by Lipid Droplet Coat Proteins. *Mol Endocrinol* 2013; **27**: 116–126.

471 Heckmann BL, Zhang X, Saarinen AM, Liu J. Regulation of G0/G1 Switch Gene 2 (G0S2) Protein Ubiquitination and Stability by Triglyceride Accumulation and ATGL Interaction. *PLoS One* 2016; **11**: e0156742.

472 Donati B, Motta BM, Pingitore P *et al.* The rs2294918 E434K variant modulates patatin-like phospholipase domain- containing 3 expression and liver damage. *Hepatology* 2016; **63**: 787–798.

473 Liu W, Anstee QM, Wang X *et al.* Transcriptional regulation of PNPLA3 and its impact on susceptibility to nonalcoholic fatty liver Disease (NAFLD) in humans. *Aging (Albany NY)* 2016; **9**: 26–40.

474 Sheron N. Calling time. The Nation's drinking as a major health issue. *Acad Med Sci* 2004.https://eprints.soton.ac.uk/194673/1/Calling_Time.pdf (accessed 5 Jun2018).

475 NCD Risk Factor Collaboration (NCD-RisC) L, Abdeen ZA, Hamid ZA *et al.* Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *Lancet (London, England)* 2017; **390**: 2627–2642.

476 Krawczyk M, Portincasa P, Lammert F. PNPLA3-associated steatohepatitis: toward a gene-based classification of fatty liver disease. *Semin Liver Dis* 2013; **33**: 369–79.

477 Meyer K, Selbach M. Quantitative affinity purification mass spectrometry: a versatile technology to study protein-protein interactions. *Front Genet* 2015; **6**: 237.

478 Wang Y, Kory N, BasuRay S, Cohen JC, Hobbs HH. PNPLA3, CGI-58, and Inhibition of Hepatic Triglyceride Hydrolysis in Mice. *Hepatology* 2019; **69**: hep.30583.

479 BasuRay S, Wang Y, Smagris E, Cohen JC, Hobbs HH. Accumulation of PNPLA3 on lipid droplets is the basis of associated hepatic steatosis. *Proc Natl Acad Sci U S A* 2019; **116**: 9521–9526.

480 Granneman JG, Moore H-PH, Krishnamoorthy R, Rathod M. Perilipin Controls Lipolysis by Regulating the Interactions of AB-hydrolase Containing 5 (Abhd5) and Adipose Triglyceride Lipase (Atgl). *J Biol Chem* 2009; **284**: 34538–34544.

481 Paul SM, Mytelka DS, Dunwiddie CT *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010; **9**: 203–214.

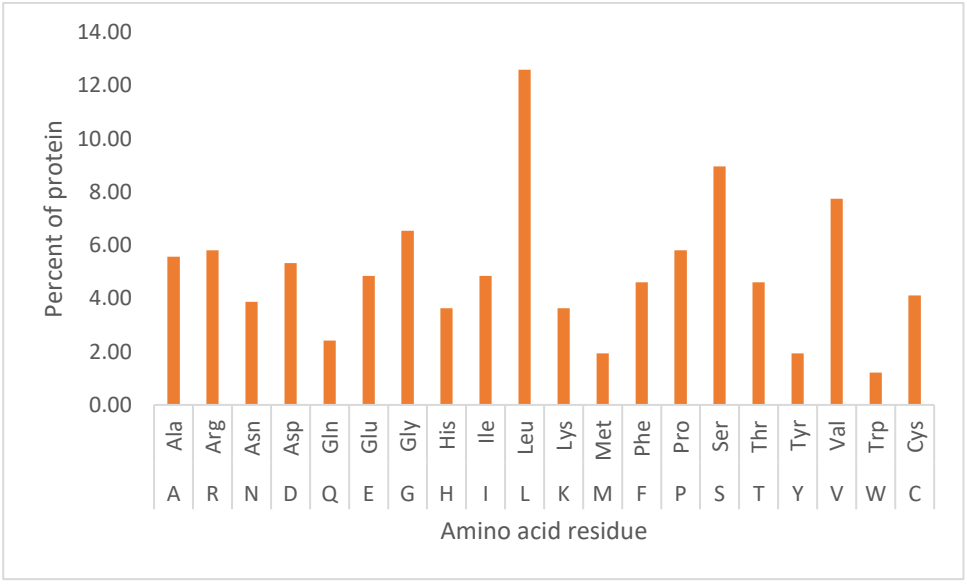
482 Segers K, Sperandio O, Sack M *et al.* Design of protein membrane interaction inhibitors by virtual ligand screening, proof of concept with the C2 domain of factor V. *Proc Natl Acad Sci U S A* 2007; **104**: 12697–702.

Appendices

Appendix I

Murine protein sequence:

```
>sp|Q91WW7|PLPL3_MOUSE Patatin-like phospholipase domain-containing
protein 3 OS=Mus musculus GN=Pnpla3 PE=2 SV=1
MYDPERRWSLSFAGCGFLGFYHVGATLCLSERAPHLLRDARTFFGCSAGALHAVTFVCSL
PLGRIMEILMDLVRKARSRNIGTLHPFFNINKCIRDGLQESLPDNVHQVISGKVHISLTR
VSDGENVLVSEFHSKDEVVDALVCSCFIPLFSGLIPPSFRGERYVDGGVSDNVPVLDAKT
TITVSPFYGEHDICPKVKSTNFFHVNITNLSLRLCTGNLQLLTRALFPSDVKVMGELCYQ
GYLDAFRFLEENGICNGPQRSLSLSLVAPEACLENGKLVGDKVPVSLCFTDENIWETLSP
ELSTALSEAIKDREGYLSKVCNLLPVRILSYIMLPCSLPVESAIAAVHRLVTWLPDIQDD
IQWLQWATSQVCARMTMCLLPSTRSRASKDDHRMLKHGHHPSPHKPQGNSAGL
```



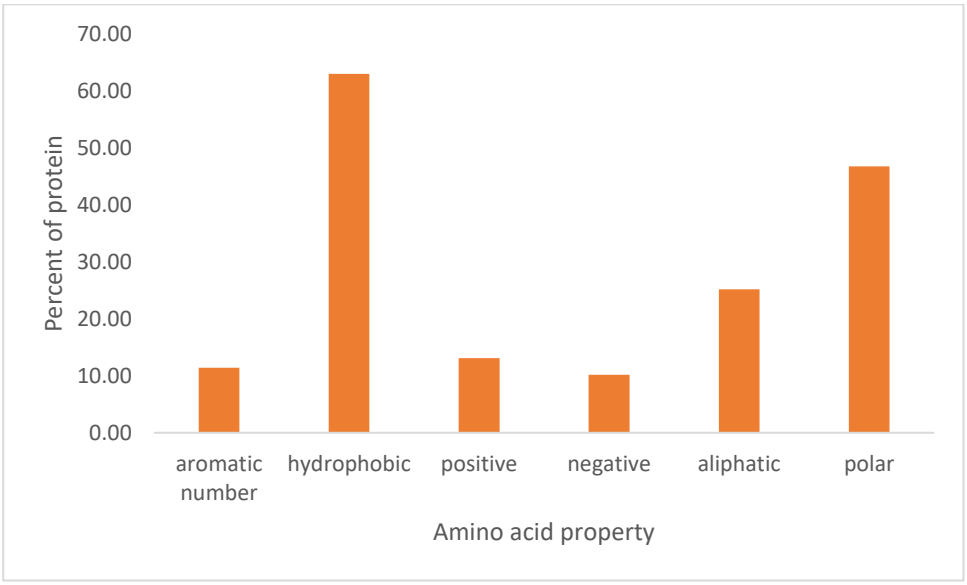
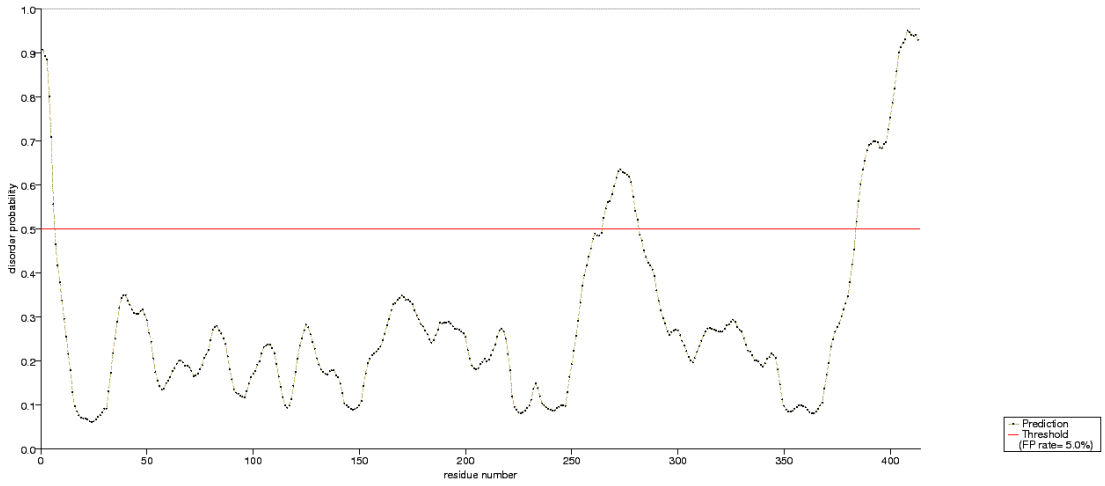


Figure A2.1 The amino acid profile of pnpla3. **Top panel)** percentage abundance of each amino acid in the protein. **Bottom panel)** The percentage abundance of each amino acid property in the protein.



1	**MYDPER**RWSL	SFAGCGFLGF	YHVGATLCLS	ERAPHLLRDA	RTFFGCSAGA	50
51	LHAVTFVCSL	PLGRIMEILM	DLVRKARSRN	IGTLHPFFNI	NKCIRDGLQE	100
101	SLPDNVHQVI	SGKVHISLTR	VSDGENVLVS	EFHSKDEVVD	ALVCSCFIPL	150
151	FSGLIPPSFR	GERYVDGGVS	DNVPVLDAKT	TITVSPFYGE	HDICPKVKST	200
201	NFFHVNITNL	SLRLCTGNLQ	LLTRALFPSD	VKVMGELCYQ	GYLDAFRFLE	250
251	ENGICNGPQR	SLSL**SLVAPE**	**ACLENGKLVG**	**D**KVPVSLCFT	DENIWETLSP	300
301	ELSTALSEAI	KDREGYLSKV	CNLLPVRILS	YIMLPCSLPV	ESAIAAVHRL	350
351	VTWLPDIQDD	IQWLQWATSQ	VCARMTMCLL	PST**RSRASKD**	**DHRMLKHGHH**	400
401	**PSPHKPQGNS**	**AGL**				450

Figure A2.2 Intrinsic disorder prediction of pnpla3 using PRDOS. **Top panel)** disorder probability represented as a dotted line on the graph. **Bottom panel)** amino acid sequence of pnpla3, with regions of disorder over 0.5 confidence.

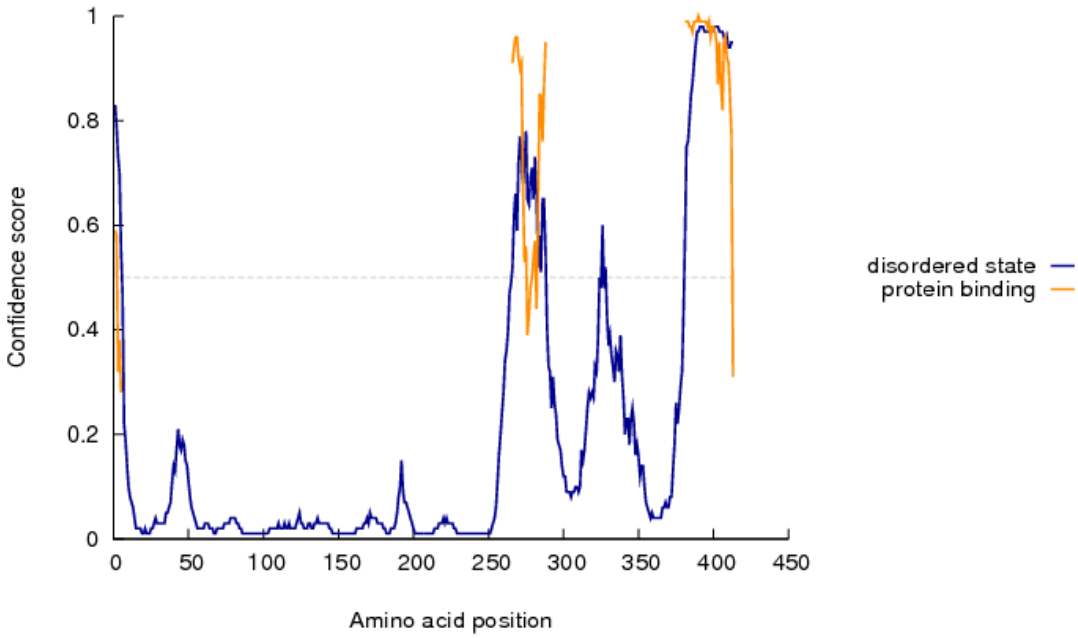


Figure A2.3 Intrinsic disorder prediction of pnpla3 using DISopred. Regions predicting disordered state are represented by the blue line and protein binding with a yellow line.

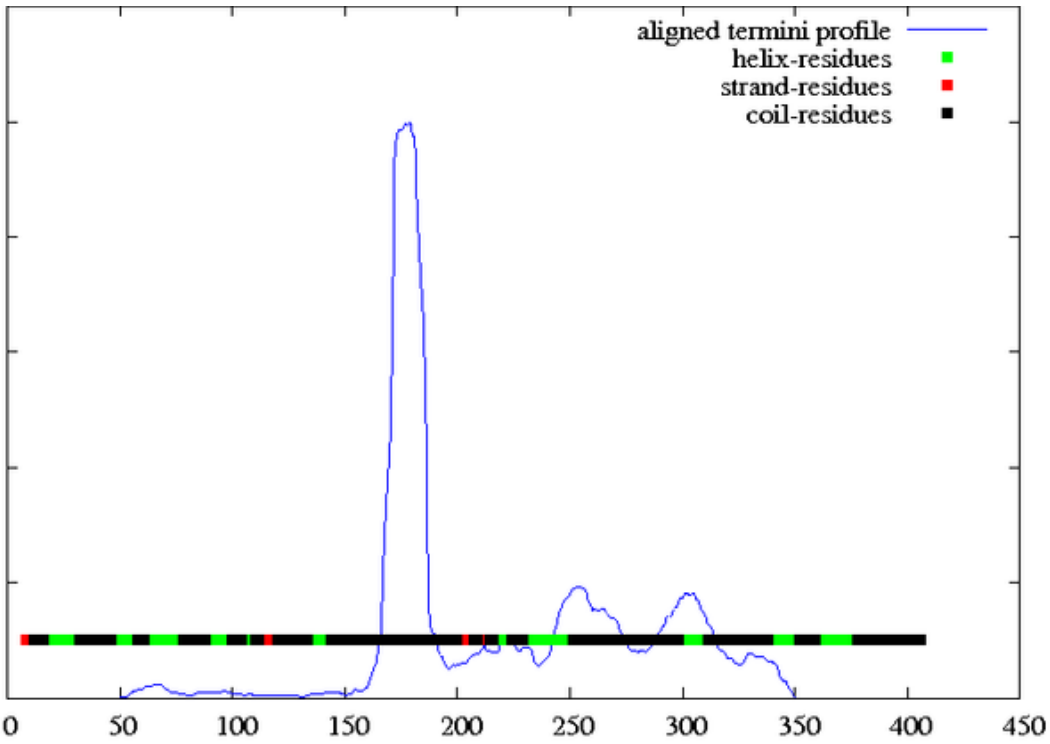


Figure A2.4 Domain boundary prediction of pnpla3 using DomPred. Predicted boundaries represented by the blue line, with the vertical axis representing confidence in the prediction. The secondary structure prediction is represented by the bar along the x-axis, with α -helices represented in green, beta-sheets represented in red and random coil represented in black.

