RATIONAL DESIGN OF REDUCED ALPHABET PROTEINS

DAVINDER KAUR DHALLA
Master of Science, Pondicherry University, 2014

A thesis submitted
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR SCIENCES

Department of Chemistry and Biochemistry
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

RATIONAL DESIGN OF REDUCED ALPHABET PROTEINS

DAVINDER KAUR DHALLA

Date of Defence: July 25, 2022

Dr. H.-J. Wieden	Professor	Ph.D.
Dr. M. Roussel	Professor	Ph.D.
Thesis Co-Supervisors		
Dr. N. Thakor	Associate Professor	Ph.D.
Thesis Examination Committee Member		
Dr. S. Wetmore	Professor	Ph.D.
Thesis Examination Committee Member		
Dr. P. Dibble	Associate Professor	Ph.D.
Internal External Examiner		
Department of Chemistry and Biochemistry		
Dr. Nediljko Budisa	Professor	Ph.D.
External Examiner		
University of Manitoba		
Winnipeg, Manitoba		
Dr. Jean-Denys Hamel	Assistant Professor	Ph.D.
Chair, Thesis Examination Committee		

DEDICATION

Thank you Almighty, for granting me countless blessings and knowledge, due to which I have been finally able to accomplish the thesis.

This thesis is dedicated to my father, uncle, and grandmother. This wouldn't have been possible without you. Thank you, Daddy, Tatshri, and Mumma!

ABSTRACT

Conventional protein design approaches generally utilize a design space limited to the standard 20 amino acid alphabet (AAA), restricting the incorporation of unnatural amino acids, and associated novel functions. Reducing the standard AAA can free up codon space for the abovementioned purpose along with accelerating *in silico* protein design. However, previously designed reduced AAA protein variants have shown little to no activity relative to their wild-type counterparts. The overarching goal of this thesis was to identify and employ the protein design rules observed in nature for the development of a generalizable RAP design pipeline. Since dynamics are integral to maintaining protein function, data presented in this thesis investigates the structural dynamics of proteins using Molecular Dynamics simulations. This insight into the structural dynamics of proteins assisted in designing RAPs and in the characterization of a receptor targeted by a virus pathogenic to humans, offering potential pharmaceutical and astrobiology applications.

CONTRIBUTIONS OF AUTHORS

Chapter 2 is an advanced draft of a manuscript being prepared for publication. The chapter reports our findings on the amino acid alphabet of proteins, particularly the incorporation of prebiotic amino acids, to understand the evolution of amino acids as life progressed on Earth and to identify the underlying protein design principles utilized in nature. The research was conceptualized by Hans-Joachim Wieden and me. I developed the methodology and software, performed script writing, data curation and processing, formal analysis, visualization, and figure preparation. The manuscript was written by me and Hans-Joachim Wieden.

Chapter 3 is an advanced draft of a format-neutral manuscript being prepared for publication. The chapter showcases our findings on how the structural dynamics of Neuropilin 2 is exploited by the Human Cytomegalovirus to gain entry into the host cells. The research was conceptualized by Hans-Joachim Wieden, me, and Dustin Smith. I carried out the homology modeling and structure validation of the Nrp2-Ca^{2+} and *apo* Nrp2 systems, performed MD simulations, subsequent trajectory, and principal component analysis, and wrote majority of the manuscript. Dustin Smith performed the backbone dihedral angle analysis, wrote associated sections of the manuscript, and contributed to the overall development of the manuscript. Hans-Joachim Wieden aided in the concept development and writing.

Chapter 4 is the early version of a format-neutral manuscript being prepared for publication. In chapter 4, we report our rational reduced alphabet protein (RAP) design pipeline, a combination of molecular dynamics simulations coupled with downstream analysis techniques, employed to design RAP variants for proteins with distinct structures and functions using a two-pronged approach. The research was conceptualized by Hans-Joachim Wieden and me. I developed the methodology and software, performed script writing, data curation and processing, formal analysis, visualization, and figure preparation. The manuscript was written by me and Hans-Joachim Wieden.

ACKNOWLEDGEMENTS

Supervisors: I thank Dr. Wieden for providing guidance during the course of the degree. I also thank Dr. Roussel for accepting to be my co-supervisor towards the end of my degree and for providing a thorough feedback on my thesis chapters.

Abhijeet: I never thought I would be blessed with such an incredibly supportive life partner. Thank you for patiently listening to my endless science talks and for acting excited even though you had little idea what I was talking about. Thank you for enduring me at my highest highs and lowest lows. The support you provided every single day made it possible for me to make it till the end. I couldn't have done it without you.

My family: Daddy and Mumma, thank you for inspiring me. You supported me while you were physically here, and although you may be out of sight now, I am sure you are still watching over me as my guardian angels. I owe everything to you, and I hope I made you proud. Tatshri Ji, thank you for teaching me how to stay focused and disciplined, without which I wouldn't have been able to finish on time. I hope you are proud of me. Aagya, my niece, thanks for being my personal stressbuster throughout stressful times. Steffy, thank you for your constant support as an elder sister and as a friend. I also thank my extended family members for their support throughout the degree.

Nehal and Stacey: Having you both as my committee members has been truly amazing. I thank you both for constantly supporting me and for providing constructive criticism which pushed me to perform better. A big thank you for being my pillars of strength during this journey.

Wieden lab members: Dustin, thank you for being my "computational" pal. I will always remember our discussions on MD, Compute Canada applications, and research ideas. I enjoyed working with you and I wish you good luck for your future endeavours. Josh and Jumai, it was a pleasure working with you both, I wish you all the best. I also thank Harland, Dylan, Jessica, Kristi, Justin, Jalyce, Emily, Fabian, Amanda,

Sydnee, Will, Dora, Luc, Taylor, Fan, and other members of the Wieden lab over the years. Thank you for all the support and good times. I will cherish the memories of working alongside you all.

Susan: Receiving encouraging emails from you is a memory I'll always hold dear. I cannot thank you enough for everything you have done to support me over the years. We the students, fail to acknowledge just how much you do for us behind the scenes, but I want to tell you that I appreciate it and would like to thank you on behalf of all of us. Thank you, work mom!

Special mentions: I thank all my friends and acquaintances in India and Canada for their support throughout my graduate journey. I owe a big thanks to the Campus Coffee Company for providing an uninterrupted flow of coffee. I thank the University of Lethbridge for providing all the scholarships over the years. Thanks to all the janitors and service staff at the UofL campus who worked tirelessly to keep the campus hygienic and safe, particularly during the COVID-19 pandemic. A big appreciation to all the scientists who developed the COVID-19 vaccines, the reason why we made it through the pandemic. To everyone else who has been a part of my graduate journey directly or indirectly, thank you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

µs	microsecond
30S	smaller subunit of a prokaryotic ribosome
3D	three dimensional
50S	larger subunit of a prokaryotic ribosome
A or Ala	Alanine
Å	Angstrom
AA	amino acid(s)
AAA	amino acid alphabet
AMBER	Assisted Model Building with Energy Refinement
apo	Definition: In an inactive, unbound state
Bx(n)	betweenness centrality
C or Cys	Cysteine
Ca^{2+}	Calcium ion with 2 positive charges or Calcium(II) ion
cm	Chorismate mutase
Cryo-EM	Cryogenic Electron Microscopy
D or Asp	Aspartic acid or Aspartate
DNA	Deoxyribonucleic acid
E or Glu	Glutamic acid or Glutamate
EAP(s)	Expanded Alphabet Protein(s)
EF-Tu	Elongation Factor-Thermo unstable
E. coli	*Escherichia coli*
ExPASy	Expert Protein Analysis System
F or Phe	Phenylalanine
fs	femtosecond
G or Gly	Glycine
GABA	gamma-aminobutyric acid
GMQE	Global Mean Quality Estimation
GROMACS	GROningen MAchine for Chemical Simulations
H or His	Histidine
H. sapiens	*Homo sapiens*
HCMV	Human Cytomegalovirus
I or Ile	Isoleucine
IF1	Initiation Factor 1
K or Lys	Lysine
K	Kelvin; unit of temperature
L or Leu	Leucine
LUJV	Lujo virus
M or Met	Methionine
M. jannaschii	*Methanocaldococcus jannaschii*
MD	Molecular Dynamics
mM	millimolar
mRNA	messenger RNA
N or Asn	Asparagine
Na^{+}	Sodium ion with 1 positive charge or Sodium(I) ion
NaCl	Sodium Chloride
nM	nanomolar
Nrp	Neuropilin
Nrp1	Neuropilin 1
Nrp2	Neuropilin 2
Nrp2-Ca^{2+}	Neuropilin bound with Calcium(II) ion
nt	Nucleotide
O or Pyl	Pyrrolysine
P or Pro	Proline

PC	Principal Component
PCA	Principal Component Analysis
PDB	Protein Data Bank
pmemd	Particle Mesh Ewald Molecular Dynamics
Q or Gln	Glutamine
R or Arg	Arginine
RA-CM	reduced alphabet chorismate mutase
RA-IF1	reduced alphabet Initiation Factor 1
RAP(s)	Reduced Alphabet Protein(s)
RA-rpS10	reduced alphabet 30S ribosomal protein S10
RA-variants	reduced alphabet variants
RCSB	Research Collaboratory for Structural Bioinformatics
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RNA	Ribonucleic acid
rpS10	30S ribosomal protein S10
rRNA	ribosomal RNA
S or Ser	Serine
SAP(s)	Small Alphabet Proteins(s)
SARS-CoV-2	Severe Acute Respiratory Syndrome-related Coronavirus-2
SECIS	Selenocysteine insertion sequence element
sdAbs	single domain antibodies
SH3	Src homology 3
SMOG	Structure-based Models for Biomolecules
T or Thr	Threonine
TIP3P	Transferable intermolecular potential with 3 points
tRNA	transfer RNA
U or Sec	Selenocysteine
UAA	Unnatural amino acids
UV	Ultraviolet
UniProtKB	Universal Protein Resource Knowledgebase
V or Val	Valine
VEGF	Vascular Endothelial Growth Factor
VMD	Visual Molecular Dynamics
W or Trp	Tryptophan
wt	wild-type
wt-cm	wild-type chorismate mutase
wt-IF1	wild-type Initiation Factor 1
wt-rpS10	wild-type 30S ribosomal protein S10
Y or Tyr	Tyrosine

CHAPTER 1: INTRODUCTION

"*The nature of life on Earth and the search for life elsewhere are two sides of the same question—the search for who we are. In the great dark between the stars there are clouds of gas and dust and organic matter. Dozens of different kinds of organic molecules have been found there by radio telescopes. The abundance of these molecules suggests that the stuff of life is everywhere. Perhaps the origin and evolution of life is, given enough time, a cosmic inevitability. On some of the billions of planets in the Milky Way Galaxy, life may never arise. On others, it may arise and die out, or never evolve beyond its simplest forms. And on some small fraction of worlds there may develop intelligences and civilizations more advanced than our own. Occasionally someone remarks on what a lucky coincidence it is that the Earth is perfectly suitable for life—moderate temperatures, liquid water, oxygen atmosphere, and so on. But this is, at least in part, a confusion of cause and effect. We Earthlings are supremely well adapted to the environment of the Earth because we grew up here. Those earlier forms of life that were not well adapted died. We are descended from the organisms that did well. Organisms that evolve on a quite different world will doubtless sing its praises too. All life on Earth is closely related. We have a common organic chemistry and a common evolutionary heritage. As a result, our biologists are profoundly limited. They study only a single kind of biology, one lonely theme in the music of life. Is this faint and reedy tune the only voice for thousands of light-years? Or is there a kind of cosmic fugue, with themes and counterpoints, dissonances and harmonies, a billion different voices playing the life music of the Galaxy?*"

— Carl Sagan, Cosmos

1.1. Overview

Most proteins are synthesized in the process of translation from the 20 canonical amino acids. Although we have some understanding of the basic biochemistry of cells, there is still little to nothing known about how life formed on Earth. This thesis attempts to understand the trajectory of evolution of proteins by studying the abundances and distribution of the amino acids in modern day proteins. Additionally, this thesis aims to recognize and extract the protein design principles in nature that would allow alteration of the amino acid composition of proteins, thereby facilitating protein engineering studies.

1.2. Amino Acids

Amino acids are the building blocks of proteins and are combined together in a specific sequence to eventually form a functional protein. All amino acids have a similar general structure containing a central carbon atom (the α-carbon) to which an amino group (NH_3^+ group), a carboxylate group (COO^- group), a hydrogen atom (H), and a R (side chain) group are attached (Figure 1.1).

Amino acids that are incorporated into polypeptides during translation are called proteinogenic or canonical amino acids (1). These include the standard 20 amino acids encoded by the universal genetic code (2) and the non-standard amino acids, selenocysteine (3) and pyrrolysine (4) (twenty first and twenty second amino acids, respectively). Unlike the standard amino acids that are found in proteins from all domains of life, the non-standard amino acids are rarely found in proteins (5, 6). All the amino acids differ from each other in the size, shape, polarity, charge and hydrophobicity of the side chains bonded to their α-carbon atoms (Figure 1.2). The differences in these side chains give the amino acids distinctive physicochemical properties, thereby governing the secondary structure propensities (7) of the amino acids. The diversity among proteins is directly related to the combinatorial possibilities of the 20 amino acids, where the physicochemical properties of the constituent amino acids influence the structure and function of a protein.