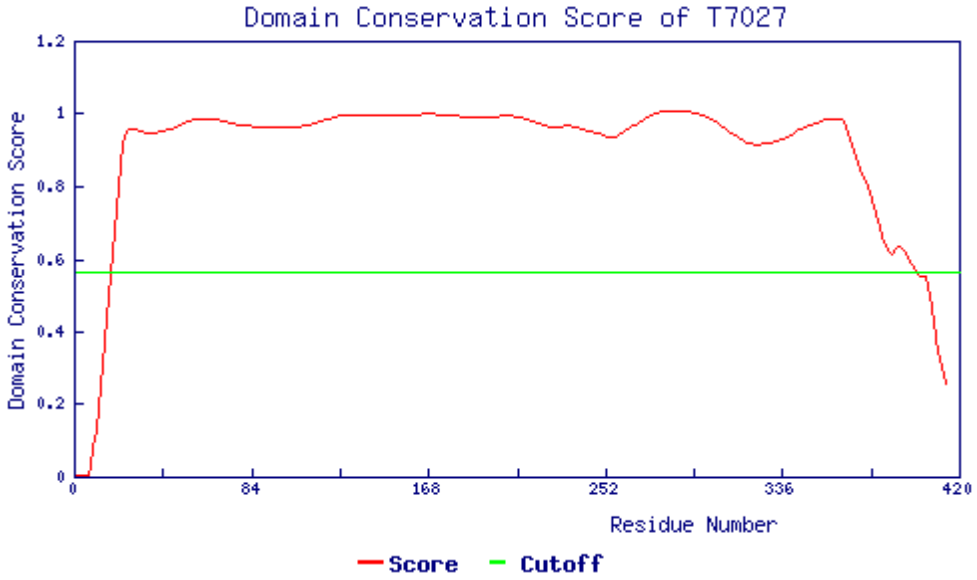


Figure A2.5 Domain boundary prediction of pnpla3 using Threadom. Continued domain conservation score shown in red, cut-off value highlighted in green.

Table A2.1 Potential pnpla3 domain boundaries as determined by multiple protein alignments using DOMSSEA.

DomSSEA Results

Score	Match	No. Doms	Boundaries
0.6884354	1ouqB	2	105
0.6829932	1q3vB	2	105
0.68123394	1pp9C	2	299
0.6801517	1ntzC	2	299
0.6798365	1nzbE	2	105
0.6759494	1101C	2	299
0.67424244	2bccC	2	299
0.6736215	1oiyD	2	201
0.67171717	1bccC	2	299

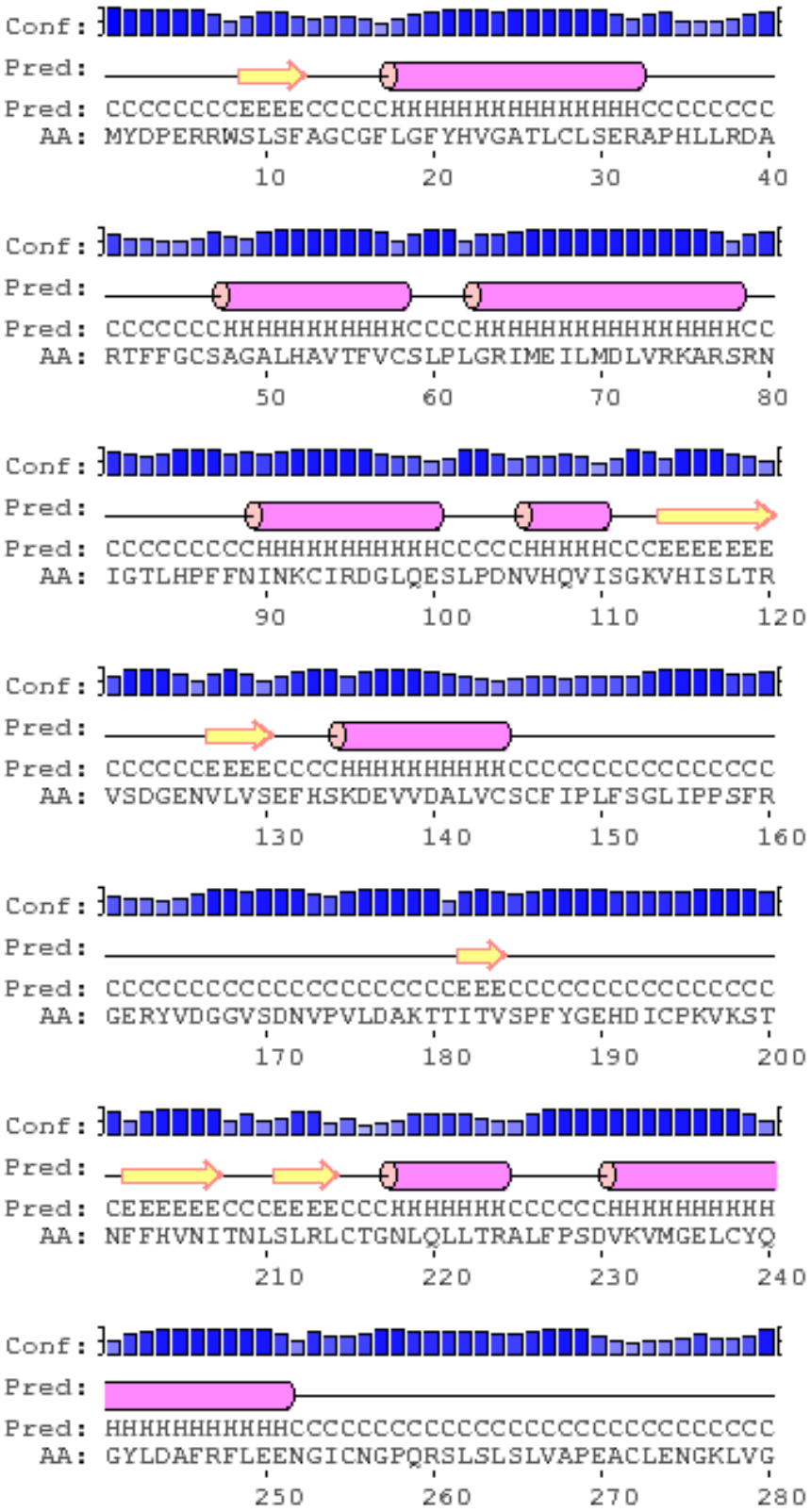


Figure A2.6 Secondary structure prediction using Psipred (Part 1). β-strands represented as yellow arrows, α-helices as pink cylinders and random coil by solid black line. The single letter amino acid code at each residue is written below the prediction, and the confidence at each point is highlighted by bar chart at the top of each line.

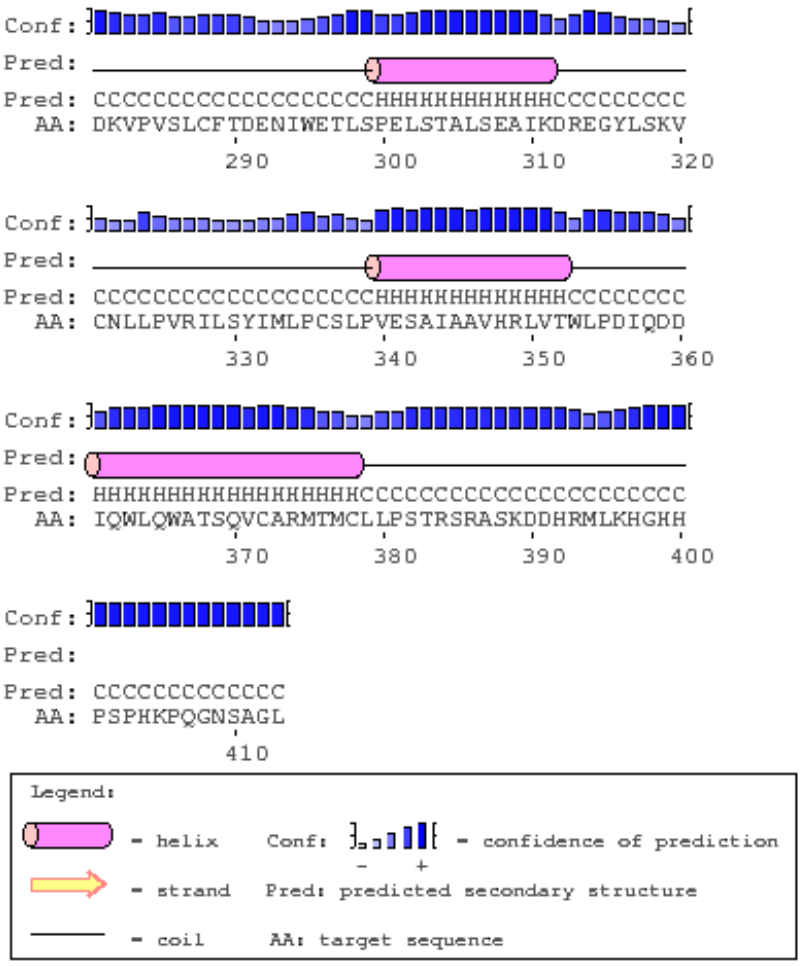


Figure A2.7 Secondary structure prediction using Psipred (Part 2). β-strands represented as yellow arrows, α-helices as pink cylinders and random coil by solid black line. The single letter amino acid code at each residue is written below the prediction, and the confidence at each point is highlighted by bar chart at the top of each line.

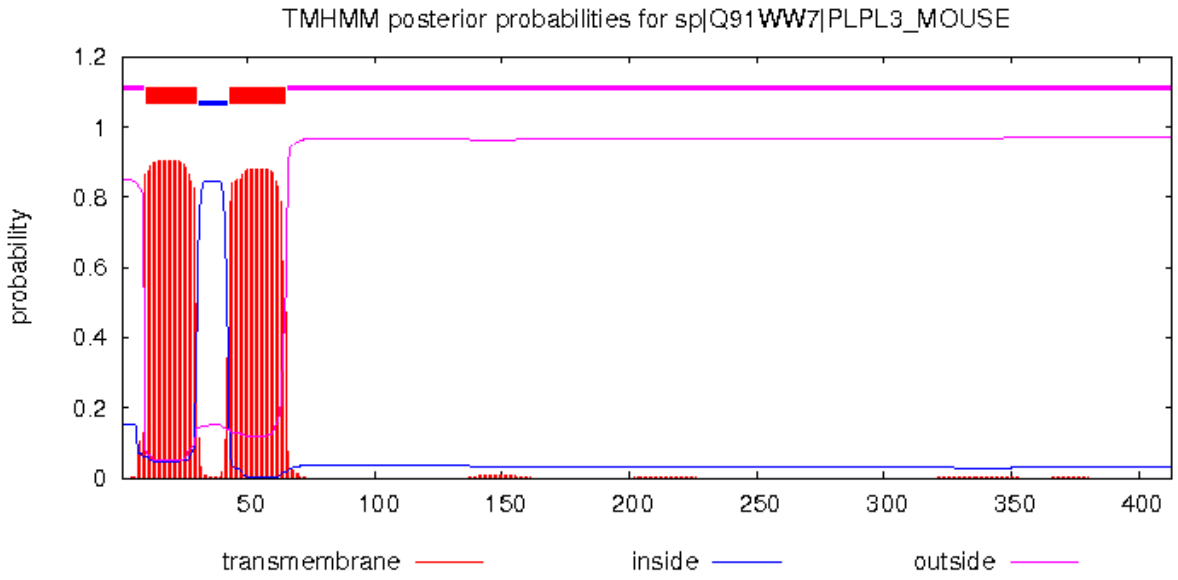


Figure A2.8 pnpla3 transmembrane helix prediction with TMHMM. Probability of cytoplasmic residues in pink and non-cytoplasmic in blue. Potential transmembrane regions highlighted with red bars.

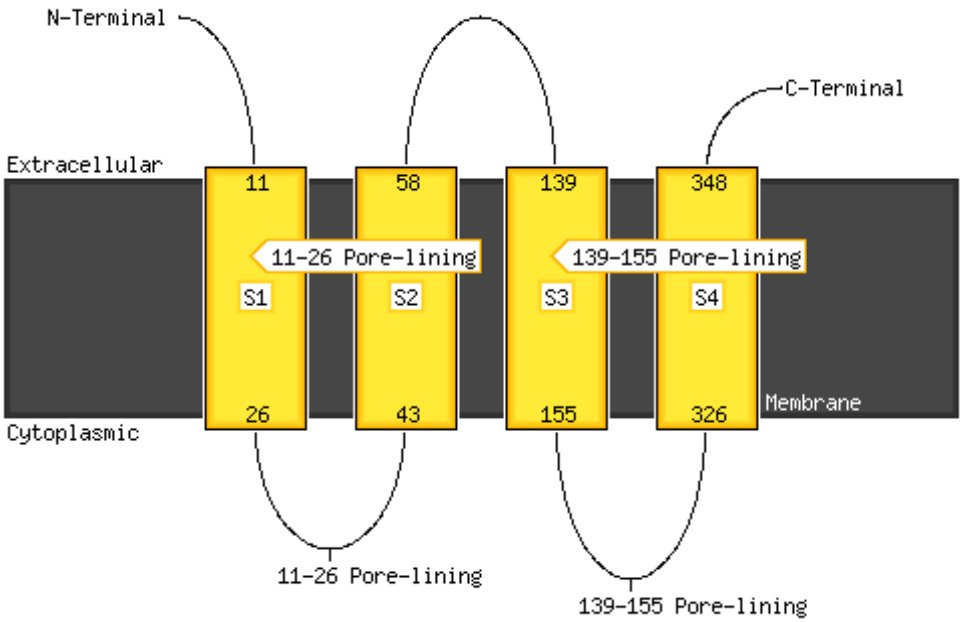


Figure A2.9 pnpla3 transmembrane prediction by Memsat-svm. Yellow squares represent predicted trans-membrane helices. The membrane depicted in dark grey, with the extracellular region depicted above the membrane and the cytoplasmic region below. Numbers describe residue at boundary of the predicted membrane helices.

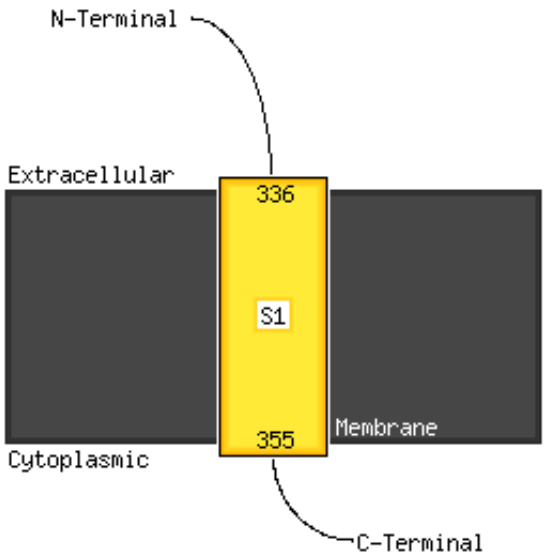


Figure A2.10 pnpla3 transmembrane prediction by Memsat3. Yellow squares represent predicted trans-membrane helices. The membrane depicted in dark grey, with the extracellular region depicted above the membrane and the cytoplasmic region below. Numbers describe residue at boundary of the predicted membrane helices.