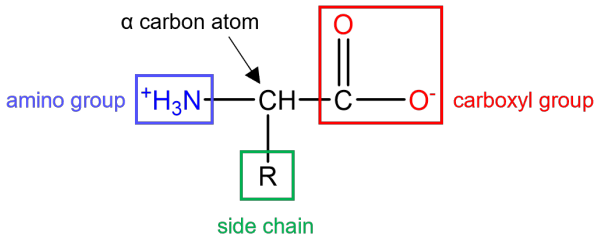


Figure 1.1 – General structure of an amino acid. The alpha(α) carbon atom in the center is linked to an amino group (blue box), a carboxylate group (red box), and a side chain R group (green box). Figure adapted from (8).

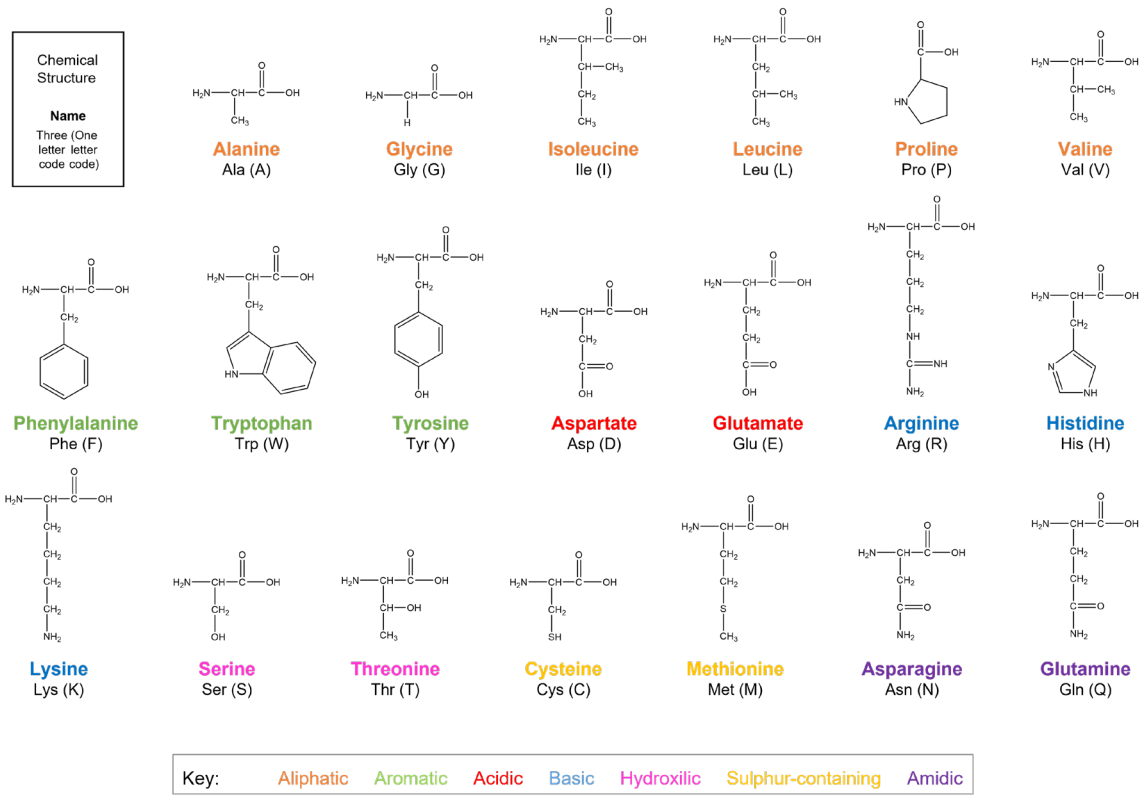


Figure 1.2 – A guide to the twenty standard amino acids. Two dimensional representation of the chemical structures of the twenty standard amino acids. The color coding corresponds to the side chain properties of the amino acids as shown in the key. Figure adapted from (9).

The standard amino acids are incorporated into proteins via a process known as translation or protein synthesis, carried out in the presence of mRNA, tRNAs and ribosomes (10). The ribosomal subunits assemble to form a functional ribosome and to read the genetic information contained in the mRNA. Different tRNAs (in complex with the Elongation Factor Tu) bring specific amino acids to the ribosome in the correct order, and with the help of rRNA, peptide bonds are formed between adjacent amino acids thereby synthesizing the polypeptide chain. However, a few organisms such as bacteria and fungi can also synthesize proteins independently of ribosomes in a process known as non-ribosomal peptide synthesis (11). On the other hand, the incorporation of nonstandard amino acids into proteins requires unique specialized mechanisms. Selenocysteine is encoded using the UGA codon, which usually is a stop codon (12), but is recoded for selenocysteine in an intricate process requiring a selenocysteine insertion sequence element (SECIS element). The selenocysteine-tRNA is delivered to the ribosome by a dedicated elongation factor SelB (13). On the other hand, pyrrolysine found in methanogenic archaea is

coded using the UAG codon, which is also a stop codon, however the incorporation of pyrrolysine uses the standard elongation factor EF-Tu (5).

Apart from the proteinogenic amino acids, more than 500 amino acids exist in nature (14), such as carnitine, GABA, levothyroxine, hydroxyproline and selenomethionine (Figure 1.3) (15, 16) and are known as non-proteinogenic amino acids. Although the non-proteinogenic amino acids are rarely found in proteins (example: hydroxyproline is incorporated into proteins via post-translational modification), they play an important role in cellular metabolic pathways or occur as intermediates in the metabolic pathways of standard amino acids. For example, gamma-aminobutyric acid (GABA) is a neurotransmitter (17), whereas ornithine and citrulline occur in the urea cycle (18), part of amino acid catabolism.

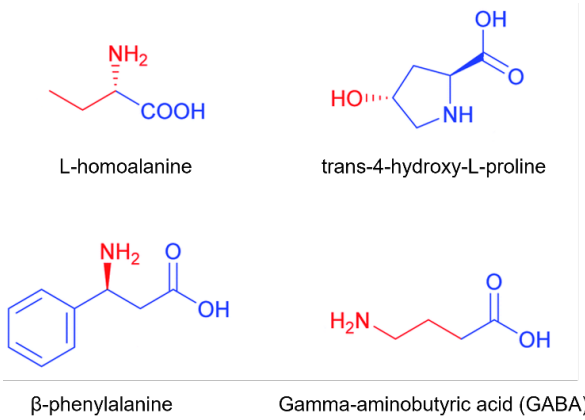


Figure 1.3 – Non-proteinogenic amino acids. Two dimensional representation of the chemical structure of non-proteinogenic amino acids. Figure adapted from (2).

1.3. Proteins

Proteins are the most abundant biomolecules in any living cell, and they play important roles in cellular structure and functions. The term protein is derived from the Greek word *πρώτειος* (*proteios*), meaning "primary" or "in the lead" (19). As the name suggests, proteins are at the forefront of a wide array of cellular functions, such as catalyzing metabolic reactions, DNA replication, maintaining cellular structure, and transporting molecules to their target cellular locations. The primary structure of a protein is comprised of a linear chain of amino acids where the individual amino acids are bonded together by peptide bonds. A peptide bond is formed when

the carboxyl group of one amino acid reacts with the amino group of the adjacent amino acid, releasing a water molecule in a reaction known as condensation or dehydration (Figure 1.4).

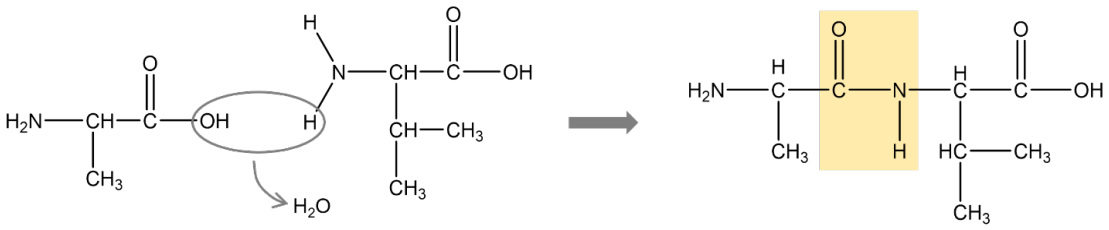


Figure 1.4 – Peptide bond formation. The carboxyl (COOH) group of one amino acid reacts with the amino (NH_2) group of the adjacent amino acid to eliminate a water molecule resulting in formation of peptide bond highlighted in yellow.

Once linked together by peptide bonds, the amino acids are called residues, and the chain of amino acids is called the polypeptide chain. The NH_2 end with a free amino group is known as the amino terminus or N-terminus, whereas the COOH end of the protein with a free carboxy group is referred to as the carboxy terminus or C-terminus (the sequence of the protein is read from N-terminus to C-terminus, left to right) (Figure 1.5). The polypeptide chain, which constitutes the primary structure of a protein, plays a crucial role in determining the physical and chemical properties of a protein. Additionally, depending on the constituent amino acids, the primary structure drives the folding of the polypeptide chain including the formation of intramolecular bonds, ultimately determining the three dimensional shape and structure of the protein.

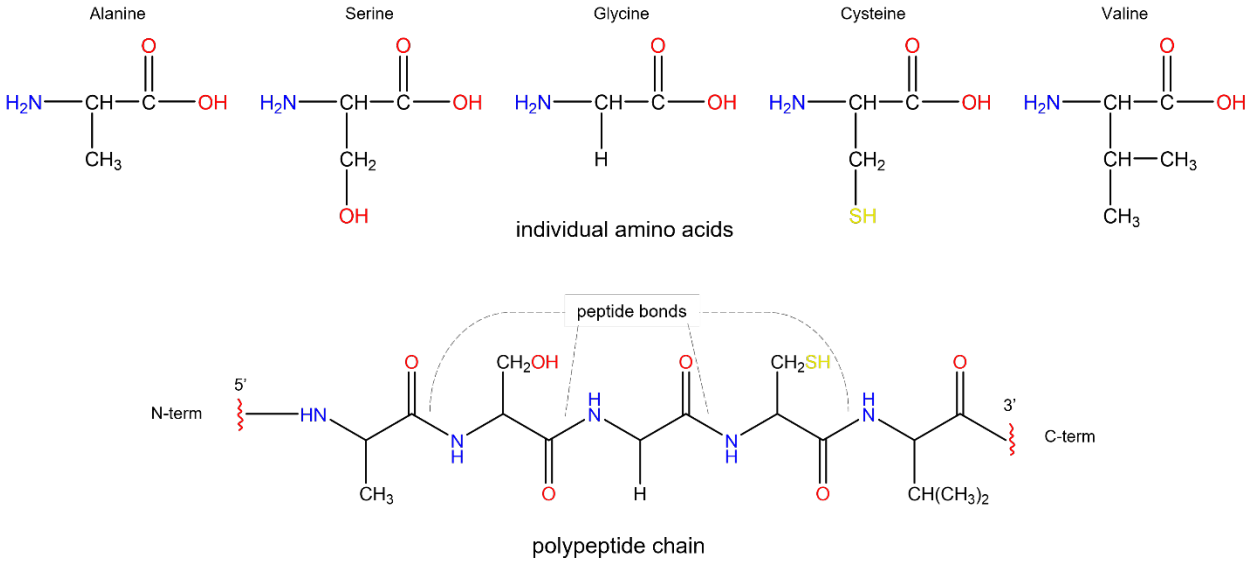


Figure 1.5 – General chemical structure of a polypeptide chain. The constituent amino acids are joined by peptide bonds by undergoing a condensation reaction involving release of a water molecule. Figure adapted from (8).

The secondary structure of a protein comprises local structures stabilized by hydrogen bonds. The most common secondary structures found in proteins include the α-helix, β-sheet, loops and turns. The next level of structural organization in proteins is known as the tertiary structure which includes the overall shape of a single protein molecule, notably the spatial relationship of the secondary structures to one another. Tertiary structure of a protein is stabilized by interactions such as salt bridges, hydrogen bonds, hydrophobic interactions, disulfide bonds, and even post-translational modifications The tertiary structure is the highest level of structural organization for a monomeric protein. However, in case of multimeric proteins that are formed by several polypeptide chains, the quaternary structure is the final level of organization, where multiple tertiary folded structures function together as a single protein complex (Figure 1.6).

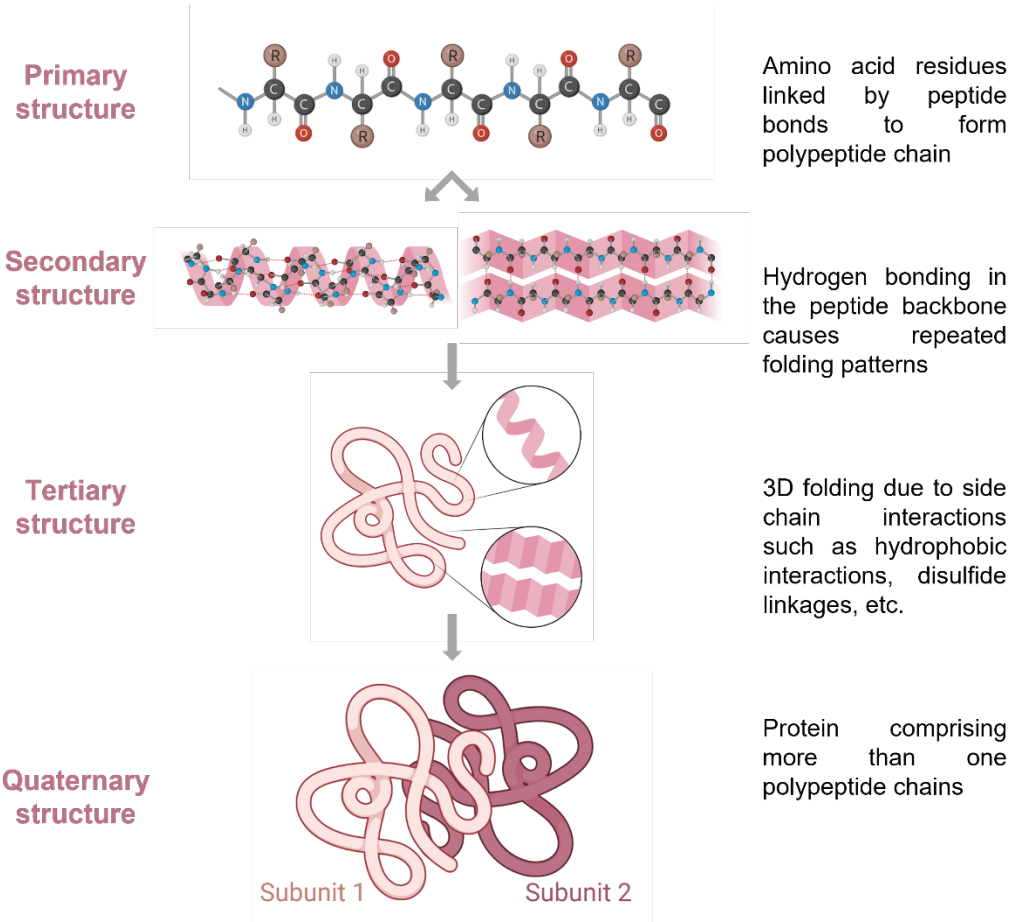


Figure 1.6 – Structural organization in proteins. The primary structure of proteins consists of a linear polypeptide chain in which residues are linked together by peptide bonds. The secondary structure of proteins is due to hydrogen bonding between the atoms of the polypeptide backbone. The tertiary structure of proteins arises from non-covalent interactions as well as disulphide interactions, ionic interactions, etc. The quaternary structure of proteins involves the formation of a complex between two or more polypeptide chains. Figure adapted from (20).

1.4. Evolution of the amino acid alphabet of proteins

The particular set of different amino acids utilized in each protein sequence is referred to as the amino acid alphabet (AAA) of the respective protein. Most proteins include all twenty amino acids, and therefore their AAA is a standard alphabet. However, several proteins such as keratin include only a subset of the twenty standard amino acids and therefore their AAA is smaller than the standard alphabet. We refer to these proteins as smaller alphabet proteins (SAPs). On the other hand, proteins such as thioredoxin reductases and methyltransferases naturally include the twenty first (selenocysteine) and twenty second (pyrrolysine) amino acids respectively, and therefore have an AAA larger than the standard alphabet. We refer to such proteins as expanded

alphabet proteins (EAPs). There is a wide variety in the AAA sizes of the proteins found in nature ranging from as small as a 3 AAA to 21 AAA (see Chapter 2, section 2.5.1). The occurrence of such a diverse range of AAA sizes indicates the current AAA is perhaps a product of evolution over billions of years (2, 21-24). Several previous studies in the fields of astrobiology, protein engineering, bio-evolution, and abiotic chemistry, along with the RNA world hypothesis have indicated that the AAA could have included fewer amino acids in the early Earth environment, containing only a subset of the current standard amino acids, which gradually evolved and expanded into the current AAA (2, 23, 25-27). According to previous evolutionary studies, this smaller subset of the standard amino acids including Gly, Ala, Asp, Ile, Val, Ser, Pro, Glu, Leu, and Thr constituted the early genetic code, and were the first set of amino acids to form in the prebiotic conditions of early Earth (28-30). Therefore, these amino acids are referred as the prebiotic amino acids (29). On the other hand, Cys, Met, Tyr, Trp, His, and Phe are believed to have been added later on into the genetic code. Additionally, Trp is speculated to be the last amino acid to be incorporated into the genetic code (31, 32). However, even after several decades of protein evolutionary studies there is no concrete evidence to prove these theories. Furthermore, it is unclear whether the structure, function or evolutionary age of a protein is reflected in the constituent amino acids and the AAA size. Therefore, studying the AAA will help us in understanding not only the evolution of proteins but also the underlying principles of protein design applied in nature. This knowledge can be utilized for the forward engineering of proteins to unlock additional functions in a near natural manner.

1.5. Tweaking the amino acid alphabet: An approach to protein engineering

Since a large number of proteins are naturally composed using an AAA smaller than the standard 20 AAA, this has inspired protein engineers to artificially reduce the AAA of the proteins of interest and to design reduced alphabet proteins (RAPs). Apart from offering an understanding of the evolution of proteins, reducing the AAA of proteins offers several other advantages. For instance, there is a high degree of overlap in the physicochemical properties of amino acids such as there being two acidic amino acids (Asp and Glu), and three aromatic amino acids (Phe, Trp,

and Tyr), etc. One or more of such overlapping amino acids can be removed from the AAA to obtain RAP variants of those proteins, as done in previous studies (Table 1.1) (33-37). Another advantage of reducing the AAA is that in place of the removed amino acid, any unnatural amino acid (UAA) of interest can be incorporated by utilizing specially engineered cells (38) to provide novel features to a target protein, thus yielding proteins with user-defined functionalities such as spectroscopic probes, UV-inducible crosslinkers, and functional groups for posttranslational modifications, etc. (39, 40) (Figure 1.7).

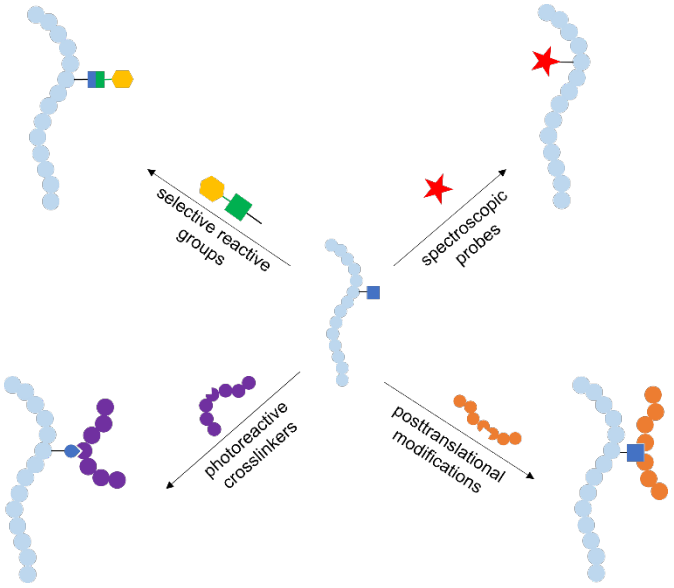


Figure 1.7 – Applications of unnatural amino acids with customized side chains. UAAs can be used to assign selectively reactive groups, photoreactive crosslinkers, posttranslational modifications, or desired probes. Figure adapted from (39).

More than 160 UAAs have been genetically encoded into the proteins to date in different organisms for various applications (39). Past studies have identified unnatural chemical groups which are uniquely reactive and can be used for site-selective labeling and modification of proteins. Such groups are useful in the preparation of protein tags which provide unprecedented control over protein labeling sites (41). On the other hand, incorporation of spectroscopic probes allows for tracking the protein structure and localization, and to detect any changes in the chemical environment of the target protein (42). Furthermore, targeted posttranslational modifications, or insertion of structural mimics allow for modification of proteins at the residue level, permitting personalized variations in the target protein (43). Lastly, with the insertion of photoreactive groups

and crosslinkers, one can extract structural and localization information of the target protein, thus providing a powerful tool for biomolecular mapping (44).

In conventional protein engineering and design, the sequence of the parent protein is reorganized to generate variants. For a protein that is 100 amino acids long with a standard 20 AAA, the total number of variants to be screened would be 100^{20}, which is practically impossible (45), whereas for a RAP, limiting the AAA size would reduce the total number of variants to be screened, thus accelerating the protein design process. Taking input from the natural SAPs, artificially designed RAPs can help us to identify the 'bare minimum' required to make functional proteins which would find applications in designing bacterial strains to run on reduced alphabet proteins. The resulting simplified enzyme components may give rise to more productive enzymatic pathways since amino acid flux will be re-directed from self-replication and synthesis of the eliminated amino acids to maximal product yield, a feature particularly useful during nutrient deficient conditions. Therefore, such organisms may act as the pioneers in setting up life on remote planets with hostile conditions where the availability of nutrients is extremely limited. However, we acknowledge that to enable the generation of such RAP strains, the elimination of amino acid(s) needs to be performed at the cellular level which would in turn affect the homeostasis of the resulting strain, a critical characteristic to be addressed when engineering organisms. One of the applications of such engineered strains are *in vivo* systems where the reduced energy requirement for cellular metabolism in an engineered strain would allow increased recombinant product yield. Such tools can be used for tasks such as bioremediation, point-of-need vaccine production, etc.

With such a wide range of applications outlined above, RAP design has become an increasingly ventured field in the past few years. Several studies have attempted RAP design on a range of diverse proteins, where they focused primarily on preserving the structure of the RAP variant (Table 1.1). However, of all the previous work on designing RAPs, most have resulted in partial/total loss of function (Table 1.1). We attribute this failure to neglecting the protein design rules in nature and the failure to preserve protein dynamics while designing RAPs. Since proteins are dynamic entities, altering the structure (including the primary structure, i.e., the amino acid sequence) would also impact the function of the protein. Therefore, we predict that it is important

to preserve both the structure and the dynamics of a protein to improve the chances of designing a functional RAP, unlike prior studies where researchers only focus on preserving the structure of the RAPs.

Table 1.1 – Summary of previous work done on reduced alphabet protein design.

Research group	Test protein	AAA size	Strategy of AAA reduction	RAP variant function
Akanuma *et al.*	Orotate phospho ribosyltransferase (OPRTase)	9	Hydrophobic, Acidic, Basic, Aromatic, Flexibility, Rigidity, Thr	Did not test function
Akio *et al.*	Green Fluorescent Protein (GFP)	19	Trp is the last amino acid to be incorporated during evolution	Complete loss of function
Kamtekar *et. al*	α-helical bundle proteins	8	Binary pattern module polar:non-polar	Did not test function
Kimura M. & Akanuma S.	Nucleoside diphosphate kinase (NDK)	13	Physicochemistry based substitution of non-catalytic residues	Complete loss of function
Kuroda and Kim	bovine pancreatic trypsin inhibitor (BPTI)	18	Alanine substitutions	Did not test function
Muller *et al.*	Chorismate mutase (cm)	11	Binary pattern module polar:non-polar + T, V	Partial loss of function
Regan and DeGrado	α-helical bundle proteins	8	Binary pattern module polar:non-polar	Did not test function
Riddle *et al.*	SRC homology 3 domain (SH3)	16	Combinatorial mutagenesis	Did not test function
Walter *et al.*	Chorismate mutase (cm)	14	Binary pattern module polar:non-polar	K_{cat} 3 fold lower, less stable

1.6. Hypothesis and Objectives

This thesis aims to understand the amino acid alphabet of proteins in nature and to identify the principles of protein design in nature which can be further utilized for the rational reduction of

the AAA of proteins to design RAPs. I hypothesize that these principles can be identified by interrogating the AAA of natural proteins and extracting the underlying design features that are at the heart of biomolecular function including folding, regulatory and enzymatic properties. This thesis further theorizes that an *in silico* molecular dynamics investigation pipeline coupled with downstream computational analyses is capable of studying protein structure and dynamics that enables efficient design of functional RAPs.

In Chapter 2, we investigate the AAA of diverse proteins from different organisms belonging to various domains of life by performing bioinformatics analysis on protein sequence entries available on the UniProt protein sequence database. Our findings showcase the distribution and size of AAA of different proteins in nature along with the design rules followed in nature depending on the function or evolutionary age of the protein. We show for the first time that the AAA size expands with increasing length (in terms of amino acid residues) of proteins and that the amino acid compositions are customized to suit different functions of proteins. Chapter 3 demonstrates our in-house developed protein dynamics analysis toolkit where molecular dynamics simulations coupled with computational analyses techniques help reveal unforeseen dynamics in Nrp2 with applications in development of antiviral therapies. Lastly, with the insight into protein design rules from chapter 2 and employing the computational protein dynamics assessment approach outlined in chapter 3, chapter 4 demonstrates the rational design of RAP variants for proteins with distinct structure and function to prove that the AAA size of proteins can be significantly reduced without compromising protein structure or dynamics, thereby generating reduced alphabet variants that have a higher likelihood of retaining function. We employ molecular dynamics simulations coupled with computational analysis techniques to investigate and compare the dynamic properties of the RAP variants with respect to the wild-type proteins, followed by ranking the RAP variants. Altogether this thesis will provide information on the amino acid alphabets of proteins and how *in silico* techniques can be employed to rationally reduce the alphabet sizes of proteins to design functional RAPs.