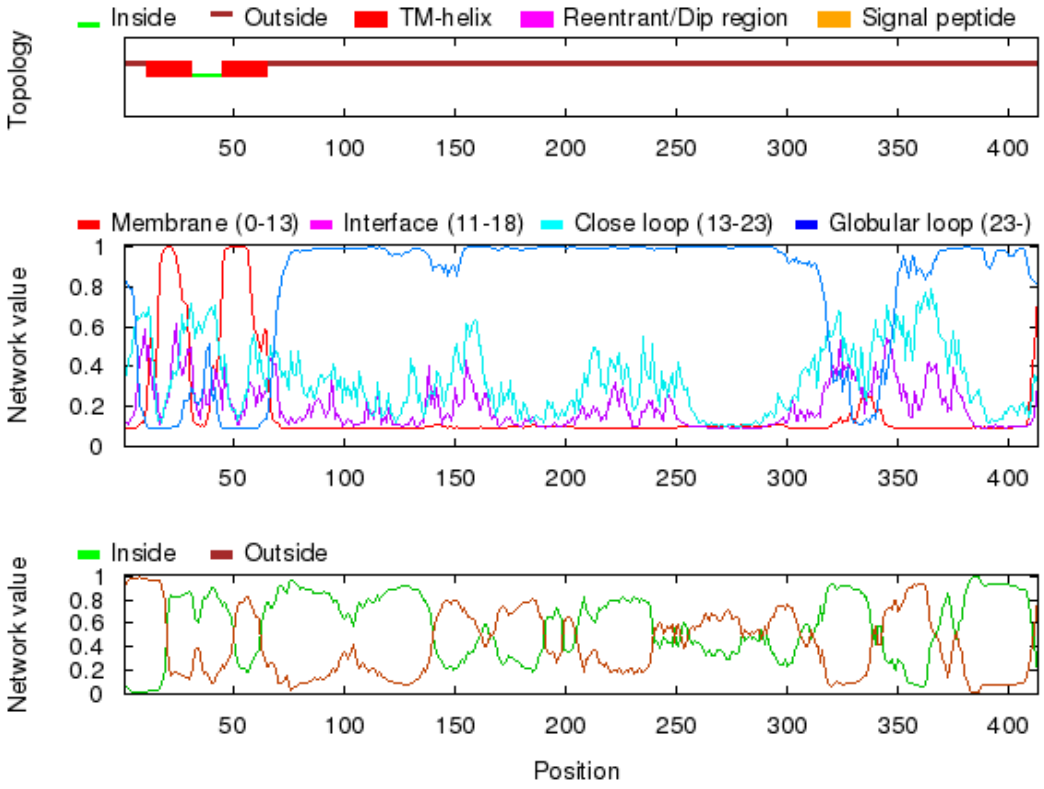


Figure A2.11 pnpla3 Transmembrane prediction with Spoctopus. **Topology output)** Graphic representation of the most likely topology as predicted by OCTOPUS. **Network output)** The two diagrams show the estimated preference for each residue to be located in different structural regions. The top diagram shows the preference of being either in: the hydrophobic part of the membrane, 0-13Å from the membrane centre; the membrane water-interface, 11-18Å from the membrane centre; a close loop region, 13-23Å from the membrane centre; a globular region, further than 23Å from the membrane. The bottom diagram shows the

estimated preference of a particular residue to be located either on the inside or outside of the membrane.

Table A2.2 pnpla3 sumoylation sites predicted by Sumoplot.

No.	Pos.	Group	Score	No.	Pos.	Group	Score
1	K402	DDHRM **LKHG** HHPSP	0.73	2	K286	GKLVG **DKVP** VSLCF	0.39

```
# Name      NN-score  Odds   Weighted   Warning
#                                by prior
#
================================================================================

sp_Q91WW7_P          0.634   1.849     0.004     -

# ============================================
```

Figure A2.12 Putative secretion signals predicted by SecretomeP 2.0.

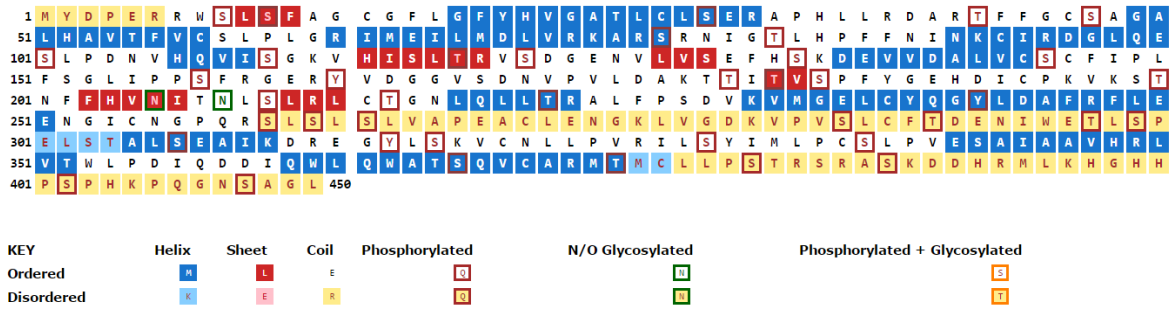


Figure A2.13 pnpla3 post-translational modification prediction with FFPred 2.0.

Table A2.3 pnpla3 biological process prediction by FFPred 2.0.

GO term	Name	Prob	SVM Reliability
GO:0019222	regulation of metabolic process	0.844	H
GO:0006810	transport	0.836	H
GO:0044281	small molecule metabolic process	0.830	H
GO:0007166	cell surface receptor signalling pathway	0.801	H
GO:0006629	lipid metabolic process	0.744	H
GO:0006796	phosphate-containing compound metabolic process	0.691	H

GO:0034220	ion transmembrane transport	0.660	H
GO:0007267	cell-cell signalling	0.660	H
GO:0009056	catabolic process	0.625	H
GO:0006066	alcohol metabolic process	0.607	H
GO:0044255	cellular lipid metabolic process	0.597	H
GO:0055085	transmembrane transport	0.558	H
GO:0030001	metal ion transport	0.541	H
GO:0050877	neurological system process	0.502	H
GO:0008152	metabolic process	0.918	L
GO:0044237	cellular metabolic process	0.917	L
GO:0050896	response to stimulus	0.911	L
GO:0007154	cell communication	0.892	L
GO:0007165	signal transduction	0.829	L
GO:0051716	cellular response to stimulus	0.823	L
GO:0023052	signalling	0.801	L
GO:0032502	developmental process	0.792	L
GO:0006807	nitrogen compound metabolic process	0.737	L
GO:0009893	positive regulation of metabolic process	0.662	L
GO:0016042	lipid catabolic process	0.659	L
GO:0007275	multicellular organismal development	0.657	L
GO:0048856	anatomical structure development	0.604	L
GO:0019538	protein metabolic process	0.578	L
GO:0009058	biosynthetic process	0.575	L
GO:0009966	regulation of signal transduction	0.560	L
GO:0030154	cell differentiation	0.540	L
GO:0031325	positive regulation of cellular metabolic process	0.509	L
GO:0010033	response to organic substance	0.506	L

Table A2.4 pnpla3 molecular function prediction by FFPred 2.0**.**

GO term	Name	Prob	SVM Reliability
GO:0003824	catalytic activity	0.913	H
GO:0001664	G-protein coupled receptor binding	0.804	H
GO:0016740	transferase activity	0.734	H
GO:0015267	channel activity	0.682	H
GO:0004872	receptor activity	0.667	H
GO:0038023	signalling receptor activity	0.665	H
GO:0015075	ion transmembrane transporter activity	0.653	H
GO:0004871	signal transducer activity	0.645	H
GO:0005216	ion channel activity	0.611	H
GO:0016788	hydrolase activity, acting on ester bonds	0.585	H
GO:0005215	transporter activity	0.557	H

GO:0022857	transmembrane transporter activity	0.548	H
GO:0005524	ATP binding	0.546	H
GO:0022891	substrate-specific transmembrane transporter activity	0.538	H
GO:0030554	adenyl nucleotide binding	0.536	H
GO:0005509	calcium ion binding	0.510	H
GO:0016747	transferase activity, transferring acyl groups other than amino-acyl groups	0.501	H
GO:0016787	hydrolase activity	0.812	L
GO:0097159	organic cyclic compound binding	0.754	L
GO:0036094	small molecule binding	0.585	L
GO:0019904	protein domain specific binding	0.564	L
GO:0005102	receptor binding	0.563	L
GO:0043169	cation binding	0.544	L

Table A2.5 pnpla3 cellular component prediction by FFPred 2.0.

GO term	Name	Prob	SVM Reliability
GO:0016021	integral component of membrane	0.993	H
GO:0016020	membrane	0.983	H
GO:0005887	integral component of plasma membrane	0.977	H
GO:0031224	intrinsic component of membrane	0.973	H
GO:0031226	intrinsic component of plasma membrane	0.904	H
GO:0005886	plasma membrane	0.900	H
GO:0098588	bounding membrane of organelle	0.862	H
GO:1902495	transmembrane transporter complex	0.770	H
GO:0012505	endomembrane system	0.766	H
GO:0031090	organelle membrane	0.750	H
GO:0005789	endoplasmic reticulum membrane	0.728	H
GO:0071944	cell periphery	0.720	H
GO:0034702	ion channel complex	0.706	H
GO:0005783	endoplasmic reticulum	0.686	H
GO:0042175	nuclear outer membrane-endoplasmic reticulum membrane network	0.622	H
GO:0005739	mitochondrion	0.500	H
GO:0005737	cytoplasm	0.916	L
GO:0043229	intracellular organelle	0.868	L
GO:0043231	intracellular membrane-bounded organelle	0.845	L
GO:0043234	protein complex	0.596	L
GO:0032991	macromolecular complex	0.535	L

Appendix II

Table A3.1 pOPIN suite protein vectors chosen for PNPLA3 expression (Adapted from Way 2016).[313]

Vector	Fusion tag	Restriction site for linearization	Forward primer extension	Reverse primer extension
pOPINE	POI-3C-HIS	NcoI, PmeI	AGGAGATATACCATG	GTGATGGTGATGTTT
pOPINE-3C-eGFP	POI-3C- eGFP-3C-HIS	NcoI, PmeI	AGGAGATATACCATG	CAGAACTTCCAGTTT
pOPINE-3C-HALO7	POI-3C- Halo7-HIS	NcoI, PmeI	AGGAGATATACCATG	CAGAACTTCCAGTTT
pOPINM	HIS-MBP-3C-POI	KpnI, HindIII	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAAGCTTTA
pOPINO	SS-POI-HIS	KpnI, PmeI	CTACCGTAGCGCAAGCT	GTGATGGTGATGTTT
pOPINS3C	HIS-SUMO- 3C-POI	KpnI, HindIII	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAGCTTTA
pOPINTRX	HIS-TRX- 3C-POI	KpnI, HindIII	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAAGCTTTA
pOPINF	HIS-3C-POI	KpnI, HindIII	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAAGCTTTA

Fusion tags are which will be co-expressed with the recombinant protein. Restriction sites are nucleotide sequences which allow for the linearization of circular plasmid DNA sequences at a specific location. Primer extension sequences are overhanging single-stranded nucleotide sequences which are specific to the plasmid for InFusion cloning. Abbreviations: POI – Protein of interest; HIS – Polyhistidine tag; MBP – Maltose binding protein; eGFP – enhanced green fluorescent protein; TRX - thioredoxin reductase; SUMO - small ubiquitin-like modifier; 3C - Rhinovirus 3C protease site; SS – signal sequence

	1	2	3	4	5	6	7	8	9	10	11	12
A	2-481	2-250	300-400	1-481	2-250	300-400	2-481	2-250	300-400	1-481	2-250	300-400
B	2-175	300-481	271-400	1-175	300-481	271-400	2-175	300-481	271-400	1-175	300-481	271-400
C	2-250	300-400	1-481	1-250	300-400	1-481	2-250	300-400	1-481	1-250	300-400	1-481
D	GFP	271-400	1-175	300-481	271-400	1-175	GFP	271-400	1-175	300-481	271-400	1-175
E	300-400	2-481	1-250	300-400	1-481	1-250	300-400	2-481	1-250	300-400	1-481	1-250
F	271-400	2-175	300-481	271-400	1-175	300-481	271-400	2-175	300-481	271-400	1-175	300-481
G	2-481	2-250	300-400	2-481	1-250	300-400	2-481	2-250	300-400	2-481	1-250	300-400
H	2-175	300-481	271-400	2-175	300-481	271-400	2-175	300-481	271-400	2-175	300-481	271-400

pOPINE	pOPINE-3C-eGFP	pOPINE-3C-HALO7	pOPINF	pOPINM	pOPINO	pOPINS3C	pPOPINTRX

Figure A3.1 Plate layout of pOPIN-PNPLA3 constructs. The plate layout of the 96-well plate used for all high throughput plasmid construction performed at the OPPF. All conditions were performed in duplicate A1-H6 and A7-H12. A positive control construct containing a pOPINE vector expressing recombinant green fluorescent protein was added to positions D1 and D7 as a positive control for all experiments (Adapted from Way 2016).[313]

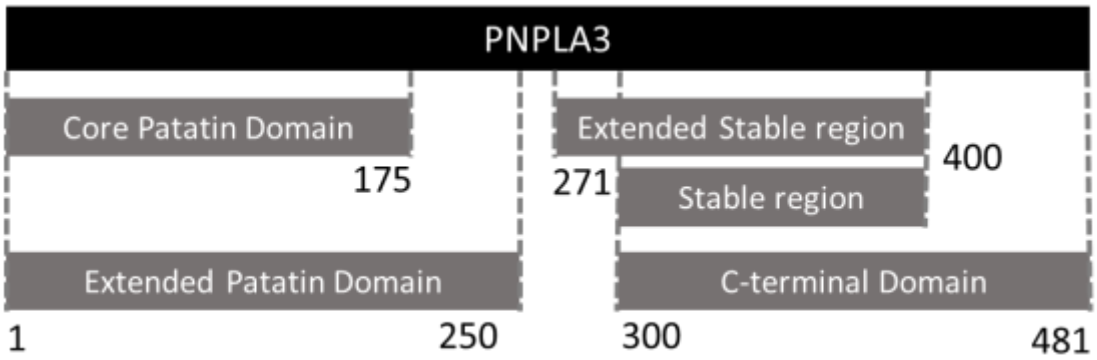


Figure A3.2 PNPLA3 inserts selected for plasmid construction and protein expression trials (Adapted from Way 2016).[313]

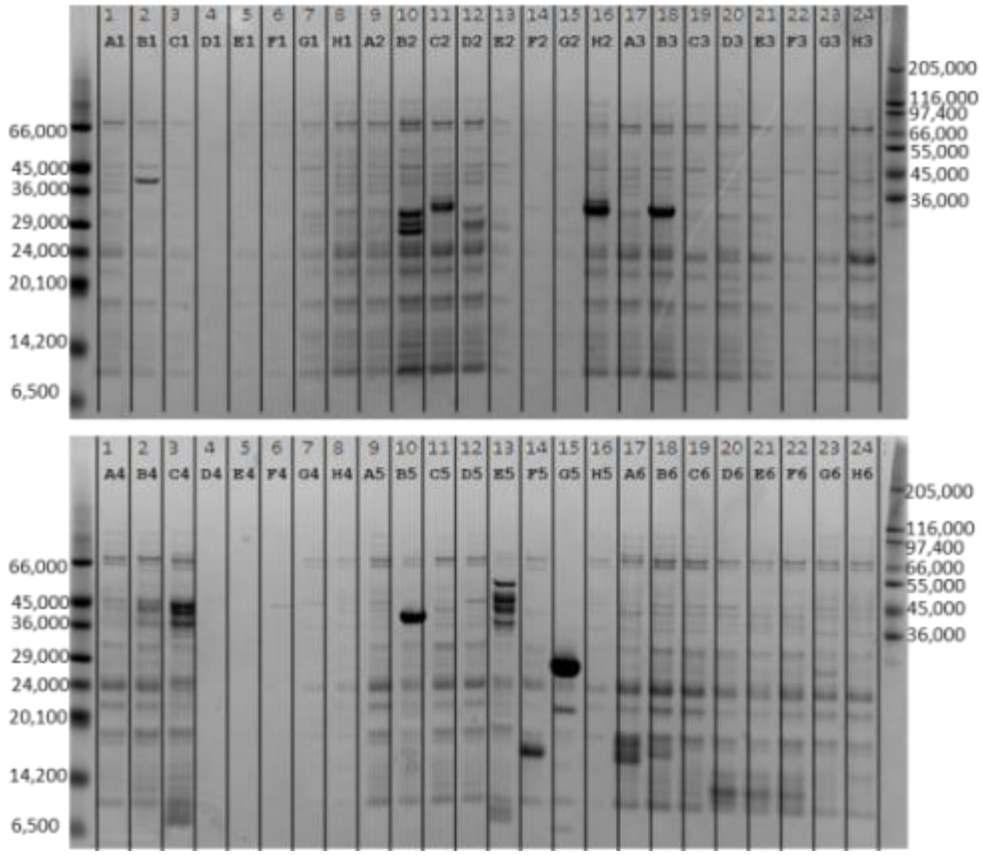


Figure A3.3 SDS-PAGE gels of affinity purified, IPTG-induced Rosetta lysates. Bands with distinct recombinant protein expression: B1 (2-275-PNPLA3-C-His), B2 (300-481-PNPLA3-GFP-C-His), C2 (300-400-PNPLA3-GFP-C-His), H2 (300-481-PNPLA3-HALO7-C-His), B3 (271-400-PNPLA3-HALO7-C-His), C4 (N-His-MBP-1-250-PNPLA3), B5 (300-481-PNPLA3-periplasmic signal sequence-C-His), E5 (N-His-MBP-1-481-PNPLA3), F5 (N-His-SUMO-1-175-PNPLA3) and G5 (GFP positive control) (Adapted from Way 2016).[313]

Plasmid DNA transformation into *E. coli* protocol (Novagen)

1. Thaw the required number of tubes of cells on ice and mix gently to ensure that the cells are evenly suspended.
2. Place the required number of 1.5-ml polypropylene microcentrifuge tubes on ice to pre-chill. Pipet 20 µl aliquots of cells into the pre-chilled tubes.
3. Add 1 µl of the DNA solution directly to the cells. Stir gently to mix.
4. Place the tubes on ice for 5 min.
5. Heat the tubes for exactly 30 s in a 42°C water bath; do not shake.
6. Place on ice for 2 min.
7. Add 80 µl of room temperature SOC Medium to each tube.
8. Incubate at 37°C while shaking at 250 rpm for 60 min prior to plating on selective medium.