CHAPTER 2: LOOKING BACK IN TIME: COMPLEXITY OF AMINO ACID ALPHABETS REVEAL ORIGIN OF NATURE'S PROTEIN DESIGN PRINCIPLES

2.1 Preface

This is an advanced draft of a manuscript being prepared for publication. This chapter showcases our findings on the amino acid alphabet of proteins, particularly the incorporation of prebiotic amino acids, to understand the evolution of amino acids as life progressed on Earth and to identify the underlying protein design principles utilized in nature. The research was conceptualized by Hans-Joachim Wieden and me. I developed the methodology and software, performed script writing, data curation and processing, formal analysis, visualization, and figure preparation. The manuscript was written by Hans-Joachim Wieden and me.

2.2 Abstract

The primary structure of proteins, their amino acid sequence, is derived from a common set of 20 amino acids, referred to as the standard amino acid alphabet (AAA). However, primordial life likely was simpler. To investigate if the current 20 AAA evolved from a smaller initial alphabet and if traces of it can be found in present-day proteins, we performed a comprehensive bioinformatics study of protein entries obtained from the UniProtKB/Swiss-Prot database. Our results reveal that proteins with smaller AAAs are still widely found in nature. In agreement with the prebiotic amino acid theory our findings show that the prebiotic amino acids are central to the AAA of present-day proteins, thus suggesting that the standard alphabet indeed has evolved from a smaller alphabet. We also demonstrate that the AAA complexity increases with the length (in terms of constituent amino acid residues) of protein sequences, the missing link supporting the theory of gradual expansion of AAA size over time. Our data suggests a critical role of the expanded AAA not only for diversification of physicochemical properties, but also enabling formation of larger proteins. Additionally, analysis of proteins with common functions reveals specific amino acid preferences, reflecting the underlying selection mechanisms and driving force for AAA diversification. In

summary, the data reported here provides important insight into the evolution of the present-day AAA, the genetic code, and previously overlooked origins of design principles in naturally occurring proteins.

2.3 Introduction

Proteins are complex and versatile biomolecular machines that perform a vast array of functions in nature, including highly selective chemical catalysis and critical regulatory as well as structural roles. This remarkable functional diversity is achieved through differences in the amino acid sequence and length (the primary structure) of proteins, which subsequently gives rise to different secondary, tertiary, and quaternary structures. Therefore, the amino acids utilized as building blocks of proteins play a central role in determining the three-dimensional structure and function of the respective protein, and ultimately reflect the evolution of proteins. The particular set of different amino acids utilized in each protein sequence is referred to as the amino acid alphabet (AAA) of the respective protein. Over the course of evolution, nature arrived at a standard set of amino acids with different physicochemical properties which are utilized differentially to derive specific protein structure(s) and function(s). For most of the present-day proteins this set consists of twenty genetically encoded proteinogenic amino acids (20 AAA). The primordial proteins likely consisted, limited by the amino acids available on primordial Earth, of a subset of this alphabet (46). This smaller subset of amino acids contains ten amino acids, alanine (A), aspartate (D), glutamate (E), glycine (G), isoleucine (I), leucine (L), proline (P), serine (S), threonine (T) and valine (V), and are known as the prebiotic amino acids. Although previous studies have demonstrated that smaller alphabet proteins consisting solely of prebiotic amino acids are able to form stable secondary and tertiary structures, it has been speculated that a progressive expansion of a smaller alphabet allowed increased conformational stability, specificity, and functional diversity (47).

Several theories have been put forward regarding the evolution of the genetic code and the evolution of the present day standard amino acids. Theories on the origin and evolution of the genetic code suggest that the selection of codons in the genetic code is non-random and governed by concepts such as stereochemistry, coevolution and error minimization (48). Other theories such

as the Alanine World model propose that the present day standard amino acids are derivatives of alanine (49, 50). The RNA world hypothesis on the other hand states that the modern genetic system (consisting of DNA, RNA, and proteins) was preceded by a simple RNA-only genetic system in the early days, where RNA performed both storage of genetic information and catalytic functions (51-56). Alternatively, the frozen accident theory proposed by Francis Crick argued that the codons were probably brought into use gradually until all of the twenty standard amino acids were incorporated into the genetic code (57-60) and proposed that any change to this code is deleterious. Although these theories attempt to address how the present day genetic code may have evolved; it is still unclear how only the current twenty standard amino acids became incorporated into the proteins and why not others. Further, it is not fully understood if the standard 20 AAA constitutes an optimized minimal alphabet required for the functional and structural diversity of current life's protein complement, or whether it contains redundancies, which if identified and removed could reduce the AAA complexity without jeopardizing the function of the resulting proteins and ultimately life as we know it.

To address these questions, we performed an in-depth bioinformatics analysis of all protein entries from the UniProt/Swiss-Prot database and mapped the AAA complexity of proteins across all domains of life. We wanted to know how prevalent small AAAs are today and if a common smaller alphabet can be identified between proteins from different life forms. We therefore analyzed the distribution of each amino acid in AAAs across different proteins and protein sub-groups. By doing so, we shed light on the relation between the amino acid alphabet composition and the function of the protein. Furthermore, we identify trends describing the inclusion or exclusion of amino acids in the AAA, particularly from a function and species perspective. Lastly, we describe a generalizable smaller AAA for specific organisms which could be used for the forward engineering of novel proteins.

2.4 Methods

2.4.1 UniProt/Swiss-Prot derived protein sequence dataset: The Swiss-Prot database (release 2021_02) has been chosen as the source of protein sequence information for this work. A

combination of Boolean operators was used to extract all protein sequence entries only from the Swiss-Prot component of the UniProtKB, which constitutes our raw dataset. The raw dataset underwent the cleanup process mentioned below to obtain our final Swiss-Prot derived dataset. In this work, all polypeptides with a sequence length of 29 amino acids and longer are classified as proteins, although this cutoff is different and lower compared to the cutoff defined by several biochemistry authors (61). Our cutoff is based on the reported average length of small proteins in bacteria of 23 amino acids (62), and the sequence length of glucagon (smallest functional proteins in Eukaryotes) which is 29 residues.

2.4.2 Scripting for file cleanup and analysis: All the scripts used for this analysis have been written in the Perl and Bash programming languages (63, 64). The first Perl script performs the initial process of removing duplicates and fragments to eliminate redundant and incomplete entries from the raw data (Appendix Figure 2.1). This process called data cleanup eliminated 93,675 entries as fragments and duplicates. The final Swiss-Prot derived dataset contains 460,602 entries corresponding to only unique full length protein sequences, carried forward to the next step for analysis. The second Perl script counts the number of occurrences of each amino acid for each entry in our dataset. Thus, for each entry, the script returns the amino acids present in the AAA, the population of each amino acid and the list of absent amino acids. In the last part, the script compiles the list of all protein sequence entries and arranges them according to the AAA size, entries with largest alphabet sizes on top and entries with the smallest alphabets at the bottom. We have also designed our scripts to returns the UniProt identifier of the entries to make the identification and further studies of entries of interest easier. The output generated after the Perl analysis is the input to a bash (65) script for clustering the entries based on their AAA size. The bash script takes the output file from the previous Perl analysis and bins all the entries corresponding to a given AAA size, thus reporting the different AAA sizes and the number of protein entries with that particular alphabet size. The above stated combination of scripts is used to study the trends of amino acid preferences and exclusions observed in protein entries from different domains of life, functional sub-groups, or species.

2.4.3 Amino acid binning/grouping: To understand the exclusion trends of different amino acids in our overall Swiss-Prot derived dataset, the exclusion fraction values are used to bin the amino acids into separate groups (Table 2.2). A conservative cutoff of 0.01 exclusion fraction value (equal to 1%) is deemed appropriate to segregate the core amino acids from the rest. Since the total number of entries in our Swiss-Prot derived dataset is 460,602, one percent of the dataset would include 4,606 sequence entries. For any given amino acid this means that the given amino acid would need to be absent/present in 4,606 entries to change its amino acid exclusion fraction value by 0.01. Because the chances of such a large change happening at random is negligible, 0.01 is established as the cutoff to segregate the core amino acids from other amino acids with higher exclusion fraction values. Subsequently, as the exclusion fraction values increase, the cutoffs between different groups starts becoming clearer with more drastic differences in the exclusion fraction values of the amino acids.

2.5 Results

As a starting point for studying the conservation and utilization of differentially complex amino acid alphabets in present-day proteins, we opted for the UniProtKB/Swiss-Prot database (66) as our source of protein sequences. Swiss-Prot (67) is the curated component of the UniProt database and contains high quality manually annotated protein sequence entries. Therefore, all protein sequence entries deposited in Swiss-Prot were included in our raw dataset (Appendix Figure 1). After removing the duplicates and fragments, our Swiss-Prot derived dataset contained a total of 460,602 unique sequence entries which constitutes 82% of all Swiss-Prot entries. Subsets of these sequences were extracted for different protein classes to analyze the distribution of AAAs in proteins from different organisms and functional classes.

2.5.1 Small alphabet proteins are widely found in nature: In order to investigate the utilization of various AAA sizes in present-day proteins, we performed a comparative analysis of the AAA variability across all proteins in the Swiss-Prot derived dataset (n=460,602) using in-house

developed scripts for the analysis (see Methods). We found that 11% (n=31,703) of all proteins are smaller alphabet proteins or SAPs (with a reduction of the AAA size by at least 2 amino acids, from 20 AAA to 18 AAA) (Figure 2.1.A). The smallest AAA for any protein in our Swiss-Prot derived dataset is a 3 AAA (sperm protamine P3 from *Murex brandaris* (eukaryote, UniProt ID: P83213) where the 54 amino-acid protein is composed of only glycine, lysine, and arginine).

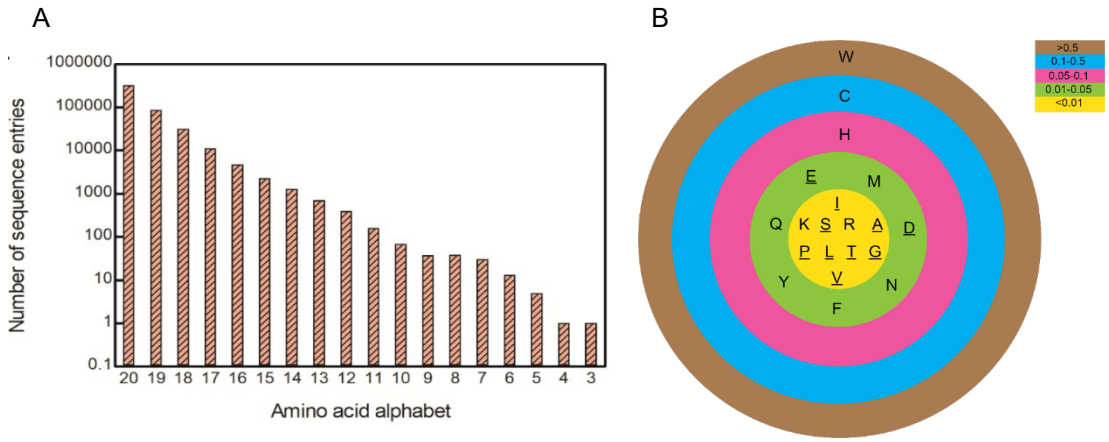


Figure 2.1 – Amino acid alphabet distribution in protein sequence entries in the Swiss-Prot derived dataset (n=460602). A. The x axis shows the amino acid alphabet, and the y axis shows the population of proteins corresponding to the given amino acid alphabet. B. Bull's eye plot showing the exclusion fraction (E_f) values for different amino acids. The amino acids with highest E_f values are shown in the outermost circle and vice versa. The core amino acids (with E_f value ≤ 0.01) are shown on the innermost circle (prebiotic amino acids are underlined).

2.5.2 Smallest alphabet sizes vary between domains of life: To identify the smallest alphabet size utilized by proteins, we analyzed the SAPs from different domains of life (Table 2.1). The smallest alphabet size among archaeal proteins in our dataset is 10 AAA (50S ribosomal protein L41e from *Pyrococcus furiosus* (Q8U232)). Similarly, the smallest alphabet size among the bacterial proteins in our dataset is 10 AAA (found in twelve proteins across nine different species of bacteria). For viruses, the smallest alphabet size is 8 AAA (DNA-binding protein from *Autographa californica nuclear polyhedrosis virus* (P06545)). Surprisingly, in our dataset the corresponding AAA size for a eukaryotic protein is 3 AAA, found in sperm protamine P3 from *Murex brandaris* (P83213).

Table 2.1 – Proteins with smallest AAA sizes across different domains of life. The size of the smallest alphabet is accompanied by the name of the protein(s), length of the protein sequence(s), and the parent organism(s)/species.

Domain	Smallest Alphabet Size	Protein name	Sequence length	Organism
Archaea	10	50S ribosomal protein L41e	37	*Pyrococcus furiosus*
Bacteria	10	50S ribosomal protein L34	45	*Prochlorococcus marinus*
		Cytochrome b6-f complex subunit 6	32	*Mastigocladus laminosus*
		Lantibiotic paenibacillin	30	*Paenibacillus polymyxa*
		Small toxic protein TisB	29	*Escherichia coli*
		SPBc2 prophage-derived membrane protein YosA	39	*Bacillus subtilis*
		Spore coat protein C	66	*Bacillus subtilis*
		Uncharacterized membrane protein YczM	29	*Bacillus subtilis*
		Uncharacterized membrane protein YuzJ	43	*Bacillus subtilis*
		UPF0391 membrane protein GbCGDNIH1_2123	59	*Granulibacter bethesdensis*
		UPF0391 membrane protein Pnap_0032	61	*Polaromonas naphthalenivorans*
		UPF0391 membrane protein XC_2938	57	*Xanthomonas campestris*
		UPF0391 membrane protein XOO1885	57	*Xanthomonas oryzae*
Eukaryota	3	Sperm protamine P3	54	*Murex brandaris*
Virus	8	DNA-binding protein	55	*Autographa californica nuclear polyhedrosis virus*

2.5.3 Amino acid exclusion hierarchy reveals a 'core' set of amino acids: Based on the observation that smaller AAAs are frequently found in nature, we wanted to know if preferences exist with respect to which of the 20 proteinogenic amino acids are not included in the smaller alphabet proteins. If the amino acids are not excluded randomly but based on their, for example, biochemical and biophysical equivalency, a hierarchy should be identifiable in which certain amino acids are more likely to be excluded from a protein's AAA than others. To this end, we analyzed the amino acid alphabets of the proteins in our dataset and calculated the probability for a particular

amino acid to be absent in smaller AAAs (18 AAA and smaller) and termed this value exclusion fraction (E_f). The exclusion fraction is defined as the fraction of protein sequences that do not contain that particular amino acid in their AAA. For example: in a dataset of 100 proteins, if a certain amino acid is absent in 45 proteins, its exclusion fraction would be 0.45. Our results show that every amino acid has a different exclusion fraction value (Table 2.2). Interestingly, these fall into 5 major classes when grouping amino acids with similar E_f values. The bull's eye plot in Figure 2.1.B representing these 5 classes illustrates that the amino acids tryptophan (W) and cysteine (C) have the highest exclusion fraction values ($E_{f,W}$ = 0.54 and $E_{f,C}$ = 0.22, respectively) by placing them at the periphery of the plot. On the other hand, some amino acids are highly unlikely to be excluded from a smaller AAA, i.e., these amino acids are almost always present in proteins and have corresponding E_f values of less than 0.01 and are therefore located in the center of the bull's eye plot (Figure 2.1.B). We describe this set of amino acids as the 'core' amino acids. In our Swiss-Prot derived dataset, this set of ten core amino acids includes **alanine**, **glycine**, **isoleucine**, lysine, **leucine**, **proline**, arginine, **serine**, **threonine**, and **valine**, representing a diverse range of physicochemical properties including aliphatic, charged, neutral, and non-polar. It is also interesting to note that out of the ten core amino acids, eight are prebiotic amino acids (indicated in bold in the list above), the set of amino acids presumed to be present on early Earth (47). It is surprising that the negatively charged prebiotic amino acids aspartate and glutamate are not found in this core group. The fact that the core set includes positively charged amino acids lysine and arginine ($E_{f,K}$ = 0.009 and $E_{f,R}$ = 0.003, respectively), but no negatively charged amino acids might indicate an underlying requirement of the core set to facilitate interactions with negatively charged environments or interaction partners such as RNA. The latter would be consistent with the RNA-world hypothesis (52, 68). Thus, our analysis shows that a general set of core amino acids can be identified in present-day proteins that contains amino acids with diverse physicochemical properties, most of which are prebiotic amino acids. Our results therefore strongly support the hypothesis that the present-day standard amino acid alphabet has evolved from the expansion of a smaller AAA that existed on early Earth.

Table 2.2 – Exclusion fraction (E_f) values of the twenty proteinogenic amino acids for the protein sequence entries in the Swiss-Prot derived dataset. The color coding of the amino acids and E_f values corresponds to the bull's eye plot in Figure 2.1.B.

Amino acid	Exclusion fraction value
W	0.54
C	0.22
H	0.067
Y	0.038
Q	0.023
F	0.019
N	0.017
D	0.015
E	0.012
M	0.011
K	0.0093
P	0.0079
I	0.0048
T	0.0033
R	0.0031
A	0.0023
G	0.0021
V	0.0018
L	0.0011
S	0.0011

2.5.4 The centrality of prebiotic amino acids is conserved across all domains of life: We subsequently wanted to investigate if the centrality of prebiotic amino acids as core amino acids identified above is a common feature of present-day proteins or is dependent on the evolutionary lineage. Therefore, we assessed the distribution of prebiotic amino acids in proteins from different domains of life (69, 70). For this, protein sequence entries from archaea, bacteria, and eukaryotes were analyzed with respect to their AAA exclusion fraction values. Because it has been argued that viruses represent the fourth domain of life (71), we also included sequence entries of viral proteins. Independently of the sequence origin (archaea, bacteria, eukaryote or viral), prebiotic amino acids have some of the lowest E_f values of ≤0.0052, ≤0.0071, ≤0.0066, ≤0.0088 in archaea, bacteria, eukaryotes, and viruses, respectively (Appendix Table 2.1, Figure 2.2). This suggests that the prebiotic amino acids indeed hold a central role in facilitating the structure and/or function of present-day proteins. We further investigated the distribution of prebiotic amino acids in smaller

alphabet proteins and observed that prebiotic amino acids are present even in proteins with the smallest alphabets of 3 AAA and 4 AAA (Appendix Figure 2). This indicates that prebiotic amino acids are likely foundational for protein folding and function even for present-day proteins with the smallest AAAs. Consistently with this and to add additional functionality to a protein, as the alphabet size increases, more prebiotic amino acids are included into the AAA of proteins (Appendix Figure 2.2). It is interesting to note that all ten prebiotic amino acids become part of the AAA pool already when the alphabet expands from 3 AAA to only 6 AAA, which further supports their critical role for maintaining protein structure and/or function and suggests a certain degree of redundancy of biophysical properties that give rise to their utilization in different combinations. It is only when the AAA reaches a size of 13 that all ten prebiotic amino acids appear together in a single protein, further supporting the hypothesis of redundancy and the central role of the prebiotic amino acids for present-day proteins, regardless of the size of the AAA.

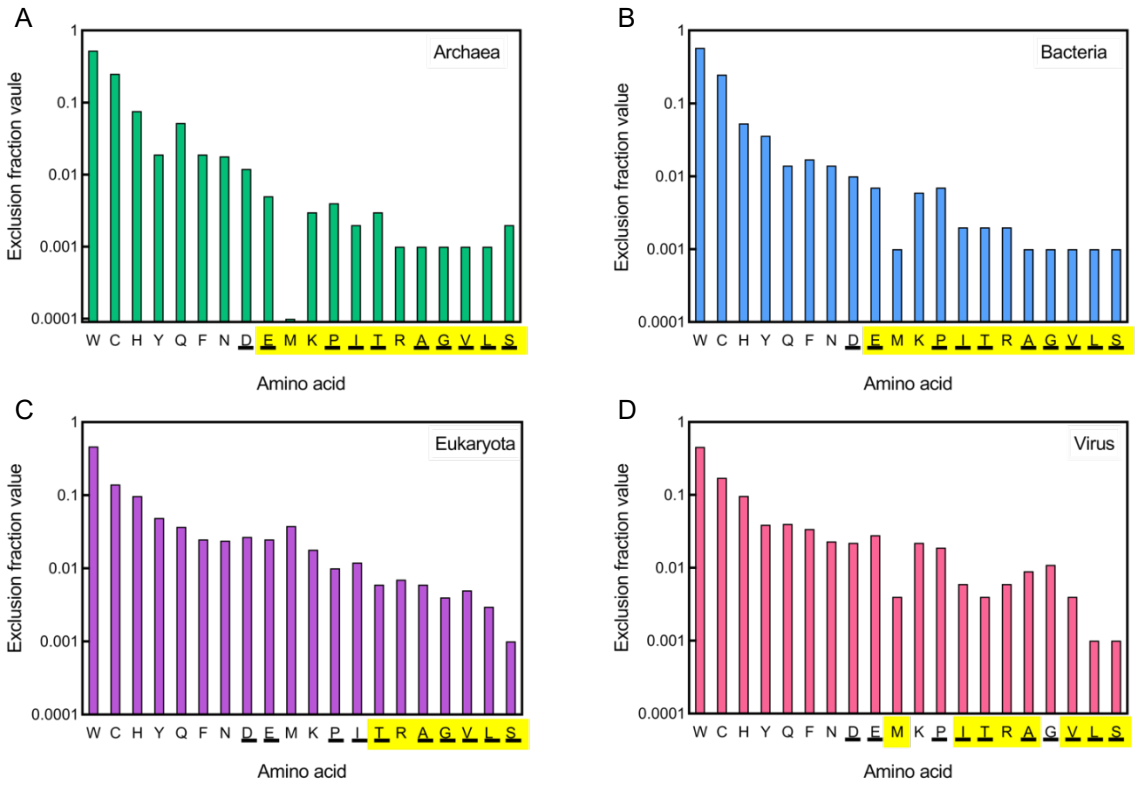


Figure 2.2 – Exclusion fraction (E_f) values of the standard twenty amino acids in proteins from different domains of life. A. Archaea B. Bacteria C. Eukaryota D. Virus. The amino acids on the x axis are arranged in order of the general exclusion trend observed for the Swiss-Prot derived dataset (Table 2.2), with the prebiotic amino acids underlined. The core amino acids are highlighted on the x axis. The y axis shows the E_f values of different amino acids on a log scale.

2.5.5 Exclusion fraction trends identify methionine's unique role in protein synthesis: In order to investigate if the observed trends in E_f values reported above represent general trends in protein evolution or are just the result of the aggregation of subsets with vastly different underlying trends (for example, organisms with different evolutionary origins), we calculated the amino acid exclusion fraction values for proteins from archaeal, bacterial, eukaryotic, and viral origins independently. Regardless of their origin, all proteins maintain comparable E_f values for individual amino acids (e.g. Glutamine: bacteria, $E_{f,Q}$ = 0.010; archaea, $E_{f,Q}$ = 0.012; eukaryotes, $E_{f,Q}$ = 0.027; viruses, $E_{f,Q}$ = 0.022) (Figure 2.2). The only exception to this trend is the amino acid methionine which is one of the most invariant amino acids in archaeal, bacterial, and viral sequences with E_f values of ≤0.001 (archaea and bacteria) and 0.004 (viruses) (Appendix Table 2.2). However, in eukaryotes, methionine has a roughly 40 times higher exclusion fraction value ($E_{f,M}$ = 0.04) and is only the 16th most invariant amino acid. These $E_{f,M}$ values are consistent with the unique role that the amino acid methionine plays for translation initiation (72, 73), as the corresponding amino acid to the canonical start codon. Although translation of a bacterial (and archaeal) mRNA commonly requires a methionine as the start codon, other start codons are also utilized (74-76) which is reflected in the fact that methionine is not completely conserved in the AAAs of archaeal and bacterial proteins ($E_{f,M}$ = 1×10^{-4} in archaea and $E_{f,M}$ = 7×10^{-4} in bacteria). Given the sensitivity of the $E_{f,M}$ value to biases due to the translation initiation mechanisms, the reduced retention (reflected in a 40-fold higher exclusion fraction value) of methionine in the AAA of eukaryotic proteins suggests that non-standard initiation is even more common than previously suggested (77). On this background the observation that the exclusion fraction value of methionine in viral sequences in our data set, which contain only approximately 10% sequences of bacteriophages, is low suggesting that viral proteins primarily utilize methionine start codons for translation initiation by the host translation machinery.

2.5.6 Exclusion fraction values as a measure of AAA complexity is sensitive to the functional sub-group of a protein: Previous studies have suggested that the present-day AAA complexity has evolved to facilitate the diversity of functions performed by proteins (2, 16, 78). Therefore, we

wanted to investigate if the E_f values of amino acids in proteins with different functions and from different domains of life shared similar amino acid exclusion trends. To this end, we extracted sub-groups of proteins from our dataset based on specific function (Appendix Figure 2.3), such as proteins with enzymatic activity, ribosomal proteins, single domain antibodies (sdAbs), transmembrane proteins, flavoproteins, and nucleotide-binding (nt-binding) proteins to determine the amino acid exclusion frequencies within each sub-group. To compare these values, we then calculated the normalized exclusion fraction (E_f^N) of amino acids (normalized exclusion fraction is the exclusion fraction of each amino acid in the given protein sub-group divided by the exclusion fraction value of that amino acid in the overall dataset) (Figure 2.3). Results reveals that proteins with different functions roughly follow the general trend of amino acid exclusion observed for the overall Swiss-Prot derived dataset, with predominantly low E_f^N values for the core and prebiotic amino acids (Table 2.3). However, each sub-group has a specific set of amino acids with high E_f^N values such as arginine in transmembrane proteins ($E_f^{N,R} = 2.6$), consistent with their corresponding specialized function (Table 2.4). For the sub-group containing proteins with enzymatic activity, the E_f^N values show a tendency to preserve a set of amino acids with diverse physicochemical properties. Ribosomal proteins are the only sub-group that does not contain aromatic amino acid in their core set. Interestingly, their core set comprises methionine, lysine, and arginine which, although they are not prebiotic amino acids, have normalized exclusion fraction values ($E_f^N \lesssim 0.1$) so small that they are almost completely conserved across all ribosomal proteins ($E_f^{N,M} = 0.0027$, $E_f^{N,K} = 0.0044$, $E_f^{N,R} = 0.14$).