Selection for transformants is accomplished by plating on media containing antibiotic for the plasmid-encoded drug resistance. Additional host-specific antibiotics may also be appropriate to insure maintenance of the host-encoded feature(s).

Spread plating protocol (microbelibrary)

Select and prepare an agar medium based upon the type of bacteria to be enumerated or selected.

After autoclaving, cool the agar to between 45°C and 50°C prior to pouring the plates to minimize the amount of condensation that forms. The thickness of the agar should be roughly 0.3 cm, which can be achieved by pouring 15 to 20 ml of media per 100 x 15 mm plate.

A convenient inoculum volume, in terms of spreading, absorption, and calculations, is 0.1 ml (100 microliters).

Since some bacteria rapidly attach to the agar surface, the inoculum should be spread soon after it is applied.

A reusable glass or metal spreader should be flame sterilized by dipping in alcohol (such as 70% isopropyl or ethanol), shaking off the excess alcohol, and igniting the residue (Fig. 2, Atlas page). The spreader is then allowed to cool.

The spreader is placed in contact with the inoculum on the surface of the plate and positioned to allow the inoculum to run evenly along the length of the spreader. Even pressure is applied to the spreader and the plate is spun, on a turntable or by hand, as illustrated in Fig. 3-7 on the Atlas page.

Avoid disturbing plates for 10 to 20 minutes after spreading. Drying time varies with the room temperature and humidity.

After the spread plates have been permitted to absorb the inocula for 10 to 20 minutes they may be inverted and incubated as desired.

Observe the plates before the colonies have had time to fully develop. Closely positioned colonies may be difficult to resolve as separate colonies later. Continue the incubation as necessary.

Miniprep of plasmid DNA protocol (Axygen)

Reagents

RNase A: 50 mg/ml. Stable at room temperature for up to 6 months. Recommend -20°C for long-term storage. If a precipitate is present, use an aliquot of Buffer S1 to resuspend and transfer to the Buffer S1 bottle.

Buffer S1: Resuspension buffer. Store at 4°C after addition of RNase A.

Buffer S2: Lysis buffer. Store at room temperature.

Buffer S3: Neutralization buffer. Store at room temperature.

Buffer W1: Wash buffer. Store at room temperature.

Buffer W2 concentrate: Desalting buffer. Before using the kit, add ethanol according to instructions on the bottle label. Either 100% or 95% denatured ethanol can be used. Store at room temperature.

Eluent: 2.5 mM Tris-HCl, pH 8.5. Store at room temperature.

Preparation before experiment

1) Before using the kit, add the RNase A to Buffer S1. Mix well and store at 4°C.

Note: If a precipitate is present, use a small volume of Buffer S1 to resuspend the RNase A and then transfer to the Buffer S1 bottle.

2) Add the volume of ethanol specified on the bottle label to the Buffer W2 concentrate and mix well. Either 100% or 95% (denatured) ethanol can be used.

3) Check Buffer S2 for precipitation before each use. If precipitation occurs, incubate at 37°C to dissolve the precipitate and then equilibrate to room temperature. After use, the bottle should be closed immediately in order to avoid neutralization of NaOH by CO2 in the air.

4) Pre-warming Eluent to 65°C may improve elution efficiency.

AxyPrep Plasmid Miniprep Spin Protocol

1. Collect 1-4 ml of overnight LB culture. Centrifuge at 12,000×g for 1 minute to pellet the bacteria. Decant or pipette off as much of the supernatant as practical.

Note: When using rich broths such as LBG or 2×YT, reduce the culture volume by half. Excessive bacteria will reduce lysis efficiency, resulting in low yield and reduced purity of the plasmid DNA. Do not exceed 2 ml of bacterial culture grown in rich broth.

2. Resuspend the bacterial pellet in 250 µl of Buffer S1 by vortexing. Please be sure that the bacteria are completely resuspended before proceeding.

Note: Be sure that RNase A has been added into Buffer S1.

3. Add 250 µl of Buffer S2, and mix by gently inverting the tube for 4-6×. Do not vortex.

Note: Vigorous shaking or vortexing will cause shearing of the bacterial genomic DNA and result in the contamination of the plasmid DNA.

Note: After use, the buffer S2 bottle should be closed immediately in order to avoid neutralization of NaOH by ambient CO2.

Note: Buffer S3 (Step 4, below) must be added within 5 minutes.

4. Add 350 µl of Buffer S3, and mix by gently inverting 6-8×. Centrifuge at 12,000×g for 10 minutes to clarify the lysate. Do not vortex.

Note: Vigorous shaking or vortexing will result in contamination with genomic DNA.

5. Place a Miniprep column into an uncapped 2 ml Microfuge tube (provided). Transfer the clarified supernatant from Step 4 into the Miniprep column. Transfer the Miniprep column and 2 ml Microfuge tube to microcentrifuge and spin at 12,000×g for 1 minute.

6. Optional step: Buffer W1 Wash Washing with Buffer W1 is required only in cases where the plasmid has been propagated in an endA+ bacterial strain. These strains often exhibit high levels of endonuclease activity which will degrade the plasmid DNA.

Proceed to Step 7 if an endA- bacterial strain is used.

Pipette 500 µl of Buffer W1 into each Miniprep column. Centrifuge at 12,000×g for 1 minute.

7. Pipette 700 µl of Buffer W2 into each Miniprep column. Centrifuge at 12,000×g for 1 minute. Note: Make sure that the volume of ethanol specified on the bottle label has been added to the Buffer W2 concentrate.

8. Optional Step: Discard the filtrate from the 2 ml Microfuge tube. Place the Miniprep column back into the 2 ml Microfuge tube. Add 700 µl of Buffer W2 to the Miniprep column and centrifuge at 12,000×g for 1 minute.

Note: Two washes with Buffer W2 are used to ensure the complete removal of salt, eliminating potential problems in subsequent enzymatic reactions.

9. Discard filtrate from the 2 ml Microfuge tube. Place the Miniprep column back into the 2 ml Microfuge tube. Centrifuge at 12,000×g for 1 minute.

10. Transfer the Miniprep column into a clean 1.5 ml Microfuge tube (provided). To elute the purified plasmid DNA, add 60~80 µl of Eluent (or deionized water) to the center of the membrane. Let it stand for 1 min at room temperature. Centrifuge at 12,000×g for 1 minute.

Sodium-dodecyl sulphate gel electrophoresis protocol

Clamp gel onto electrophoresis tank. Carefully remove binding clips and the comb from gel. Place gel / glass plate sandwich into electrophoresis core. The short glass plate should face the centre, or inside of the core. If the plate sandwich does not fit in the core, check the direction of the short glass plate and the rubber gasket at the centre of the core to make sure everything is correct. If only one gel is to be run, place the buffer damn on the other side of the core, otherwise, place the second glass plate sandwich on the other side of the core.

Place the core assembly into the running tank. Add 1X Electrophoresis buffer to the core. Buffer should be added to the top of the assembly. Add 1 – 2 inches of 1X Electrophoresis buffer to the running tank.

Rinse out wells with buffer in preparation for sample loading.

Thaw protein samples rapidly in room temp water bath. Add 1/5 vol 6X Sample Buffer to protein samples. Heat samples for 5-10 minutes in the 95C dry bath. (Optional) Spin down samples in microfuge for 5 minutes.

Load samples into wells using 200 µL pipet tip. Take care not to separate glass plates by wedging the tip too far into the well. Normally, 25 µL of sample can be loaded into each well.

Fill empty wells with 1X Sample dye. Attach electrodes so that proteins will move towards the anode (+ or red lead). Run gel at 100-200 V until dye front reaches the bottom of the gel. Running time will vary based upon percentage cross linking of gel and buffer composition.

10X Electrophoresis Buffer: 30 g Tris, 145 g Glycine, 10 g SDS, bring to 1 L with H2O.

Coomassie stain

Disconnect electrodes, dump out electrophoresis buffer, and remove gel sandwich from tank. Remove side spacers and carefully pry apart the plates so that the gel remains on one plate. Pour staining solution into pipet tip box.

Invert plate with gel into the staining solution and gently allow gel to "float" off of plate into solution. Cover and gently agitate gel on gel rocker for 15-30 minutes (longer will increase sensitivity but requires a longer destaining period)

Remove staining solution (can be reused many times) and rinse gel with ddH2O to remove excess stain. DO NOT PLACE WATER STREAM ONTO GEL ITSELF. Add water to the pipet tip box and let it gently wash over gel surface

Add destaining solution and agitate on gel rocker for 10-15 minutes. Change destaining solution and agitate until proper level of destaining is achieved (can destain overnight if needed)

Coomassie Staining Solution: 0.1% Coomassie brilliant blue R-250, 40% Methanol, 10% Acetic Acid

Destaining Solution: 40% Methanol, 10% Acetic Acid

Silver stain (Sigma Aldrich)

Preparation Instructions

The use of ultrapure water is essential for low background and high sensitivity staining.

1. Fixing solution. Add 50 ml of ethanol and 10 ml of acetic acid to 40 ml of ultrapure water.

2. 30% Ethanol solution. Add 30 ml of ethanol to 70 ml of ultrapure water.

3. Sensitizer solution. Add 1 ml of ProteoSilver Sensitizer to 99 ml of ultrapure water. The prepared solution should be used within 2 hours. A precipitate may form in the ProteoSilver Sensitizer. This precipitate will not affect the performance of the solution. Simply allow the precipitate to settle and remove 1 ml of the supernatant.

4. Silver solution. Add 1 ml of ProteoSilver Silver Solution to 99 ml of ultrapure water. The prepared solution should be used within 2 hours.

5. Developer solution. Add 5 ml ProteoSilver Developer 1 and 0.1 ml ProteoSilver Developer 2 to 95 ml of ultrapure water. The developer solution should be prepared immediately (<20 minutes) before use.

Storage/Stability: All kit components are stable at room temperature for at least 1 year.

Direct silver staining procedure

All steps should be carried out at room temperature on an orbital shaker at 60 to 70 rpm. The gel should be stained in a glass or plastic tray, which has been cleaned with detergent and rinsed thoroughly. Clean, disposable gloves should be worn and changed before each step to prevent fingerprints on the gel. The volumes indicated in this procedure are for mini gels. The volumes should be tripled for large format (13 x 16 cm) gels. The staining process may be halted at the Fixing step by leaving the gel in the Fixing solution overnight if there is not enough time to complete the staining protocol. Staining

1. Fixing - After electrophoresis of the proteins in the mini polyacrylamide gel, place the gel into a clean tray with 100 ml of the Fixing solution for 20 minutes.

Note: A clearer background can be achieved by a longer fixing time (40 minutes to overnight).

2. Ethanol wash - Decant the Fixing solution and wash the gel for 10 minutes with 100 ml of the 30% Ethanol solution.

3. Water wash – Decant the 30% Ethanol solution and wash the gel for 10 minutes with 200 ml of ultrapure water.

4. Sensitization – Decant the water and incubate the gel for 10 minutes with 100 ml of the Sensitizer solution.

5. Water wash – Decant the Sensitizer solution and wash the gel twice, each time for 10 minutes with 200 ml of ultrapure water.

6. Silver equilibration – Decant the water and equilibrate the gel for 10 minutes with 100 ml of the Silver solution.

7. Water wash – Decant the Silver solution and wash the gel for 1 to 1.5 minutes with 200 ml of ultrapure water.

Note: Washing for longer than 1.5 minutes will result in decreased sensitivity.

8. Gel development – Decant the water and develop the gel with 100 ml of the Developer solution. Development times of 3 to 7 minutes are sufficient to produce the desired staining intensity for most gels. Development times as long as 10 to 12 minutes may be required to detect bands or spots with very low protein concentrations (0.1 ng/mm2).

Note: Over development of the gel will increase the background staining.

9. Stop - Add 5 ml of the ProteoSilver Stop Solution to the developer solution to stop the developing reaction and incubate for 5 minutes. Bubbles of CO_2 gas will form in the mixture.

10. Storage – Decant the Developer/Stop solution and wash the gel for 15 minutes with 200 ml of ultrapure water. Store the gel in fresh, ultrapure water.

Double staining

Silver Staining following Coomassie Brilliant Blue Staining Double staining (Coomassie brilliant blue and Silver) can increase the detection sensitivity 2-4 fold over that observed with silver staining alone. The Coomassie brilliant blue stained gel must be destained until the background (gel with no protein) is essentially clear. After Coomassie brilliant blue destaining, begin the silver staining at the Fixing step (step 1) of the staining procedure.

Western blotting protocol

Western blots only require ~1ug total protein loaded per well/lane.

1. Prior to removing gel from plates, soak the following in Western blotting transfer buffer: 2 x blotting paper, 4 x sponges, 1 x nitrocellulose membrane.

2. After running SDS-PAGE, remove the gel from the glass plates, cut off wells and place inside the cathode core (deeper unit, denoted by -ve symbol on the outside), horizontally at the base of the unit, away from the electrodes:

-Sponge x2 (top: +ve anode)

-Blotting paper

-Nitrocellulose membrane

-Gel

-Blotting paper

-Sponge x2 (bottom: -ve cathode)

(Try not to handle nitrocellulose membrane more than necessary - wear gloves and/or use forceps.) Roll over top blotting paper with a universal/50ml tube/etc. to flatten, ensure good contact and remove air bubbles.

3. The whole sandwich should rise about 0.5 cm above the sides of the cathode base unit; if not, add more sponges until this is reached.

4. Fit the upper anode core on top of the cathode core, compressing the sandwich until they meet at the gasket.

5. Slide this into the electrode side of the tank and place the clamp behind: depress the white lever to lock into place.

6. Place the entire tank in an ice box with enough ice to surround the bottom and all sides of the unlidded unit.

7. Fill the core with just enough transfer buffer to cover the top of the sandwich (more will generate unnecessary heat).

8. Fill the outer chamber of the tank with 650 ml dH2O (this is just to cool the core as much as possible).

9. Lid the tank, ensuring the electrodes line up when pushing down.

10. Run at 100V (constant) for 1.5-2 hours (depending on the size of the protein of interest - smaller ones may pass through the membrane if over-transferred.) Before leaving, check for presence of bubbles in the buffer in the core.

11. Check every now and then that the unit is still cooled (rearrange ice if melted/not enough/etc.).

12. After transferring, remove nitrocellulose membrane carefully from the sandwich and place in a Tupperware box and block by adding just enough 5% milk/TBS-Tween (0.1%) to cover the membrane on an orbital shaker for 1 hour (although 10 mins may be enough if in a hurry).

13. Pour off blocking solution and incubate with primary antibody in 5% milk/TBS-T (usually 1 in 2000 dilution) for ~2 hours/overnight; or, if using one-step HRP-conjugated antibody, leave for at least 3 hours/overnight.

14. If using secondary antibody, pour off first antibody solution and replace with secondary antibody in 5% milk/TBS-T (usually 1 in 5000 dilution) for ~1.5 hours.

15. Wash 3 x 2 mins with TBS-T.

16. During last TBS-T wash, make up 12 ml developer solution with AEC staining kit:

    6 drops acetate buffer (vial 1)

    3 drops AEC chromogen (vial 2)

    3 drops 3% H2O2 (vial 3)

17. Pour off last TBS-T wash and replace with developer solution: incubate for 5-10 mins.

18. Stop developing by washing a few times with dH2O.

Buffers

Western Blot Transfer Buffer (10x): 25mM Tris, 192mM glycine, 0.1% SDS, pH 8.3, 30.3g Tris, 144g Glycine (no need to adjust pH: additional acid/base will increase buffer conductivity)

Western Blot Transfer Buffer (1x): 100ml 10x Transfer buffer, 200ml Methanol, 700ml ddH2O

TBS (10x): 1x = 150mM NaCl, 10mM Tris pH 8.0, 87.66g NaCl, 12.11g Tris, ~4ml HCl - to pH 8

TBS-T (1x): 100ml 10x TBS, 10ml 10% TWEEN-20, 890ml ddH2O

Blocking Buffer: 5% non-fat milk in TBS-T

Nickel affinity chromatography

1. Fill the pump tubing or syringe with distilled water.

2. Remove the stopper and connect the column to the chromatography system tubing, syringe (use the connector provided) or laboratory pump “drop-to-drop” to avoid introducing air into the system.

3. Wash out the ethanol with 3 to 5 column volumes of distilled water.

4. Equilibrate the column with at least 5 column volumes of binding buffer. In some cases, a blank run is recommended before final equilibration/ sample application.

5. Apply the unclarified lysate with a pump or a syringe. Continuous stirring of the sample during sample loading is recommended to prevent sedimentation. Typical loading volumes of unclarified lysate (highly dependent on specific sample, sample pre-treatment, and temperature at sample loading).

Note: Sample loading at 4°C may increase the viscosity of the sample. An adverse effect of increased sample viscosity is that maximum back pressure for the column is reached at a lower sample volume loading on the column. Do not exceed the binding capacity of the column. Large volumes may increase back pressure, making the use of a syringe more difficult.

6. Wash with binding buffer until the absorbance reaches a steady baseline (generally at least 10 to 15 column volumes). Note: Purification results are improved by using imidazole in sample and binding buffer.

7. Elute with elution buffer using a one-step procedure or a linear gradient. For step elution, five column volumes of elution buffer are usually sufficient. A shallow gradient, e.g., a linear gradient over 20 column volumes or more, can separate proteins with similar binding strengths.

8. Collect eluted sample from appropriate elution tubing.

9. Wash out elution buffer with 3 to 5 column volumes of distilled water.

10. Store column in 20% ethanol.

Size exclusion chromatography

1. Fill the pump tubing or syringe with distilled water.

2. Remove the stopper and connect the column to the chromatography system tubing, syringe (use the connector provided) or laboratory pump “drop-to-drop” to avoid introducing air into the system.

3. Wash out the ethanol with 3 to 5 column volumes of distilled water.

4. Equilibrate the column with at least 5 column volumes of running buffer. In some cases, a blank run is recommended before final equilibration/ sample application.

5. Apply the sample with a pump or a syringe. Typical loading volumes of unclarified lysate highly dependent upon specific column capacity.

6. Monitor UV absorbance 280 nm from the time of injection either collecting absorbance peaks manually or through the use of a fraction collector.

7. Ensure sample has been fully eluted.

8. Wash out running buffer with 3 to 5 column volumes of distilled water.

9. Store column in 20% ethanol.

Appendix III

REFERENCE information on running properties (Adapted from AmberTools 17).[446]

ntx=5,

Option to read the initial coordinates, velocities and box size from the inpcrd file. Option 1 must be used when one is starting from minimized or model-built coordinates. If an MD restrt file is used as inpcrd, then option 5 is generally used (unless you explicitly wish to ignore the velocities that are present). = 1 (default) Coordinates, but no velocities, will be read; either formatted (ASCII) files or NetCDF files can be used, as the input file type will be auto-detected. = 5 Coordinates and velocities will be read from either a NetCDF or a formatted (ASCII) coordinate file. Box information will be read if ntb > 0. The velocity information will only be used if irest = 1 (see below)

dt=0.002, - timestep in fs

ntc=2,

Flag for SHAKE to perform bond length constraints. (See also NTF in the Potential function section. In particular, typically NTF = NTC.) The SHAKE option should be used for most MD calculations. The size of the MD timestep is determined by the fastest motions in the system. SHAKE removes the bond stretching freedom, which is the fastest motion, and consequently allows a larger timestep to be used. For water models, a special "three-point" algorithm is used. Consequently, to employ TIP3P set NTF = NTC = 2. Since SHAKE is an algorithm based on dynamics, the minimizer is not aware of what SHAKE is doing; for this reason, minimizations generally should be carried out without SHAKE. One exception is short minimizations whose purpose is to remove bad contacts before dynamics can begin. For parallel versions of sander only intramolecular atoms can be constrained. Thus, such atoms must be in the same chain of the originating PDB file. = 1 SHAKE is not performed (default) = 2 bonds involving hydrogen are constrained = 3 all bonds are constrained (not available for parallel or qmmm runs in sander)

ntf=2,

ntf Force evaluation. Note: If SHAKE is used (see NTC), it is not necessary to calculate forces for the constrained bonds. = 1 complete interaction is calculated (default) = 2 bond interactions involving H-atoms omitted (use with NTC=2) = 3 all the bond interactions are omitted (use with NTC=3) = 4 angle involving H-atoms and all bonds are omitted = 5 all bond and angle interactions are omitted = 6 dihedrals involving H-atoms and all bonds and all angle interactions are omitted = 7 all bond, angle and dihedral interactions are omitted = 8 all bond, angle, dihedral and non-bonded interactions are omitted

ntb=1,

This variable controls whether or not periodic boundaries are imposed on the system during the calculation of non-bonded interactions. Bonds spanning periodic boundaries are not yet supported. There is no longer any need to set this variable, since it can be determined from igb and ntp parameters. The “proper” default for ntb is chosen (ntb=0 when igb > 0, ntb=2 when ntp > 0, and ntb=1 otherwise). This behavior can be overridden by supplying an explicit value, although this is discouraged to prevent errors. The allowed values for NTB are = 0 no periodicity is applied and PME is off (default when igb > 0) = 1 constant volume (default when igb and ntp are both 0, which are their defaults) = 2 constant pressure (default when ntp > 0) If NTB is nonzero then there must be a periodic boundary in the topology file. Constant pressure is not used in minimization (IMIN=1, above).

igb=0,

igb Flag for using the generalized Born or Poisson-Boltzmann implicit solvent models. See Section 4 for information about using this option. Default is 0.

ntp=0,

Flag for constant pressure dynamics. This option should be set to 1 or 2 when Constant Pressure periodic boundary conditions are used. = 0 No pressure scaling (Default) = 1 md with isotropic position scaling = 2 md with anisotropic (x-,y-,z-) pressure scaling: this should only be used with orthogonal boxes (i.e. with all angles set to 90 degrees). Anisotropic scaling is primarily intended for non-isotropic systems, such as membrane simulations, where the surface tensions are different in different directions; it is generally not appropriate for solutes dissolved in water.) = 3 md with semiisotropic pressure scaling: this is only available with constant surface tension (csurften > 0) and orthogonal boxes. This links the pressure coupling in the two directions tangential to the interface.

ntt=1,

Switch for temperature scaling. Note that setting ntt=0 corresponds to the microcanonical (NVE) ensemble (which should approach the canonical one for large numbers of degrees of freedom). Some aspects of the "weak-coupling ensemble" (ntt=1) have been examined, and roughly interpolate between the microcanonical and canonical ensembles.[333, 334] The ntt=2 and 3 options correspond to the canonical (constant T) ensemble.

= 0 Constant total energy classical dynamics (assuming that ntb<2, as should probably always be the case when ntt=0).

= 1 Constant temperature, using the weak-coupling algorithm. A single scaling factor is used for all atoms. Note that this algorithm just ensures that the total kinetic energy is appropriate for the desired temperature; it does nothing to ensure that the temperature is even over all parts of the molecule. Atomic collisions will tend to ensure an even temperature distribution, but this is not guaranteed, and there are many subtle problems that can arise with weak temperature coupling.[336] Using ntt=1 is especially dangerous for generalized Born simulations, where there are no collisions with solvent to aid in thermalization.) Other temperature coupling options (especially ntt=3) should be used instead.

= 2 Andersen temperature coupling scheme, in which imaginary "collisions" randomize the velocities to a distribution corresponding to temp0 every vrand steps. Note that in between these "massive collisions", the dynamics is Newtonian. Hence, time correlation functions (etc.) can be computed in these sections, and the results averaged over an initial canonical distribution. Note also that too high a collision rate (too small a value of vrand) will slow down the speed at which the molecules explore configuration space, whereas too low a rate means that the canonical distribution of energies will be sampled slowly. A discussion of this rate is given by Andersen

= 3 Use Langevin dynamics with the collision frequency γ given by gamma_ln, discussed below. Note that when γ has its default value of zero, this is the same as setting ntt = 0. Since Langevin simulations are highly susceptible to "synchronization" artifacts, you should explicitly set the ig variable (described below) to a different value at each restart of a given simulation

tautp=100,

Time constant, in ps, for heat bath coupling for the system, if ntt = 1. Default is 1.0. Generally, values for TAUTP should be in the range of 0.5-5.0 ps, with a smaller value providing tighter coupling to the heat bath and, thus, faster heating and a less natural trajectory. Smaller values of TAUTP result in smaller fluctuations in kinetic energy, but larger fluctuations in the total energy. Values much larger than the length of the simulation result in a return to constant energy conditions.

ig=-1,

The seed for the pseudo-random number generator. The MD starting velocity is dependent on the random number generator seed if NTX .lt. 3 .and. TEMPI .ne. 0.0. The value of this seed also affects the set of pseudo-random values used for Langevin dynamics or Andersen coupling, and hence should be set to a different value on each restart if ntt = 2 or 3. If ig=-1, (the default) the random seed will be based on the current date and time, and hence will be different for every run. It is recommended that, unless you specifically desire reproducibility, that you set ig=-1 for all runs involving ntt=2 or 3.

Cut = 8

This is used to specify the nonbonded cutoff, in Angstroms. For PME, the cutoff is used to limit direct space sum, and 8.0 is usually a good value. When igb>0, the cutoff is used to truncate nonbonded pairs (on an atom-by-atom basis); here a larger value than the default is generally required. A separate parameter (RGBMAX) controls the maximum distance between atom pairs that will be considered in carrying out the pairwise summation involved in calculating the effective Born radii, see the generalized Born section below

rgbmax=8,

rgbmax This parameter controls the maximum distance between atom pairs that will be considered in carrying out the pairwise summation involved in calculating the effective Born radii. Atoms whose associated spheres are farther way than rgbmax from given atom will not contribute to that atom’s effective Born radius. This is implemented in a "smooth" fashion (thanks mainly to W.A. Svrcek-Seiler), so that when part of an atom’s atomic sphere lies inside rgbmax cutoff, that part contributes to the low-dielectric 59 4. The Generalized Born/Surface Area Model region that determines the effective Born radius. The default is 25 Å, which is usually plenty for single-domain proteins of a few hundred residues. Even smaller values (of 10-15 Å) are reasonable, changing the functional form of the generalized Born theory a little bit, in exchange for a considerable speed-up in efficiency, and without introducing the usual cut-off artifacts such as drifts in the total energy. The rgbmax parameter affects only the effective Born radii (and the derivatives of these values with respect to atomic coordinates). The cut parameter, on the other hand, determines the maximum distance for the electrostatic, van der Waals and "off-diagonal" terms of the generalized Born interaction. The value of rgbmax might be either greater or smaller than that of cut: these two parameters are independent of each other. However, values of cut that are too small are more likely to lead to artifacts than are small values of rgbmax; therefore one typically sets rgbmax <= cut

Model_simulation1.sh

```
# Execute your Bash code below:
# note if restarting ntx=5 and irest=1

function create_files {

cat > Min1.in <<EOF
Initial minimisation
&cntrl
Imin=1, maxcyc=50000, ncyc=50000, irest=0,ntb=1,
Cut=8, rgbmax=10, ntpr=100, ntwx=0, ntx=1, ntr=1,
/
Holdfixedprotein
1000
RES 1 239
END
END
EOF

cat > Min2.in <<EOF
Initial minimisation
&cntrl
Imin=1, maxcyc=50000, ncyc=50000, irest=0,ntb=1,
Cut=8, rgbmax=10, ntpr=100, ntwx=0, ntx=1, ntr=1,
/
Holdfixedprotein
500
RES 1 239
END
END
EOF

cat > Min3.in <<EOF
Initial minimisation
&cntrl
```

```
Imin=1, maxcyc=50000, ncyc=50000, irest=0,ntb=1,
Cut=8, rgbmax=10, ntpr=100, ntwx=0, ntx=1, ntr=1,
/
Holdfixedprotein
200
RES 1 239
END
END
EOF

cat > Min4.in <<EOF
Initial minimisation
&cntrl
Imin=1, maxcyc=75000, ncyc=50000, irest=0,ntb=1,
Cut=8, rgbmax=10, ntpr=100, ntwx=0, ntx=1,
/
EOF


cat > Heat1.in <<EOF
Heat system
&cntrl
 imin=0, ntx=1, irest=0,
 nstlim=40000, dt=0.0005, ntc=1,
 ntpr=100, ntwx=100,
 cut=8, ntb=1, igb=0, ntp=0,
 ntt=3, gamma_ln=1.0, ig=-1,
 tempi=000.0, temp0=100.0,
cut=8, rgbmax=8, nmropt=1,
/
&wt
   type='TEMP0',
   istep1=0,
   istep2=40000,
   value1=000.0,
   value2=100.0,
```

```
/
&wt type='END',
/
EOF

cat > Heat2.in <<EOF
Heat system
&cntrl
 imin=0, ntx=1, irest=0,
 nstlim=40000, dt=0.0005, ntc=1,
 ntpr=100, ntwx=100,
 cut=8, ntb=1, igb=0, ntp=0,
 ntt=3, gamma_ln=1.0, ig=-1,
 tempi=100.0, temp0=250.0,
cut=8, rgbmax=8, nmropt=1,
/
&wt
   type='TEMP0',
   istep1=0,
   istep2=40000,
   value1=100.0,
   value2=250.0,
/
&wt type='END',
/
EOF

cat > Heat3.in <<EOF
Heat system
&cntrl
 imin=0, ntx=5, irest=1,
 nstlim=30000, dt=0.0005, ntc=1,
 ntpr=100, ntwx=100,
 cut=8, ntb=2, igb=0, ntp=1,
 ntt=3, gamma_ln=1.0, ig=-1,
 tempi=250.0, temp0=280.0,
```

```
cut=8, rgbmax=8, nmropt=1,
/
&wt
   type='TEMP0',
   istep1=0,
   istep2=30000,
   value1=250.0,
   value2=280.0,
/
&wt type='END',
/
EOF

cat > Heat4.in <<EOF
Heat system
&cntrl
 imin=0, ntx=5, irest=1,
 nstlim=30000, dt=0.0005, ntc=1,
 ntpr=100, ntwx=100,
 cut=8, ntb=2, igb=0, ntp=1,
 ntt=3, gamma_ln=1.0, ig=-1,
 tempi=280.0, temp0=310.0,
cut=8, rgbmax=8, nmropt=1,
/
&wt
   type='TEMP0',
   istep1=0,
   istep2=30000,
   value1=280.0,
   value2=310.0,
/
&wt type='END',
/
EOF

cat > Equil.in <<EOF
```

```
Heat system
&cntrl
  imin=0, ntx=5, irest=1,
  nstlim=1500000, dt=0.002, ntc=2, ntf=2,
  ntpr=100, ntwx=100,
  cut=8, ntb=2, igb=0, ntp=1,
  ntt=3, gamma_ln=1.0, ig=-1,
  tempi=310.0, temp0=310.0,
cut=8, rgbmax=8,
/
EOF

cat > Prod.in <<EOF
Heat system
&cntrl
  imin=0, ntx=5, irest=1,
  nstlim=10000000, dt=0.002, ntc=2, ntf=2,
  ntpr=1000, ntwx=1000,
  cut=8, ntb=1, igb=0, ntp=0,
  ntt=1, tautp=100, ig=-1,
  tempi=310.0, temp0=310.0,
cut=8, rgbmax=8,
/
EOF
}

#run step 1
function run_system {
pmemd.cuda -O -i Min1.in -o Min1.out -p PNPLA3.prmtop -c PNPLA3.inpcrd -r PNPLA3Min1.rst -inf
        Min1.mdinfo -ref PNPLA3.inpcrd
ambpdb -p PNPLA3.prmtop -c PNPLA3Min1.rst > PNPLA3Min1.pdb
pmemd.cuda -O -i Min2.in -o Min2.out -p PNPLA3.prmtop -c PNPLA3Min1.rst -r PNPLA3Min2.rst -inf
        Min2.mdinfo -ref PNPLA3Min1.rst
ambpdb -p PNPLA3.prmtop -c PNPLA3Min2.rst > PNPLA3Min2.pdb
pmemd.cuda -O -i Min3.in -o Min3.out -p PNPLA3.prmtop -c PNPLA3Min2.rst -r PNPLA3Min3.rst -inf
        Min3.mdinfo -ref PNPLA3Min2.rst
ambpdb -p PNPLA3.prmtop -c PNPLA3Min3.rst > PNPLA3Min3.pdb
```

```
pmemd.cuda -O -i Min4.in -o Min4.out -p PNPLA3.prmtop -c PNPLA3Min3.rst -r PNPLA3Min4.rst -inf
        Min4.mdinfo

ambpdb -p PNPLA3.prmtop -c PNPLA3Min4.rst > PNPLA3Min4.pdb

pmemd.cuda -O -i Heat1.in -o Heat1.out -p PNPLA3.prmtop -c PNPLA3Min4.rst -r Heat1.rst -inf
        Heat1.mdinfo -x Heat1.mdcrd

ambpdb -p PNPLA3.prmtop -c Heat1.rst > Heat1.pdb

pmemd.cuda -O -i Heat2.in -o Heat2.out -p PNPLA3.prmtop -c Heat1.rst -r Heat2.rst -inf Heat2.mdinfo -x
        Heat2.mdcrd

ambpdb -p PNPLA3.prmtop -c Heat2.rst > Heat2.pdb

pmemd.cuda -O -i Heat3.in -o Heat3.out -p PNPLA3.prmtop -c Heat2.rst -r Heat3.rst -inf Heat3.mdinfo -x
        Heat3.mdcrd

ambpdb -p PNPLA3.prmtop -c Heat3.rst > Heat3.pdb

pmemd.cuda -O -i Heat4.in -o Heat4.out -p PNPLA3.prmtop -c Heat3.rst -r Heat4.rst -inf Heat4.mdinfo -x
        Heat4.mdcrd

ambpdb -p PNPLA3.prmtop -c Heat4.rst > Heat4.pdb

pmemd.cuda -O -i Equil.in -o Equil.out -p PNPLA3.prmtop -c Heat4.rst -r Equil.rst -inf Equil.mdinfo -x
        Equil.mdcrd

ambpdb -p PNPLA3.prmtop -c Equil.rst > Equil.pdb

pmemd.cuda -O -i Prod.in -o Prod.out -p PNPLA3.prmtop -c Equil.rst -r Prod.rst -inf Prod.mdinfo -x
        Prod.mdcrd

ambpdb -p PNPLA3.prmtop -c Prod.rst > Prod.pdb

}


#run all

create_files

run_system
```