In summary, we identified that the amino acid exclusion trends are similar across different functional sub-groups of proteins. However, the E_f^N values of individual amino acids vary between protein functional classes. Investigation of the size and composition of the AAA of proteins from different sub-groups analyzed here shows that the preference for selecting and excluding amino acids is related to the functional diversity of the respective protein class.

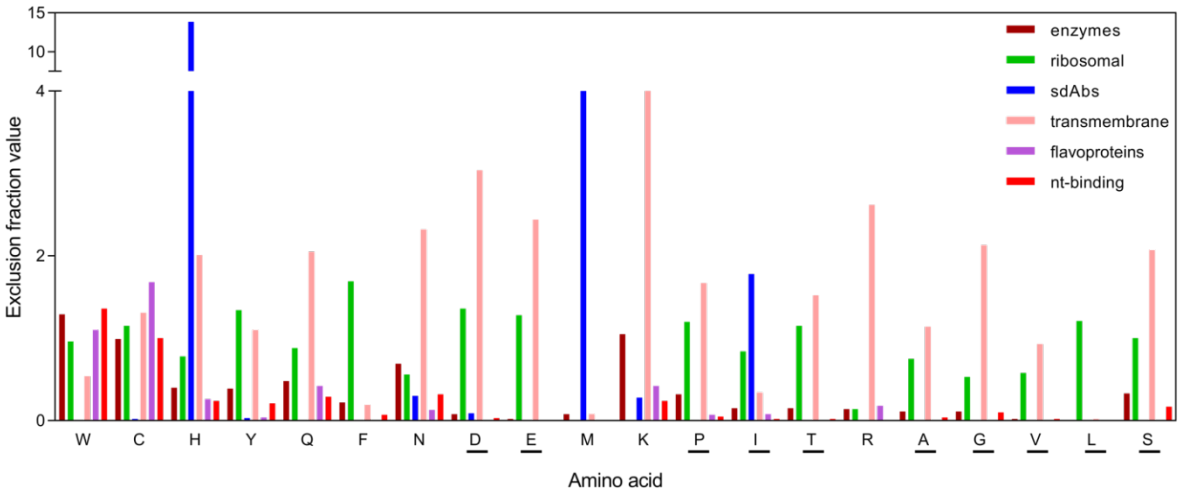


Figure 2.3 – Normalized exclusion fraction (E_f^N) values for the standard twenty amino acids in different protein sub-groups. Amino acids are arranged in the order of the E_f^N values observed for all proteins in the Swiss-Prot derived dataset (prebiotic amino acids are underlined).

Table 2.3 – Normalized exclusion fraction (E_f^N) values of the twenty proteinogenic amino acids in different protein sub-groups. The leftmost column is arranged in the decreasing order of exclusion fraction (E_f) values observed for amino acids in the Swiss-Prot derived dataset (prebiotic amino acids are shown in orange). An exact value of zero is denoted by 0.

Amino acid	Enzymatic proteins	Ribosomal proteins	sdAbs	Transmembrane proteins	Flavo-proteins	nt-binding proteins
W	1.3	0.96	0.0047	0.54	1.1	1.4
C	0.99	1.2	0.023	1.3	1.7	1.0
H	0.40	0.78	13.8	2.0	0.26	0.24
Y	0.39	1.34	0.034	1.1	0.044	0.21
Q	0.48	0.88	0	2.0	0.42	0.29
F	0.22	1.7	0	0.19	0	0.068
N	0.69	0.56	0.30	2.3	0.13	0.32
D	0.079	1.4	0.087	3.0	0	0.030
E	0.024	1.3	0	2.4	0	0.0026
M	0.077	0.0027	4.4	0.076	0	0

Amino acid	Enzymatic proteins	Ribosomal proteins	sdAbs	Transmembrane proteins	Flavo-proteins	nt-binding proteins
K	1.0	0.0044	0.28	4.0	0.42	0.24
P	0.32	1.2	0	1.7	0.072	0.053
I	0.15	0.84	1.8	0.34	0.078	0.017
T	0.15	1.2	0	1.5	0	0.020
R	0.14	0.14	0	2.6	0.18	0.013
A	0.11	0.75	0	1.1	0	0.042
G	0.11	0.53	0	2.1	0	0.10
V	0.025	0.58	0	0.93	0	0.017
L	0.0078	1.2	0	0.017	0	0
S	0.33	1.0	0	2.1	0	0.17

2.5.7 Amino acid alphabet distribution rules are universal for all species: To identify whether the evolutionary age affects amino acid distribution, particularly the prebiotic ones, we investigated the proteins from four prokaryotic (*Escherichia coli*, *Salmonella, Pseudomonas, Aeromonas*) and mammalian (*Homo sapiens, Pan, Rattus, Macaca)* species, two evolutionarily distant lineages in the tree of life (Figure 2.4, Appendix Table 2.3, 2.4). We observed that 8±3% of the proteins in the prokaryotic set are small alphabet proteins whereas 5±2% of mammalian proteins are SAPs. This finding indicates that the AAA complexities are comparable and have not changed over the course of evolution. The only difference lies in the E_f values of proline and isoleucine. In *E. coli* proteins, $E_{f,P}$ = 0.014, whereas in *H. sapiens* $E_{f,P}$ = 0.005. Therefore, according to the cutoffs defined in Table 2.2, proline is a part of the core set in *H. sapiens*, but not in *E. coli*. On the other hand, it is interesting that isoleucine is highly conserved in *E. coli* proteins with $E_{f,I}$ = 0.003. This is probably because isoleucine is oddly rich in the number of alternative routes (nine, to be specific) for its biosynthesis in *E. coli* (79). Furthermore, it is also evident that a core set of amino acids is dominated by the presence of prebiotic amino acids. Over the course of time and evolution, the AAA size has

expanded with the addition of amino acids with specific properties to increase the functional space accessible to the resulting protein and ultimately to life, also for evolutionarily old branches of the evolutionary tree. The latter suggests that expansion of the AAA predates the emergence of modern-day organisms independent of the age of the respective branch in the tree of life.

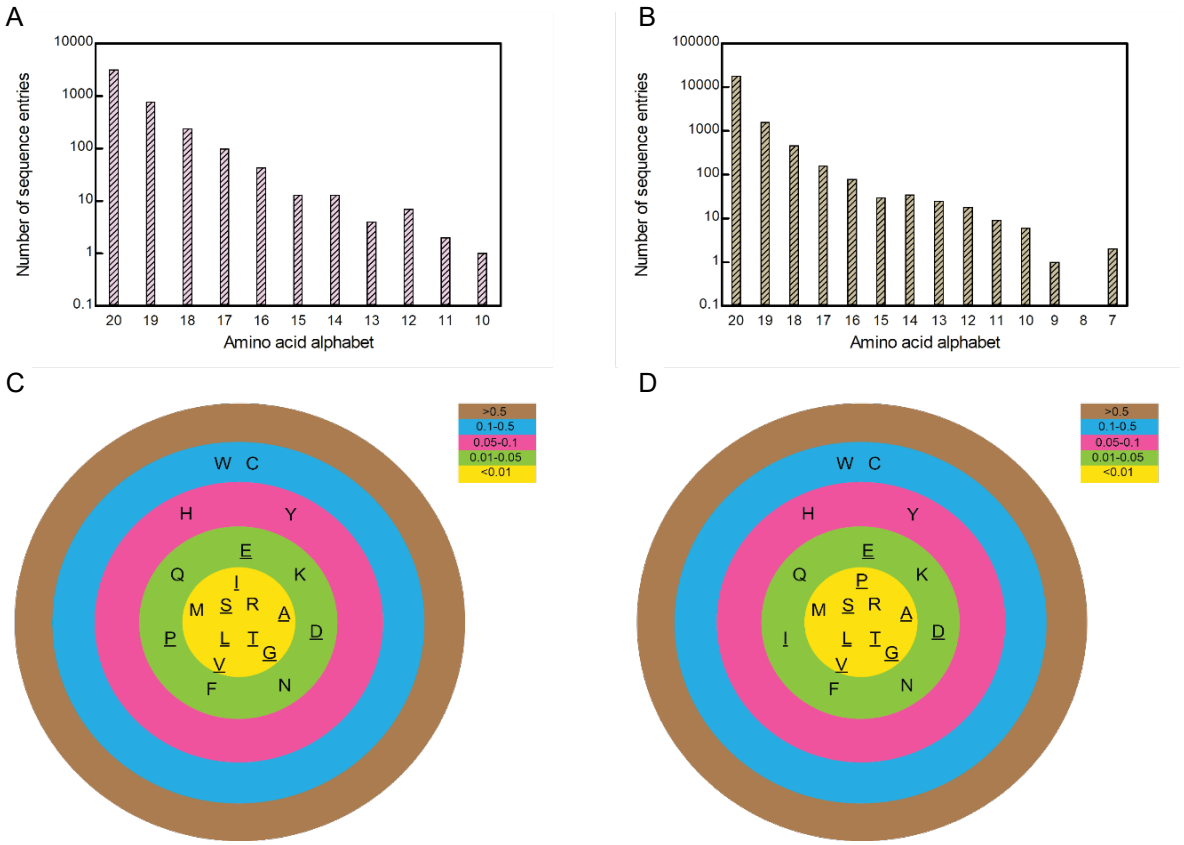


Figure 2.4 – Amino acid alphabet distribution for all proteins in *E. coli* and *H. sapiens*. A and B: The x axis shows the amino acid alphabet, and the y axis shows the population of proteins corresponding to the given amino acid alphabet in *E. coli* and *H. sapiens*, respectively. C and D: Bull's eye plots showing the exclusion fraction (E_f) values for different amino acids in *E. coli* and *H. sapiens*, respectively. The amino acids with highest E_f values are shown on the outermost circle and vice versa. The core amino acids (with E_f value ≤ 0.01) are shown on the innermost circle (prebiotic amino acids are underlined).

2.5.8 AAA size increases with protein length: In order to address the question regarding the driving force behind a gradual expansion of the amino acid alphabets (in addition to offering further physicochemical features and functional groups to facilitate additional roles, e.g., catalysis), we wanted to know if any correlation exists between the AAA complexity and the length of the protein utilizing a more complex AAA. To this end we calculated the protein length distributions for a given amino acid alphabet size (Figure 2.5). Interestingly, a majority of the proteins utilizing alphabet size smaller than 17 AAA rarely exceed protein sequence length of 100 amino acids. When the AAA size reaches 17, an increasing number of protein sequences are of length of more than 100 residues with almost all being above the length of 100 residues in proteins with 20 AAA (46% of 17 AAA proteins, 74% of 18 AAA proteins, 92% of 19 AAA proteins, and 99.5% of 20 AAA proteins). Notably, the distribution of protein sequence length is not smooth for a number of smaller AAA, such as the 10 AAA proteins which exhibit a trimodal distribution with the first, second and third modes respectively centered at lengths of 35, 53 and 65 residues, respectively. The latter might reflect preferences for domain sizes available to proteins based on these alphabets.

By analogy to the hypotheses that proposed a gradual expansion of the amino acid alphabet (16, 47), the observed correlation between AAA complexity and increased protein sequence length suggests that the present-day long polypeptide sequences of proteins evolved from shorter peptides. This in turn raises the mechanistic question of why larger AAAs (>16) are required for this. One possibility is the need for increased availability of amino acids with related physico-chemical properties. This would facilitate the synthesis of proteins from a pool of “monomers” where these monomers are less likely to be depleted. Such an availability would allow for longer sequences and avoid unwanted lag due to stalling or misincorporation. Another possibility is that the addition of more functional groups allowed interactions between the secondary structures or protein domains within a longer peptide that are conducive to sequential reactions, or for the formation of multidomain proteins.

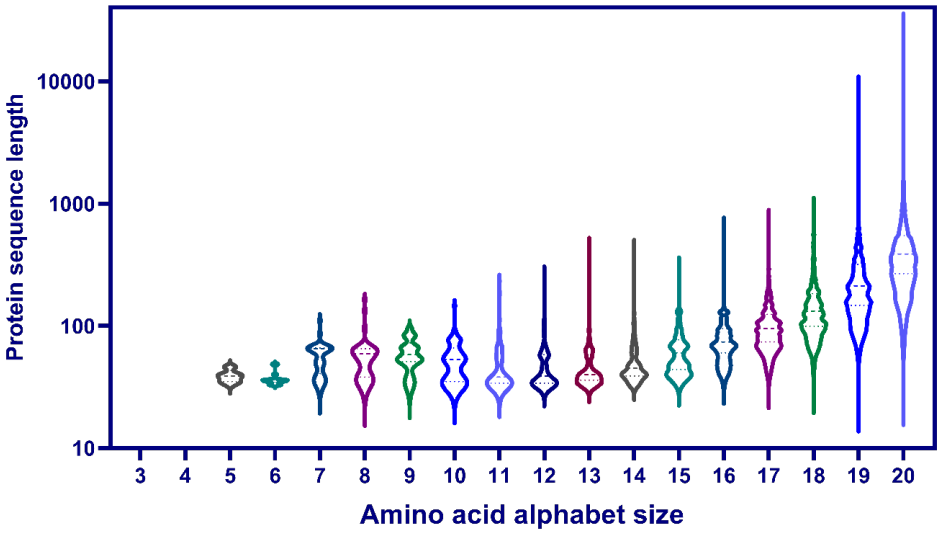


Figure 2.5 – Violin plot showing the relation between protein sequence length and amino acid alphabet size. The x-axis shows the AAA size of the proteins, and the y-axis shows the sequence length of proteins in log scale. For each AAA size, the shape of the violin plot shows the distribution of the entries corresponding to different sequence lengths. A wider shape indicates a cluster of entries with a similar sequence length, and the dotted lines shows the first, second and third modes from bottom to top respectively.