Extract_all_results.sh

```
#! bin/bash

# script to extract all data from the simulation, from minimisation, heating and equilibration,
        to production.


#load amber modules

export AMBERHOME=/home/will/AMBER/amber16

source /home/will/AMBER//amber16/amber.sh


# extract basic data from the run outputs.

bash Extract_min_results_and_plot_script.sh

perl extract_heating_results.pl Heat1.out Heat2.out Heat3.out Heat4.out Equil.out

perl extract_production_results.pl Prod.out


# run cpptraj data extraction

cpptraj -i Extract_heat.cpptraj

cpptraj -i Extract_heat2.cpptraj

cpptraj -i Extract_prod.cpptraj

cpptraj -i Extract_prod2.cpptraj
```

Extract_min_results_and_plot.sh (Adapted from ambertools 17).[446]

```
#! bin/bash
# script to extract key data from the minimization of an amber simulation and plt important graphs
# extract data into individual files of column data

# min 1
grep -A1 " NSTEP " Min1.out | grep -v " NSTEP " | grep " " | awk '{print$1,$1}' > rstep1.dat
grep -A1 " ENERGY " Min1.out | grep -v " ENERGY " | grep " " | awk '{print$1,$2}' > energy1.dat
grep -A1 " RMS " Min1.out | grep -v " RMS " | grep " " | awk '{print$1,$3}' > rms1.dat
grep -A1 " GMAX " Min1.out | grep -v " GMAX " | grep " " | awk '{print$1,$4}' > gmax1.dat

# min 2
grep -A1 " NSTEP " Min2.out | grep -v " NSTEP " | grep " " | awk '{print$1,$1}' > rstep2.dat
grep -A1 " ENERGY " Min2.out | grep -v " ENERGY " | grep " " | awk '{print$1,$2}' > energy2.dat
grep -A1 " RMS " Min2.out | grep -v " RMS " | grep " " | awk '{print$1,$3}' > rms2.dat
grep -A1 " GMAX " Min2.out | grep -v " GMAX " | grep " " | awk '{print$1,$4}' > gmax2.dat

# min 3
grep -A1 " NSTEP " Min3.out | grep -v " NSTEP " | grep " " | awk '{print$1,$1}' > rstep3.dat
grep -A1 " ENERGY " Min3.out | grep -v " ENERGY " | grep " " | awk '{print$1,$2}' > energy3.dat
grep -A1 " RMS " Min3.out | grep -v " RMS " | grep " " | awk '{print$1,$3}' > rms3.dat
grep -A1 " GMAX " Min3.out | grep -v " GMAX " | grep " " | awk '{print$1,$4}' > gmax3.dat

# min 4
grep -A1 " NSTEP " Min4.out | grep -v " NSTEP " | grep " " | awk '{print$1,$1}' > rstep4.dat
grep -A1 " ENERGY " Min4.out | grep -v " ENERGY " | grep " " | awk '{print$1,$2}' > energy4.dat
grep -A1 " RMS " Min4.out | grep -v " RMS " | grep " " | awk '{print$1,$3}' > rms4.dat
grep -A1 " GMAX " Min4.out | grep -v " GMAX " | grep " " | awk '{print$1,$4}' > gmax4.dat

# concatonate the files into single files

cat rstep1.dat rstep2.dat rstep3.dat rstep4.dat > rstep.dat
cat energy1.dat energy2.dat energy3.dat energy4.dat > energy.dat
cat rms1.dat rms2.dat rms3.dat rms4.dat > rms.dat
cat gmax1.dat gmax2.dat gmax3.dat gmax4.dat > gmax.dat
```

```
# change the increments to show step progression (assumes input is 100 increments, at 0.5 picosecond
        cycles. multiplied by 2 to change to 100 picosecond intervals or 0.1Ns)


awk 'BEGIN{OFS=" "; col1=0}{print col1, $2; col1+=100}' rstep.dat > Mrstep.dat

awk 'BEGIN{OFS=" "; col1=0}{print col1, $2; col1+=100}' energy.dat > Menergy.dat

awk 'BEGIN{OFS=" "; col1=0}{print col1, $2; col1+=100}' rms.dat > Mrms.dat

awk 'BEGIN{OFS=" "; col1=0}{print col1, $2; col1+=100}' gmax.dat > Mgmax.dat
```

extract_heating_results.pl (Adapted from ambertools 17).[446]

```
#!/usr/bin/perl

if ($#ARGV < 0) {
   print "  Incorrect usage...\n";
   exit;
}
foreach $i ( 0..$#ARGV ) {
   $filein = $ARGV[$i];
   $checkfile = $filein;
   $checkfile =~ s/\.Z//;
   if ( $filein ne $checkfile ) {
         open(INPUT, "zcat $filein |") ||
           die "Cannot open compressed $filein -- $!\n";
   } else {
         open(INPUT, $filein) || die "Cannot open $filein -- $!\n";
   }
   print "Processing sander output file ($filein)...\n";
   &process_input;
   close(INPUT);
}

print "Starting output...\n";
@sortedkeys = sort by_number keys(%TIME);
@sortedavgkeys = sort by_number keys(%AVG_TIME);

foreach $i ( TEMP, TSOLUTE, TSOLVENT, PRES, EKCMT, ETOT, EKTOT, EPTOT, DENSITY, VOLUME ) {
   print "Outputing Hsummary.$i\n";
   open(OUTPUT, "> Hsummary.$i");
   %outarray = eval "\%$i";
   foreach $j ( @sortedkeys ) {
         print OUTPUT "$j  ", $outarray{$j}, "\n";
   }
   close (OUTPUT);

   print "Outputing Hsummary_avg.$i\n";
```

```
    open(OUTPUT, "> Hsummary_avg.$i");
    %outarray = eval "\%AVG_$i";
    foreach $j ( @sortedavgkeys ) {
        print OUTPUT "$j  ", $outarray{$j}, "\n";
    }
    close (OUTPUT);

    print "Outputing Hsummary_rms.$i\n";
    open(OUTPUT, "> Hsummary_rms.$i");
    %outarray = eval "\%RMS_$i";
    foreach $j ( @sortedavgkeys ) {
        print OUTPUT "$j  ", $outarray{$j}, "\n";
    }
    close (OUTPUT);
}
##########################################
sub by_number {
    if ($a < $b) {
        -1;
    } elsif ($a == $b) {
        0;
    } elsif ($a > $b) {
        1;
    }
}

sub process_input {

    $status = 0;
    $debug = 0;
    while ( <INPUT> ) {
        $string = $_;

        print $_ if ( ! /NB-upda/ && $debug );

        if (/A V E R A G E S/) {
```

```
    $averages = 1;
    ($averages_over) = /.*O V E R.*(\d*).*S T E P S/;
}

$rms = 1 if (/R M S/);

if (/NSTEP/) {
    ($time, $temp, $pres) =
  /NSTEP =.*TIME.* =(.*\d*\.\d*).*TEMP.* =(.*\d*\.\d*).*PRESS = (.*\d*\.\d*)/;
    if ( $debug ) {
  print $_;
  print "time is $time, temp is $temp, pres is $pres\n";
    }
    $_ = <INPUT>;

    if (/Etot/) {
  ($etot, $ektot, $eptot) =
     /Etot.*=(.*\d*\.\d*).*EKtot.*=(.*\d*\.\d*).*EPtot.*=(.*\d*\.\d*)/;
  if ( $debug ) {
     print $_;
     print "Etot is $etot, ektot is $ektot, eptot is $eptot\n";
  }
  $_ = <INPUT>;
    }
    if (/BOND.*ANGLE.*DIHED/) {
  ($bond, $angle, $dihedral) =
     /BOND.*=(.*\d*\.\d*).*ANGLE.*=(.*\d*\.\d*).*DIHED.*=(.*\d*\.\d*)/;
  if ( $debug ) {
     print $_;
     print "bond is $bond, angle is $angle, dihedral is $dihedral\n";
  }
  $_ = <INPUT>;
    }
    if (/1-4 NB/) {
  ($nb14, $eel14, $nb) =
     /1-4 NB.*=(.*\d*\.\d*).*1-4 EEL.*=(.*\d*\.\d*).*VDWAALS.*=(.*\d*\.\d*)/;
```

```
        if ( $debug ) {
           print $_;
           print "nb14 is $nb14, eel14 is $eel14, vdwaals is $nb\n";
        }
        $_ = <INPUT>;
         }
         if (/EELEC/) {
        ($eel, $ehbond, $constraint) =
           /EELEC.*=(.*\d*\.\d*).*EHBOND.*=(.*\d*\.\d*).*CONSTRAINT.*=(.*\d*\.\d*)/;
        if ( $debug ) {
           print $_;
           print "eel is $eel, ehbond is $ehbond, constraint is $constraint\n";
        }
        $_ = <INPUT>;
#
#        check to see if EAMBER is in the mdout file (present when
#        NTR=1)
#
            if ( /EAMBER/ ) {
           $_ = <INPUT>;
        }
         }
         if (/EKCMT/) {
        ($ekcmt, $virial, $volume) =
           /EKCMT.*=(.*\d*\.\d*).*VIRIAL.*=(.*\d*\.\d*).*VOLUME.*=(.*\d*\.\d*)/;
        if ( $debug ) {
           print $_;
           print "Ekcmt is $ekcmt, virial is $virial, volume is $volume\n";
        }
        $_ = <INPUT>;
         }
         if (/T_SOLUTE/) {
        ($tsolute, $tsolvent) =
           /T_SOLUTE =(.*\d*\.\d*).*T_SOLVENT =(.*\d*\.\d*)/;
        if ( $debug ) {
           print $_;
```

```
    print "Temp solute is $tsolute, temp solvent is $tsolvent\n";
  }
  $_ = <INPUT>;
   }

   if (/Density/) {
  ($density) = /.*Density.*=(.*\d*\.\d*)/;
  if ( $debug ) {
    print $_;
    print "Density is $density\n";
  }
  $_ = <INPUT>;
   }

#     update arrays

    if ( $averages == 1 ) {
  $AVG_TIME{$time}      = $time;
  $AVG_TEMP{$time}      = $temp;
  $AVG_PRES{$time}      = $pres;
  $AVG_ETOT{$time}      = $etot;
  $AVG_EKTOT{$time}     = $ektot;
  $AVG_EPTOT{$time}     = $eptot;
  $AVG_BOND{$time}      = $bond;
  $AVG_ANGLE{$time}     = $angle;
  $AVG_DIHEDRAL{$time}   = $dihedral;
  $AVG_NB14{$time}      = $nb14;
  $AVG_EEL14{$time}     = $eel14;
  $AVG_NB{$time}        = $nb;
  $AVG_EEL{$time}       = $eel;
  $AVG_EHBOND{$time}    = $ehbond;
  $AVG_CONSTRAINT{$time} = $constraint;
  $AVG_EKCMT{$time}     = $ekcmt;
  $AVG_VIRIAL{$time}    = $virial;
  $AVG_VOLUME{$time}    = $volume;
  $AVG_TSOLUTE{$time}   = $tsolute;
```

```
$AVG_TSOLVENT{$time}   = $tsolvent;
$AVG_DENSITY{$time}    = $density;

$averages = 0;
  } elsif ( $rms == 1 ) {
$RMS_TIME{$time}      = $time;
$RMS_TEMP{$time}      = $temp;
$RMS_PRES{$time}      = $pres;
$RMS_ETOT{$time}      = $etot;
$RMS_EKTOT{$time}     = $ektot;
$RMS_EPTOT{$time}     = $eptot;
$RMS_BOND{$time}      = $bond;
$RMS_ANGLE{$time}     = $angle;
$RMS_DIHEDRAL{$time}   = $dihedral;
$RMS_NB14{$time}      = $nb14;
$RMS_EEL14{$time}     = $eel14;
$RMS_NB{$time}        = $nb;
$RMS_EEL{$time}       = $eel;
$RMS_EHBOND{$time}    = $ehbond;
$RMS_CONSTRAINT{$time} = $constraint;
$RMS_EKCMT{$time}     = $ekcmt;
$RMS_VIRIAL{$time}    = $virial;
$RMS_VOLUME{$time}    = $volume;
$RMS_TSOLUTE{$time}   = $tsolute;
$RMS_TSOLVENT{$time}   = $tsolvent;
$RMS_DENSITY{$time}    = $density;

$rms = 0;
  } else {
$TIME{$time}      = $time;
$TEMP{$time}      = $temp;
$PRES{$time}      = $pres;
$ETOT{$time}      = $etot;
$EKTOT{$time}     = $ektot;
$EPTOT{$time}     = $eptot;
$BOND{$time}      = $bond;
```

```
            $ANGLE{$time}     = $angle;
            $DIHEDRAL{$time}   = $dihedral;
            $NB14{$time}      = $nb14;
            $EEL14{$time}     = $eel14;
            $NB{$time}        = $nb;
            $EEL{$time}       = $eel;
            $EHBOND{$time}    = $ehbond;
            $CONSTRAINT{$time} = $constraint;
            $EKCMT{$time}     = $ekcmt;
            $VIRIAL{$time}    = $virial;
            $VOLUME{$time}    = $volume;
            $TSOLUTE{$time}   = $tsolute;
            $TSOLVENT{$time}   = $tsolvent;
            $DENSITY{$time}   = $density;
             }

          }
   }
}
```