2.6 Discussion

Even after decades of protein evolution studies, it is still debatable whether the current AAA has been present since life first emerged on Earth or whether it is a product of evolution. If it is the latter, several theories such as the evolution of the genetic code, the alanine world hypothesis, etc. put forward scenarios hypothesizing how the evolution of the genetic code may have occurred (48-53, 57, 59). Particularly for the evolution of amino acids, it has been argued that the prebiotic amino acids made up the primary AAA set to appear on Earth, which gradually expanded to include more amino acids and evolved into the present day standard AAA (36, 47, 80). As much as these hypotheses and speculations seem correct, there has been little to no evidence to confirm these theories. In this work, we have addressed these questions by performing an in-depth bioinformatic study on roughly half a million protein sequences, a dataset derived from the UniProt/Swiss-Prot database. Analysis of the amino acid alphabets of protein sequences in our dataset puts on display the diversity of AAA explored by nature to design proteins, which ranges from AAAs as large as

twenty to as small as a three amino acids. The findings reported in this work support the gradual evolution of the AAA of proteins. Proteins with smaller AAAs tend to have smaller sequence length, which is in accordance with the fact that early proteins likely consisted of short length peptides. Moreover, as the length of the protein increases, the AAA size also increases, consistently with early proteins having been short sequence peptides which were composed of a smaller amino acid alphabet, presumably the prebiotic ones, as proposed by previous studies. Gradually over time, the amino acid alphabet pool of the proteins continued getting bigger with the incorporation of additional amino acids, perhaps to expand the structural and functional landscape of proteins, which also permitted an increase in the sequence length. On that note, it may be theorized that the alphabet expansion is perhaps occurring even in the present world where we witness the inclusion of the twenty-first and twenty-second amino acids, namely selenocysteine and pyrrolysine, in the amino acid alphabets of some present-day proteins (6, 16, 81-83), although further studies need to be performed to test such theories.

In contrast to the frozen code theory which states that the genetic code is universal, and any attempt to change it would be lethal, our results show that the AAA of proteins demonstrate a great extent of flexibility. Our findings reveal that small alphabet proteins are abundantly found in nature, thus proving that the present-day standard alphabet is not the only possible alphabet size and that smaller alphabet proteins exist in abundance. Additionally, the exclusion fraction values also show that some amino acids such as tryptophan and histidine have higher chances of being absent from the alphabet, whereas others such as the prebiotic ones are hightly conserved in the alphabet. Such a preferential trend of amino acid exclusion supports the theory of protein evolution facilitated by the gradual inclusion of amino acids into the alphabet. The discovery that prebiotic amino acids are dominantly conserved in proteins from all domains of life suggests that the prebiotic amino acids have always been and continue to be an integral part of all proteins regardless of the evolutionary age of the parent organism or species. It is notable that the prebiotic amino acids do not include any positively charged amino acid. Since one of the most important functions of the positively charged amino acids is to facilitate ribosome activity, it can be speculated that these amino acids were not vitally important in the prebiotic scenario because the ribosomes may not

have come into existence yet. Such a scenario contradicts the RNA-world hypothesis because if RNA came first, then the original peptide-bond catalysis machinery would have been a ribozyme, i.e., a primitive ribosome.

Investigation of the amino acid alphabets of proteins with diverse functions also revealed that the amino acid composition is, to an extent, reflected in the function of the protein. Findings such as conservation of positively charged amino acids in ribosomal proteins, non-polar amino acids in transmembrane proteins etc. establish that specific amino acids are preferred to promote a specific function. Additionally, a strong conservation of the prebiotic amino acids in proteins with diverse functions emphasizes that the prebiotic amino acids form the core of proteins, which along with other amino acids, allow the proteins to explore diverse functions. Lastly, to tie our findings into the current trend of designing engineered strains or species, we investigated amino acid distribution and their exclusion trends in species that lie on two extremes of the evolutionary scale, to identify if the design strategies differ for different species. Results such as similar trends of amino acid exclusion, conservation of prebiotic amino acids in the alphabet, and a similar set of core amino acids point towards the fact that even over a course of several million years, some amino acids are preferred and preserved by nature in all organisms ranging from the most primitive ones to the most evolved ones.

Understanding the evolution of the AAA might help us to design novel proteins as well as to reduce the amino acid complexity to facilitate protein engineering. Results also suggest that the alphabet design criteria followed by nature is fairly conserved across species and can be utilized for protein engineering and synthesizing small alphabet proteins for an organism in a near natural manner (Figure 2.6). Subsequently, such SAPs would help design reduced alphabet strains, capable of performing cellular functions using SAPs. This will not only reduce the energy and resource requirement of the organism but will also facilitate its growth and proliferation in an environment where the availability of resources is sparse, such as in space or on a remote planet. Taken together, our findings demonstrate that not all proteins require all twenty standard amino acids, and possibly the standard alphabet we observe today is required only to facilitate protein-function complexity and/or rapid and error-free synthesis.

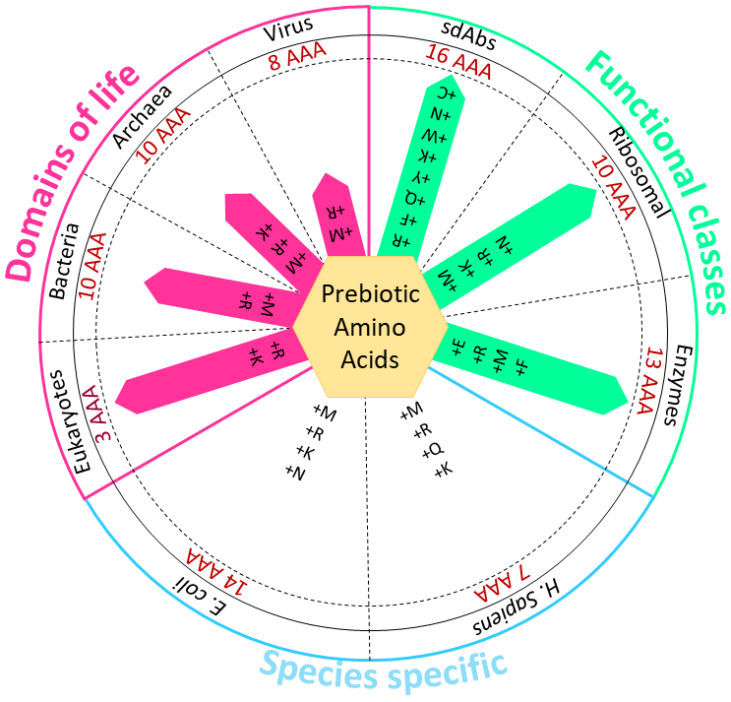


Figure 2.6 – Protein design principles utilized by nature. The three main categories of our analysis are shown in three distinct colors and include proteins from different domains of life (84), proteins with different functions (green), and proteins from specific species (85). The central hexagon represents the prebiotic amino acids (A, D, E, G, I, L, P, S, T, V), which make the core of the AAA of all proteins. The outermost circle in black shows the classes analyzed within a category and the second circle with brown text show the smallest AAA size possible for the given protein sub-group. The lengths of the pink arrows show the evolutionary age of the lineage (small = evolved early; long = evolved late). The radially arranged amino acids are the ones important in the AAA of the given sub-group apart from the prebiotic AAs, based on their exclusion fraction values.

CHAPTER 3: MOLECULAR DYNAMICS GUIDED INVESTIGATION OF NRP2 REVEALS LARGE-SCALE MOTIONS REQUIRED FOR HCMV PENTAMER PROTEIN BINDING

3.1 Preface

This chapter is an advanced draft of a format-neutral manuscript being prepared for publication. This chapter showcases our findings on how the structural dynamics of Neuropilin 2 is exploited by the Human Cytomegalovirus to gain entry into the host cells. The research was conceptualized by Hans-Joachim Wieden, me, and Dustin Smith. I carried out the homology modeling and structure validation of the Nrp2-Ca^{2+} and *apo* Nrp2 systems, performed MD simulations, subsequent trajectory, and principal component analysis, and wrote majority of the manuscript. Dustin Smith performed the backbone dihedral angle analysis, wrote associated sections of the manuscript, and contributed to the overall development of the manuscript. Hans-Joachim Wieden aided in the concept development, data analysis and writing.

3.2 Abstract

Neuropilins 1 and 2 (Nrp1 and Nrp2) are essential cellular receptors in vertebrates that interact with a variety of molecules to facilitate downstream signaling pathways. These receptors can be bound and exploited by viruses such as HCMV and SARS-CoV-2, enabling viral entry into a host cell and allowing viral propagation. The HCMV pentamer has been shown to form an interaction interface with the Ca^{2+}-containing loop of domain a2, as well as a loop region in domain b2 of Nrp2. However, in order for the HCMV pentamer-Nrp2 complex to form, a large-scale conformational rearrangement to displace domain a1 from the a2b1b2 core in Nrp2 is required. Here we employ molecular dynamics simulations to show for the first time an opening motion sampled by the a1 domain of Nrp2, which exposes the surface required for the HCMV pentamer-Nrp2 complex to form. Our findings demonstrate that domain a1 opening is a product of strongly coordinated motion of the Nrp2 core formed by the a2b1b2 domains that repel and trigger the a1 domain opening. We speculate that the HCMV pentamer can gain access to the open form of Nrp2 via spontaneous

displacement of a1, or by “pushing a button” in Nrp2 which facilitates the a1 domain displacement. The insights of Nrp2 structural dynamics reported in this work have broad implications for the development of antiviral therapies for Nrp2, which are also expandable to homologous Nrp1.

3.3 Introduction

The Neuropilin (Nrp) family of transmembrane proteins is comprised of essential multifunctional cell surface receptors in vertebrates, playing key roles in several signaling pathways (86, 87). Sharing 44% sequence similarity, Nrp1 and Nrp2 serve as receptors for a wide range of factors and ligands such as Vascular Endothelial Growth Factor (VEGF) receptors, semaphorin ligands, and other receptors to promote downstream signaling (88-91). Recent studies have reported that Nrps can be exploited by several viruses such as Human Cytomegalovirus (HCMV), Lujo virus (LUJV), and Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) to gain entry into the host cells (92-94). Additionally, the pleiotropic nature of *nrp* gene causes dysregulation leading to many pathological disorders such as cancer and cardiovascular diseases (95-98). The involvement of both Nrp1 and Nrp2 in such a wide range of physiological, pathological and signaling roles (88, 98) thereby sparks a great deal of scientific interest as understanding their structure, dynamics and function may potentially guide the development of therapeutic strategies.

The extracellular region of Nrp2 is composed of five domains: the a1 and a2 (a1a2) domains essential for semaphorin binding, b1 and b2 domains (b1b2) necessary for VEGF binding, and lastly the c domain required only for receptor dimerization (91). Previous findings have reported that the a1a2 domains belong to a family of CUB domains (99, 100), known to contain a conserved Ca^{2+} binding site. Though both a1 and a2 domains belong to the CUB domain family, the Ca^{2+} binding site has only been reported in the a2 domain, a conserved feature that is shared across distantly related species (91, 101-104). A recent study has shown that a loop region in the a2 domain coordinates a Ca^{2+} ion, which in turn interfaces with the HCMV pentamer protein (105). Structural studies on the Nrp2 fragment (a1a2b1b2) in complex with antibodies shows that Nrp2 undergoes an unexpected domain rearrangement in which the domains a2, b1, and b2 form a

tightly packed core and where the a1 domain is displaced away from the core (91). Additionally, the interface between the a1 domain and the a2b1b2 core is small, non-conserved and lacks strong interactions (91). In summary, the aforementioned studies indicate the a1 domain of Nrp2 is flexible and its displacement is required in order to form the HCMV pentamer-Nrp2 complex. It is therefore important to understand the conformational dynamics of this motion as it has implications in transmembrane signaling and mediating viral cell entry, as previously proposed (92).

To further understand the role of the underlying conformational dynamics in Nrp2 function with respect to the formation of the HCMV pentamer-Nrp2 complex, we employed molecular dynamics simulations and downstream computational analyses to study the structural flexibility of Nrp2. To produce a comprehensive suite reflecting the conformational flexibility of Nrp2, we examined the dynamic properties of the protein in the presence and absence of the Ca^{2+} ion in the Ca^{2+}-containing loop of the a2 domain. Examination of Nrp2 in the presence and absence of Ca^{2+} allowed observation of equilibrium-shifts in the dynamic properties at sites both proximal and distal to the Ca^{2+}-binding site upon Ca^{2+}-binding, including in the a1 domain. Revealed for the first time, we describe a large-scale opening and rearrangement of the a1 domain with respect to the a2b1b2 core along a hinge formed by the a1a2 loop, providing a mechanism by which the a1 domain may be displaced to form the HCMV pentamer-Nrp2 complex. Together, these results provide insight into the dynamic properties of Nrp2 and other homologous receptors, detailing how these dynamic motions are critical to their function and may be exploited for viral recognition.