extract_production_results.pl (Adapted from ambertools 17).[446]

```
#!/usr/bin/perl

if ($#ARGV < 0) {
   print "  Incorrect usage...\n";
   exit;
}

foreach $i ( 0..$#ARGV ) {
   $filein = $ARGV[$i];
   $checkfile = $filein;
   $checkfile =~ s/\.Z//;
   if ( $filein ne $checkfile ) {
         open(INPUT, "zcat $filein |") ||
            die "Cannot open compressed $filein -- $!\n";
   } else {
         open(INPUT, $filein) || die "Cannot open $filein -- $!\n";
   }
   print "Processing sander output file ($filein)...\n";
   &process_input;
   close(INPUT);
}

print "Starting output...\n";
@sortedkeys = sort by_number keys(%TIME);
@sortedavgkeys = sort by_number keys(%AVG_TIME);

foreach $i ( TEMP, TSOLUTE, TSOLVENT, PRES, EKCMT, ETOT, EKTOT, EPTOT, DENSITY, VOLUME ) {
   print "Outputing Psummary.$i\n";
   open(OUTPUT, "> Psummary.$i");
   %outarray = eval "\%$i";
   foreach $j ( @sortedkeys ) {
         print OUTPUT "$j  ", $outarray{$j}, "\n";
   }
   close (OUTPUT);
```

```
    print "Outputing Psummary_avg.$i\n";
    open(OUTPUT, "> Psummary_avg.$i");
    %outarray = eval "\%AVG_$i";
    foreach $j ( @sortedavgkeys ) {
        print OUTPUT "$j  ", $outarray{$j}, "\n";
    }
    close (OUTPUT);

    print "Outputing Psummary_rms.$i\n";
    open(OUTPUT, "> Psummary_rms.$i");
    %outarray = eval "\%RMS_$i";
    foreach $j ( @sortedavgkeys ) {
        print OUTPUT "$j  ", $outarray{$j}, "\n";
    }
    close (OUTPUT);


}


sub by_number {
    if ($a < $b) {
        -1;
    } elsif ($a == $b) {
        0;
    } elsif ($a > $b) {
        1;
    }
}

sub process_input {

    $status = 0;
    $debug = 0;
    while ( <INPUT> ) {
        $string = $_;
```

```
print $_ if ( ! /NB-upda/ && $debug );

if (/A V E R A G E S/) {
   $averages = 1;
   ($averages_over) = /.*O V E R.*(\d*).*S T E P S/;
}

$rms = 1 if (/R M S/);

if (/NSTEP/) {
   ($time, $temp, $pres) =
 /NSTEP =.*TIME.* =(.*\d*\.\d*).*TEMP.* =(.*\d*\.\d*).*PRESS = (.*\d*\.\d*)/;
   if ( $debug ) {
 print $_;
 print "time is $time, temp is $temp, pres is $pres\n";
   }
   $_ = <INPUT>;

   if (/Etot/) {
 ($etot, $ektot, $eptot) =
    /Etot.*=(.*\d*\.\d*).*EKtot.*=(.*\d*\.\d*).*EPtot.*=(.*\d*\.\d*)/;
 if ( $debug ) {
    print $_;
    print "Etot is $etot, ektot is $ektot, eptot is $eptot\n";
 }
 $_ = <INPUT>;
  }
  if (/BOND.*ANGLE.*DIHED/) {
 ($bond, $angle, $dihedral) =
    /BOND.*=(.*\d*\.\d*).*ANGLE.*=(.*\d*\.\d*).*DIHED.*=(.*\d*\.\d*)/;
 if ( $debug ) {
    print $_;
    print "bond is $bond, angle is $angle, dihedral is $dihedral\n";
 }
 $_ = <INPUT>;
```

```
  }
  if (/1-4 NB/) {
($nb14, $eel14, $nb) =
   /1-4 NB.*=(.*\d*\.\d*).*1-4 EEL.*=(.*\d*\.\d*).*VDWAALS.*=(.*\d*\.\d*)/;
if ( $debug ) {
   print $_;
   print "nb14 is $nb14, eel14 is $eel14, vdwaals is $nb\n";
}
$_ = <INPUT>;
  }
  if (/EELEC/) {
($eel, $ehbond, $constraint) =
   /EELEC.*=(.*\d*\.\d*).*EHBOND.*=(.*\d*\.\d*).*CONSTRAINT.*=(.*\d*\.\d*)/;
if ( $debug ) {
   print $_;
   print "eel is $eel, ehbond is $ehbond, constraint is $constraint\n";
}
$_ = <INPUT>;
#
#          check to see if EAMBER is in the mdout file (present when
#          NTR=1)
#
     if ( /EAMBER/ ) {
   $_ = <INPUT>;
}
  }
  if (/EKCMT/) {
($ekcmt, $virial, $volume) =
   /EKCMT.*=(.*\d*\.\d*).*VIRIAL.*=(.*\d*\.\d*).*VOLUME.*=(.*\d*\.\d*)/;
if ( $debug ) {
   print $_;
   print "Ekcmt is $ekcmt, virial is $virial, volume is $volume\n";
}
$_ = <INPUT>;
  }
  if (/T_SOLUTE/) {
```

```
($tsolute, $tsolvent) =
   /T_SOLUTE =(.*\d*\.\d*).*T_SOLVENT =(.*\d*\.\d*)/;
if ( $debug ) {
   print $_;
   print "Temp solute is $tsolute, temp solvent is $tsolvent\n";
}
$_ = <INPUT>;
 }

 if (/Density/) {
($density) = /.*Density.*=(.*\d*\.\d*)/;
if ( $debug ) {
   print $_;
   print "Density is $density\n";
}
$_ = <INPUT>;
 }

#     update arrays

  if ( $averages == 1 ) {
$AVG_TIME{$time}      = $time;
$AVG_TEMP{$time}      = $temp;
$AVG_PRES{$time}      = $pres;
$AVG_ETOT{$time}      = $etot;
$AVG_EKTOT{$time}     = $ektot;
$AVG_EPTOT{$time}     = $eptot;
$AVG_BOND{$time}      = $bond;
$AVG_ANGLE{$time}     = $angle;
$AVG_DIHEDRAL{$time}  = $dihedral;
$AVG_NB14{$time}      = $nb14;
$AVG_EEL14{$time}     = $eel14;
$AVG_NB{$time}        = $nb;
$AVG_EEL{$time}       = $eel;
$AVG_EHBOND{$time}    = $ehbond;
$AVG_CONSTRAINT{$time} = $constraint;
```

```
$AVG_EKCMT{$time}     = $ekcmt;
$AVG_VIRIAL{$time}    = $virial;
$AVG_VOLUME{$time}    = $volume;
$AVG_TSOLUTE{$time}   = $tsolute;
$AVG_TSOLVENT{$time}  = $tsolvent;
$AVG_DENSITY{$time}   = $density;

$averages = 0;
  } elsif ( $rms == 1 ) {
$RMS_TIME{$time}      = $time;
$RMS_TEMP{$time}      = $temp;
$RMS_PRES{$time}      = $pres;
$RMS_ETOT{$time}      = $etot;
$RMS_EKTOT{$time}     = $ektot;
$RMS_EPTOT{$time}     = $eptot;
$RMS_BOND{$time}      = $bond;
$RMS_ANGLE{$time}     = $angle;
$RMS_DIHEDRAL{$time}  = $dihedral;
$RMS_NB14{$time}      = $nb14;
$RMS_EEL14{$time}     = $eel14;
$RMS_NB{$time}        = $nb;
$RMS_EEL{$time}       = $eel;
$RMS_EHBOND{$time}    = $ehbond;
$RMS_CONSTRAINT{$time} = $constraint;
$RMS_EKCMT{$time}     = $ekcmt;
$RMS_VIRIAL{$time}    = $virial;
$RMS_VOLUME{$time}    = $volume;
$RMS_TSOLUTE{$time}   = $tsolute;
$RMS_TSOLVENT{$time}  = $tsolvent;
$RMS_DENSITY{$time}   = $density;

$rms = 0;
  } else {
$TIME{$time}      = $time;
$TEMP{$time}      = $temp;
$PRES{$time}      = $pres;
```

```
            $ETOT{$time}      = $etot;
            $EKTOT{$time}     = $ektot;
            $EPTOT{$time}     = $eptot;
            $BOND{$time}      = $bond;
            $ANGLE{$time}     = $angle;
            $DIHEDRAL{$time}   = $dihedral;
            $NB14{$time}      = $nb14;
            $EEL14{$time}     = $eel14;
            $NB{$time}        = $nb;
            $EEL{$time}       = $eel;
            $EHBOND{$time}    = $ehbond;
            $CONSTRAINT{$time} = $constraint;
            $EKCMT{$time}     = $ekcmt;
            $VIRIAL{$time}    = $virial;
            $VOLUME{$time}    = $volume;
            $TSOLUTE{$time}   = $tsolute;
            $TSOLVENT{$time}   = $tsolvent;
            $DENSITY{$time}   = $density;
              }

          }
    }
}
```

Extract_heat.cpptraj

```
parm PNPLA3.prmtop
trajin Heat1.mdcrd
trajin Heat2.mdcrd
trajin Heat3.mdcrd
trajin Heat4.mdcrd
trajin Equil.mdcrd

autoimage
trajout autoimaged_heat.nc netcdf

run
quit
```

Extract_heat2.cpptraj

```
parm PNPLA3.prmtop
trajin autoimaged_heat.nc

rms HeatRMSD :1:481&!@H= first out rmsdHeat.dat mass
atomicfluct HeatRMSF :1-481@CA out rmsfHeat.dat byres
rms HeatRMSDnearresidue :148<@6 first out rmsdHeatnearresidue.dat mass
distance Heat47148 :47 :148 out distanceHeat.dat

run
quit
```

Extract_prod.cpptraj

```
parm PNPLA3.prmtop
trajin Prod.mdcrd

autoimage
trajout autoimaged_prod.nc netcdf

run
quit
```

Extract_prod2.cpptraj

```
#remove_bugs
parm PNPLA3.prmtop
reference Prod.rst [R0]
trajin autoimaged_Prod.nc 1 1
unstrip
run
clear trajin
trajin autoimaged_Prod.nc
reference Prod.rst [R1]

#####   TESTS BELOW    #######
#distance serine asp
distance Proddist47to166 :47 :166 out 3Proddist47to166.dat

#distance serine 148
distance Proddist47to148 :47 :148 out 3Proddist47to148.dat

#distance asp 148
distance Proddist166to148 :166 :148 out 3Proddist166to148.dat



#distance potential loop closed position residues
distance Proddistance148to212 :148 :212 out 3Proddistance148to212.dat

### RMS based ####

#rmsd global not hydrogens.
rms ProdRMSDnH :1-481&!@H= first out 1rmsdProdnH.dat

#rmsd global backbone
rms ProdRMSDoNCCAN :1-481@C,CA,N first out 1rmsdProdback.dat

#rmsd global alpha carbons
```

```
rms ProdRMSD :1-481@CA first out 1rmsdProdalpha.dat mass

#rmsd Patatin domain 5-179
rms Prodpatatin5to179RMSD :5-179@C,CA,N first out 3Prodpatatin5to179RMSD.dat mass

#rmsd extended patatin domain where applicable 5-239
rms Prodpatatin5to239RMSD :5-179@C,CA,N first out 3Prodpatatin5to239RMSD.dat mass

#rmsd residue 148
rms Prodpatatin148RMSD :148@C,CA,N first out 3Prodpatatin148RMSD.dat mass

#rmsd residue 47
rms Prodpatatin47RMSD :47@C,CA,N first out 3Prodpatatin47RMSD.dat mass


#rmsd residue 166
rms Prodpatatin166RMSD :166@C,CA,N first out 3Prodpatatin166RMSD.dat mass

#rmsf global
atomicfluct ProdRMSF :1-481@C,CA,N out 1rmsfProd.dat byres

#rmsd active site region - within 6A of S47
rms ProdRMSD6from47 :47<@6 first out 3ProdRMSD6from47.dat mass

#rmsd active site region - within 12A of S47
rms ProdRMSD12from47 :47<@12 first out 3ProdRMSD12from47.dat mass


#rmsd 148 variant - within 6 ANG
rms ProdRMSD6from148 :148<@6 first out 3ProdRMSD6from148.dat mass

#rmsd 148 variant - within 12 ANG
rms ProdRMSD12from148 :148<@12 first out 3ProdRMSD12from148.dat mass
```

#rmsd flexible loop only - 147-151 residues

rms Prodpatatinflexloop147to151RMSD :147-151@C,CA,N first out 3Prodpatatinflexloop147to151RMSD.dat mass

run

quit

100ns_production.sh

```
function create_files {

cat > Prod1.in <<EOF
&cntrl
 imin=0, ntx=5, irest=1,
 nstlim=10000000, dt=0.002, ntc=2, ntf=2,
 ntpr=1000, ntwx=1000,
 cut=8, ntb=1, igb=0, ntp=0,
 ntt=1, tautp=100, ig=-1,
 tempi=310.0, temp0=310.0,
cut=8, rgbmax=8,
/
EOF

cat > Prod2.in <<EOF
&cntrl
 imin=0, ntx=5, irest=1,
 nstlim=10000000, dt=0.002, ntc=2, ntf=2,
 ntpr=1000, ntwx=1000,
 cut=8, ntb=1, igb=0, ntp=0,
 ntt=1, tautp=100, ig=-1,
 tempi=310.0, temp0=310.0,
cut=8, rgbmax=8,
/
EOF

cat > Prod3.in <<EOF
&cntrl
 imin=0, ntx=5, irest=1,
 nstlim=10000000, dt=0.002, ntc=2, ntf=2,
 ntpr=1000, ntwx=1000,
 cut=8, ntb=1, igb=0, ntp=0,
 ntt=1, tautp=100, ig=-1,
```

```
 tempi=310.0, temp0=310.0,
cut=8, rgbmax=8,
/
EOF

cat > Prod4.in <<EOF
&cntrl
 imin=0, ntx=5, irest=1,
 nstlim=10000000, dt=0.002, ntc=2, ntf=2,
 ntpr=1000, ntwx=1000,
 cut=8, ntb=1, igb=0, ntp=0,
 ntt=1, tautp=100, ig=-1,
 tempi=310.0, temp0=310.0,
cut=8, rgbmax=8,
/
EOF

cat > Prod5.in <<EOF
&cntrl
 imin=0, ntx=5, irest=1,
 nstlim=10000000, dt=0.002, ntc=2, ntf=2,
 ntpr=1000, ntwx=1000,
 cut=8, ntb=1, igb=0, ntp=0,
 ntt=1, tautp=100, ig=-1,
 tempi=310.0, temp0=310.0,
cut=8, rgbmax=8,
/
EOF
}

#run the heating step 1
function run_system {
```

```
pmemd.cuda -O -i Prod1.in -o Prod1.out -p PNPLA3.prmtop -c Prod.rst -r Prod1.rst -inf Prod1.mdinfo -x Prod1.mdcrd

ambpdb -p PNPLA3.prmtop -c Prod1.rst > Prod1.pdb

pmemd.cuda -O -i Prod2.in -o Prod2.out -p PNPLA3.prmtop -c Prod1.rst -r Prod2.rst -inf Prod2.mdinfo -x Prod2.mdcrd

ambpdb -p PNPLA3.prmtop -c Prod2.rst > Prod2.pdb

pmemd.cuda -O -i Prod3.in -o Prod3.out -p PNPLA3.prmtop -c Prod2.rst -r Prod3.rst -inf Prod3.mdinfo -x Prod3.mdcrd

ambpdb -p PNPLA3.prmtop -c Prod3.rst > Prod3.pdb

pmemd.cuda -O -i Prod4.in -o Prod4.out -p PNPLA3.prmtop -c Prod3.rst -r Prod4.rst -inf Prod4.mdinfo -x Prod4.mdcrd

ambpdb -p PNPLA3.prmtop -c Prod4.rst > Prod4.pdb

pmemd.cuda -O -i Prod5.in -o Prod5.out -p PNPLA3.prmtop -c Prod4.rst -r Prod5.rst -inf Prod5.mdinfo -x Prod5.mdcrd

ambpdb -p PNPLA3.prmtop -c Prod5.rst > Prod5.pdb

}


#run all

create_files

run_system
```

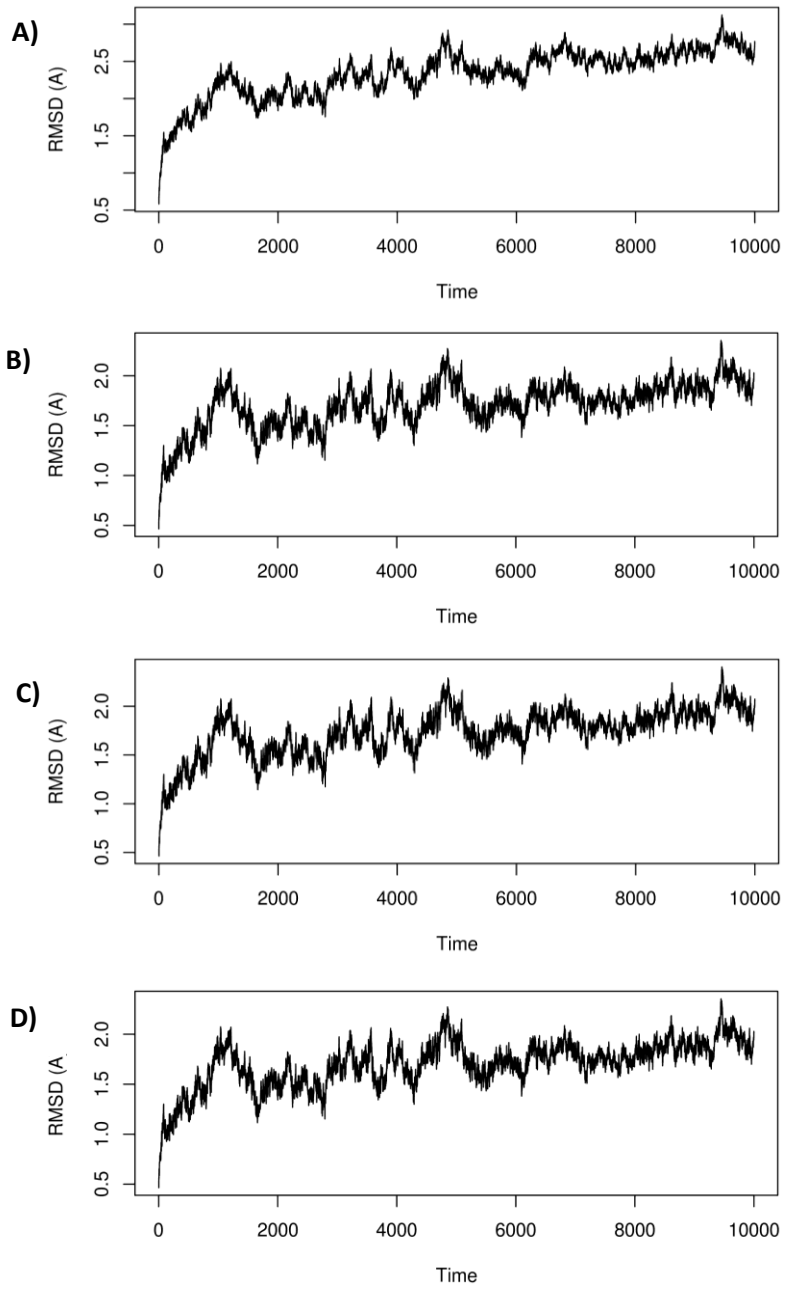


Figure A5.1 Root mean square deviation results for 20ns simulation of Model 1 without shake. Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in 0.5ps frames. **A)** RMSD of entire protein chain. **B)** RMSD based on backbone chain atoms C, CA and N. **C)** RMSD based upon α carbons. **D)** RMSD including only the patatin domain (residues 5-179).

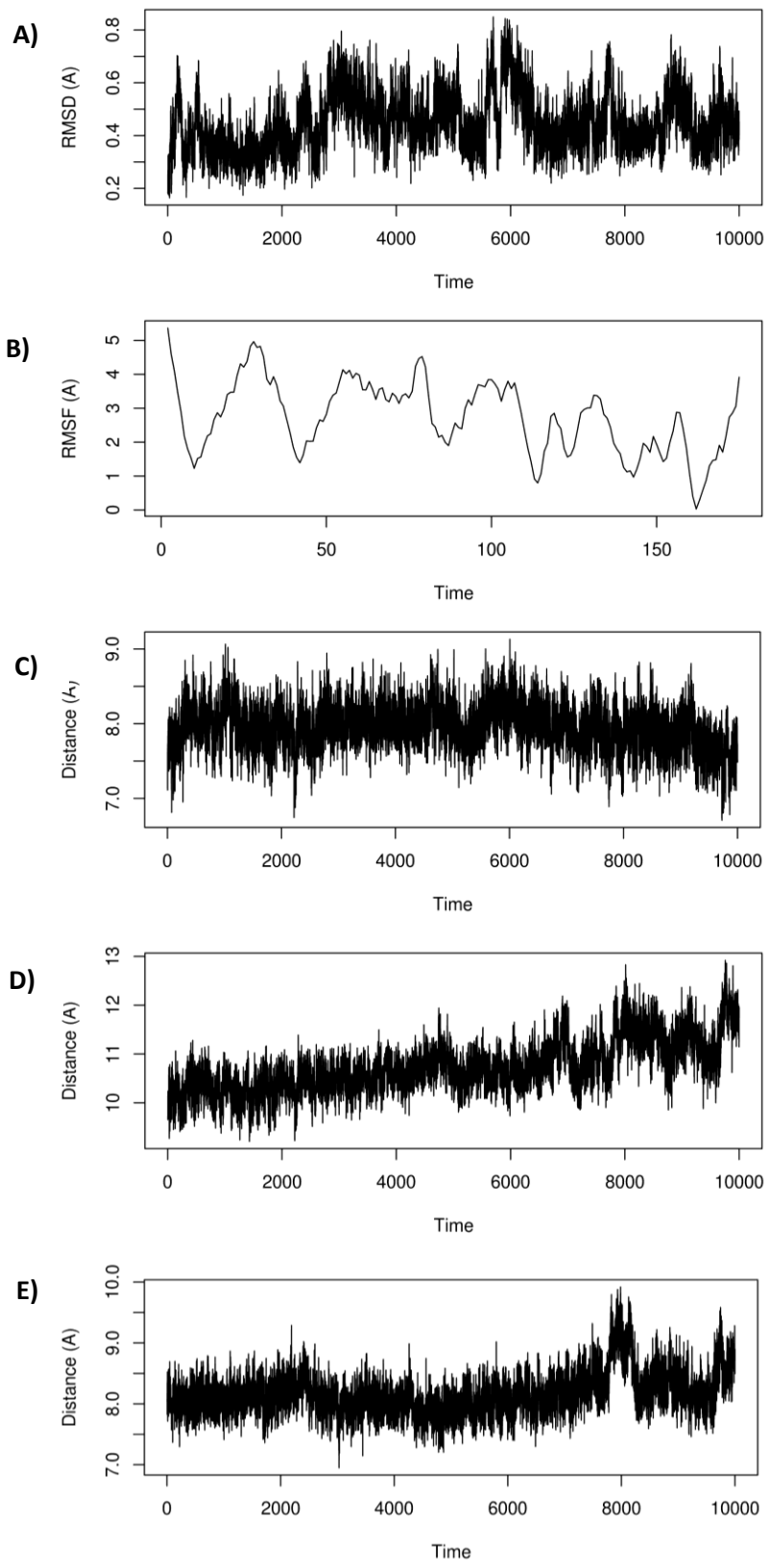


Figure A5.2 Results for 20ns simulation of Model 1 without shake (part 2). Time represented in 0.5ps frames. **A)** RMSD of I148M containing loop (residues 147-151). **B)** Root mean square fluctuations along the protein chain averaged for full simulation. **C)** Distance between S46 and D166. **D)** Distance between S46 and I148. **E)** Distance between D166 and I148.

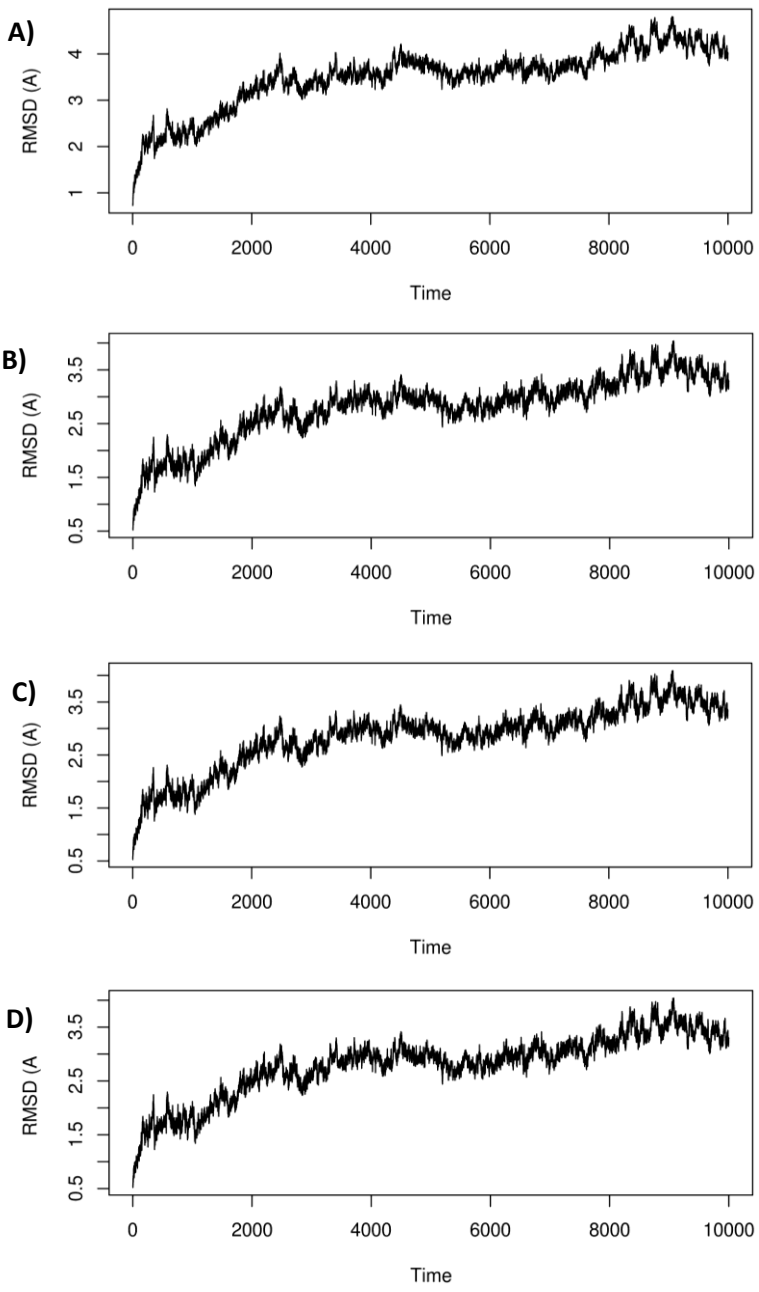


Figure A5.3 Root mean square deviation results for 20ns simulation of Model 1 under NPT. Graphs showing the root mean square deviation (RMSD) from the initial starting structure over the time course of the simulation. Time represented in 2ps frames. **A)** RMSD of entire protein chain. **B)** RMSD based on backbone chain atoms C, CA and N. **C)** RMSD based upon α carbons. **D)** RMSD including only the patatin domain (residues 5-179).

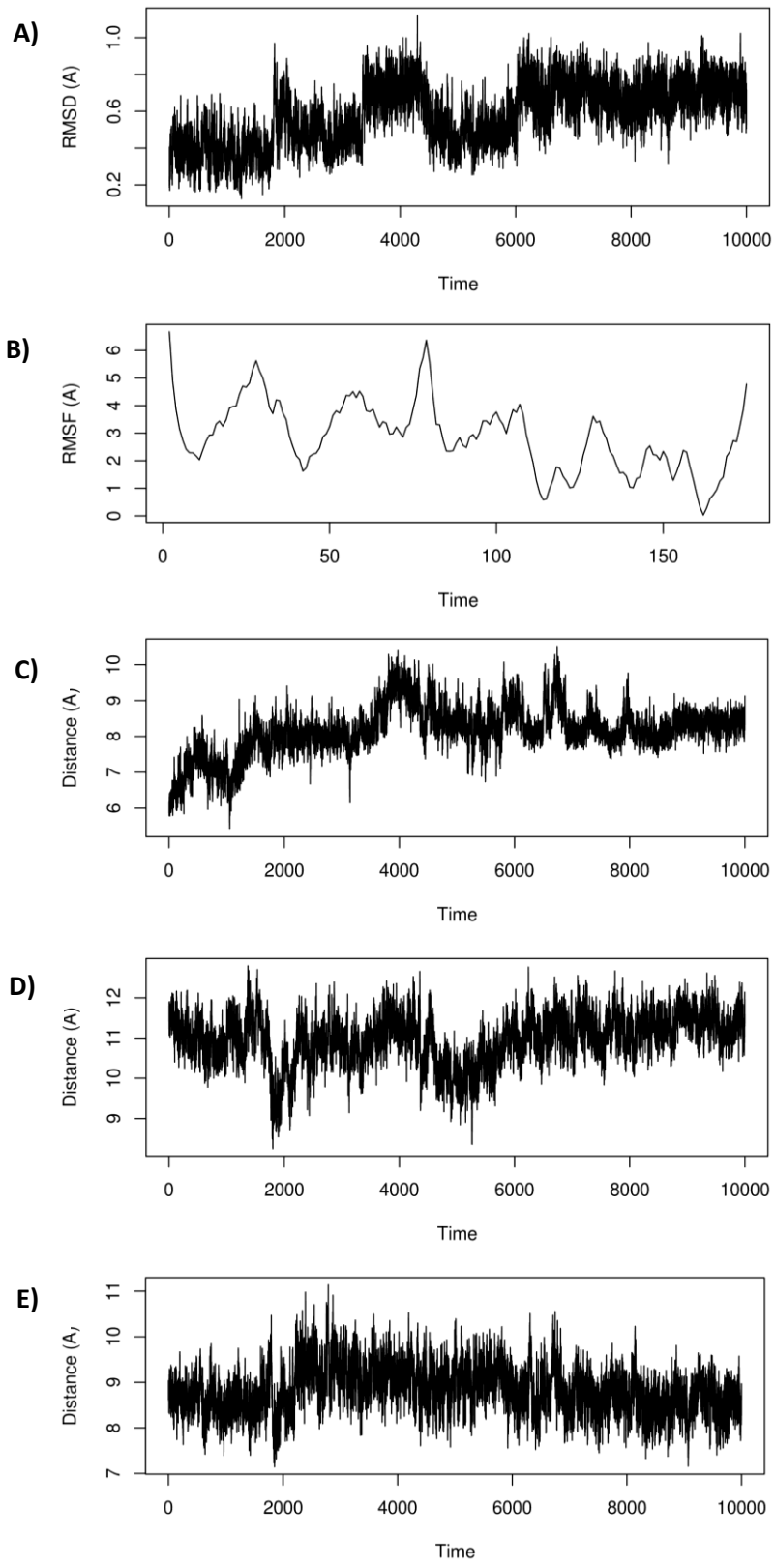


Figure A5.4 Results for 20ns simulation of Model 1 NPT throughout simulation (part 2). Time represented in 2ps frames. **A)** RMSD of I148M containing loop (residues 147-151). **B)** Root mean square fluctuations along the protein chain averaged for full simulation. **C)** Distance between S46 and D166. **D)** Distance between S46 and I148. **E)** Distance between D166 and I148.