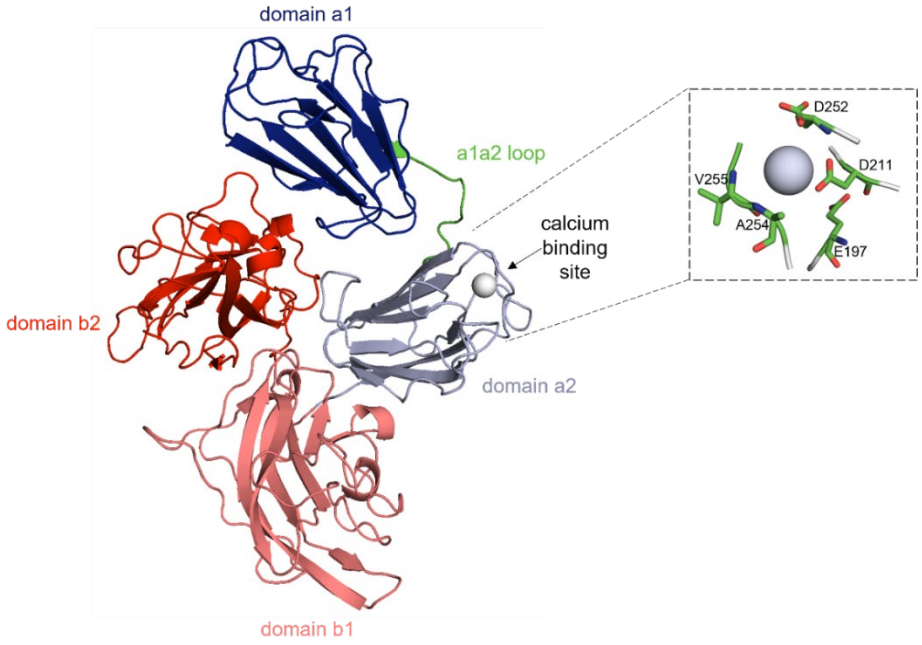
3.4 Methods

3.4.1 Homology modeling and structure validation: The amino acid sequence for *apo* Nrp2 was retrieved from the Universal Protein Resource Knowledgebase (106) from the UniProt entry number O60462. The Expasy SWISS-MODEL (107) was used to generate the homology model of a1a2b1b2 domains of *apo* Nrp2 using a crystal structure of Nrp2 a1a2b1b2 domains in complex with a semaphorin-blocking Fab (PDB ID: 2QQK) as the template structure obtained from RCSB Protein Data Bank (PDB) (108). The four-domain Nrp2 homology model was validated using Ramachandran plots (109). The homology model generated for the *apo* Nrp2 showed a high

percentage of residues in the allowed regions of the Ramachandran plot with the Global Mean Quality Estimation (GMQE) value of 0.87, reflecting a high-level of confidence in the generated model (110, 111).

3.4.2 Ca^{2+} ion placement in the Ca^{2+} binding site: According to previous reports, Ca^{2+} binds in the Ca^{2+}-binding loop of the a2 domain of Nrp2 (105) and is coordinated by the side chains of three acidic amino acids Glu197, Asp211, and Asp252 and two main chain carbonyl oxygens of Ala254 and Val255 (91). Using this information, our model structure of Nrp2-Ca^{2+} system was generated using the metal addition protocol in Pymol (112), by strategically placing the Ca^{2+} ion at the specified position in the a2 domain (Figure 3.1.A). The atomic parameters and topology description for the Ca^{2+} ion were prepared using the Ca^{2+} parameters protocol provided in the AMBER molecular dynamics package (113).

A



B

```
1   PCGGRLNSKDAGYITSPGYPQDYPSHQNCEWIVYAPEPNQKIVLNFNPHF  50
51  EIEKHDCKYDFIEIRDGDSESADLLGKHCGNIAPPTIISSGSMLYIKFTS 100
101 DYARQGAGFSLRYEIFKTGSEDCSKNFTSPNGTIESPGFPEKYPHNLDCT 150
151 FTILAKPKMEIILQFLIFDLEHDPLQVGEGDCKYDWLDIWDGIPHVGPLI 200
201 GKYCGTKTPSELRSSTGILSLTFHTDMAVAKDGFSARYYLVHQEPLENFQ 250
251 CNVPLGMESGRIANEQISASSTYSDGRWTPQQSRLHGDDNGWTPNLDSNK 300
301 EYLQVDLRFLTMLTAIATQGAISRETQNGYYVKSYKLEVSTNGEDWMVYR 350
351 HGKNHKVFQANNDATEVVLNKLHAPLLTRFVRIRPQTWHSGIALRLELFG 400
401 CRVTDAPCSNMLGMLSGLIADSQISASSTQEYLWSPSAARLVSSRSGWFP 450
451 RIPQAQPGEEWLQVDLGTPKTVKGVIIQGARGGDSITAVEARAFVRKFKV 500
501 SYSLNGKDWEYIQDPRTQQPKLFEGNMHYDTPDIRRFDPIPAQYVRVYPE 550
551 RWSPAGIGMRLEVLGCDW                                  568
```

Figure 3.1. – Overview of the structure of the Nrp2 extracellular domains a1a2b1b2. A. The domains have been sequentially colored from blue (domain a1) to red (domain b2). The loop region between domain a1 and domain a2 is referred as a1a2 loop and is colored in green. The Ca^{2+} ion is shown as a grey sphere in the Ca^{2+} binding site of domain a2. The positive charge of the Ca^{2+} ion interacts with the side chain of acidic amino acids and backbone carbonyl oxygens of aliphatic amino acids (inset). B. Sequence representation of the a1a2b1b2 domains of human Nrp2. Residue numbering is similar to numbering in crystal structures reported previously. Domains are represented using colored bars corresponding to the domain colors in panel A. Amino acids are color coded based on the secondary structure (blue: sheets, orange: helices, grey: intra-domain or inter-domain loops and turns). Amino acids crucial for Ca^{2+} binding and interaction with the HCMV pentamer protein are highlighted in cyan.

3.4.3 Molecular dynamics simulations: The Nrp2-Ca^{2+} and *apo* Nrp2 MD simulation systems were prepared using the tLEaP module of the AMBER 16 program package and the FF14SB force field (114). The total charges on the systems were neutralized using sodium (Na^{+}) ions and the proteins were solvated in an explicit TIP3P cubic water box with periodic boundary conditions, ensuring that the edge of the box extends at least 10 Å from the edge of the solute in each direction. A nonbonded cut-off of 10 Å was applied and the SHAKE algorithm was used to constrain hydrogen atoms. The solvent molecules and ions were initially minimized with 2500 steps of steepest descent followed by 2500 steps of conjugate gradient minimization, while the proteins were constrained using a 50 kcal mol^{-1} $Å^{-2}$ force constant. Both systems were then minimized using 2500 steps of unrestrained steepest descent, followed by 2500 steps of unrestrained conjugate gradient minimization. Subsequently, the systems were heated from 0 K to 300 K during a 100 ps solute restrained (2 kcal mol^{-1} $Å^{-2}$) heating phase, with a 2 fs time step. Thereafter, 10 ns unconstrained trial MD simulations were performed under NPT conditions using a 2 fs time step. Finally, using the conformation of the last frames of the trial MD runs as the starting point, 100 ns MD production simulations were performed on 20 replicates of the Nrp2-Ca^{2+} and *apo* Nrp2 systems using a time step of 2 fs and the pmemd module of the AMBER 16 program package (114). The cpptraj program was used to combine the resulting MD data to generate trajectories (115) and to calculate the root mean square deviation (RMSD) and root mean square fluctuation (RMSF) values of all replicates for both the Nrp2-Ca^{2+} and *apo* Nrp2 systems (data not shown). Due to the large size of the systems investigated (90,427 atoms including water), only three replicates for each Nrp2-Ca^{2+} and *apo* Nrp2 systems were selected for an extended simulation time of 1 µs, by using the RMSF trends for the 100 ns simulations as the selection criteria. All replicates of the Nrp2-Ca^{2+} system demonstrated comparable RMSF trends (Appendix Figure 3.1.B), therefore three replicates were selected at random to perform extended simulations. On the other hand, one of the *apo* Nrp2 replicates demonstrated significantly higher RMSF values in comparison to the other replicates, and therefore was selected, whereas all the other replicates demonstrated similar RMSF trends, so two of these replicates were selected at random for further extended simulations. Together, 20 shorter replicates and 3 longer extensions resulted in a total of 5 µs of simulation data collected for each of the Nrp2-

Ca^{2+} and *apo* Nrp2 systems. The angle, and axis of rotation describing the predominant opening motion of *apo* Nrp2 was quantified using the DynDom program (116, 117). The degree of "tilting" observed in domain a1 during the opening motion (i.e. rotation about an alternative axis) was examined via in-house developed algorithms adapted from prior publications (118). RMSF heat maps were constructed using Pymol (112). Visual inspection of the MD trajectories was performed using the molecular visualization software VMD (119).

3.4.4 Backbone dihedral angle analysis: Backbone amino acid dihedral angles for Ca^{2+}-bound and *apo* Nrp2 simulations were adapted from workflows described previously (120-122). Phi (Φ, between atoms C-N-Cα-C) and psi (Ψ, between N-Cα-C-N) angles were calculated using Tcl scripts in VMD (122) and transformed into a 180 x 180 matrix for each amino acid position (matrices each contained 32,400 bins, each bin 2° × 2°). To calculate the number of dihedral microstates populated for a given amino acid position, the number of bins with non-zero occupancy were counted using an in-house developed R script. Data reported in Figure 3.3.J is coloured according to percentage change in the number of microstates, where A_{Ca} indicated number of occupied bins in the Nrp2-Ca^{2+}, and A_{Apo} indicated the number of occupied bins in *apo* Nrp2 (Equation 1). Changes with magnitude less than 10% in the number of dihedral microstates for a given position were coloured in grey. As well, amino acid positions where changes in number of dihedral microstates were less than the magnitude of the standard deviation between replicates (i.e., a 0% change within error) were also coloured grey.

$$Percent\ difference\ dihedral\ microstates = \frac{(A_{Apo} - A_{Ca})}{A_{Ca}} \times\ 100\% \qquad \text{(Equation 1)}$$

Backbone dihedral populations in Nrp2-Ca^{2+} vs. *apo* Nrp2 were further compared by quantifying the amount of non-overlapping populations on their respective Ramachandran plots. The percentage of non-overlapping Φ/Ψ angle populations between the Nrp2-Ca^{2+} vs. *apo* Nrp2 was calculated using the sum of the absolute difference between the Ca^{2+}-bound [B_{Ca}] and the *apo* [B_{Apo}] matrices, resulting in a single value for each amino acid position. Values for each amino acid position were divided by the theoretical maximum value obtainable (1,000,000) and multiplied by

100% to determine the percentage of non-overlapping Φ/Ψ angles (Equation 2). Data reported in Figure 3.3.K illustrates colour-coded values for each individual amino acid, where grey values represented percentage non-overlapping Φ/Ψ angles <10% or a percentage non-overlapping Φ/Ψ angles whose standard deviation between three replicates was larger than the percent non-overlapping Φ/Ψ angles observed (i.e., 0% within error).

$$Percentage\ non-overlapping\ \Phi/\Psi\ angles = \frac{\sum|[B_{Ca}]-[B_{Apo}]|}{1{,}000{,}000} \times\ 100\% \qquad \text{(Equation 2)}$$

3.4.5 Principal Component Analysis: Principal component analysis (PCA) (123) was employed to detect any correlations (positive or negative) between the movement of the residues in the *apo* and Nrp2-Ca^{2+} systems. To this aim, we used the protein dynamics (ProDy) interface available in the normal mode wizard option under the Extension tab on VMD (119). On the ProDy interface window, we selected the PCA calculation option and provided the parameter topology files and the combined trajectory (netcdf) files of the two systems as the input. The PCA calculation was submitted thereafter for each replicate of both Nrp2-Ca^{2+} and *apo* Nrp2 systems. The generated output included the residue cross-correlation heatmaps for each replicate of the two systems.

3.5 Results

3.5.1 Molecular dynamics simulations of *apo* Nrp2 reveal large-scale hinge-bending motion displacing domain a1 from the a2b1b2 core: Previous studies have reported that Nrp2 can undergo a large-scale rearrangement where domain a1 moves away from the a2b1b2 core (91, 124), however, the mechanism by which this motion occurs remained unknown. We hypothesized that such a domain rearrangement would likely be rare to observe on the timescale of 1 μs MD simulations. We therefore performed MD simulations on the Nrp2-Ca^{2+} and *apo* Nrp2 to investigate whether the presence or absence of Ca^{2+} allows for the exploration of additional local or/and global motions on the conformational landscape of Nrp2. RMSD and RMSF analysis of both the Ca^{2+}-bound and *apo* Nrp2 simulations revealed that the systems adopted stable conformations

throughout, and that domain a1 displayed a high degree of inherent flexibility (Appendix Figures 3.1-3.4). For the *apo* Nrp2 system, one simulation replicate displayed an exceptionally large increase in RMSD over the course of the simulation in addition to high RMSF values in domain a1 (Appendix Figures 3.2A,B). Upon further inspection, we noted that the *apo* Nrp2 simulation sampled a large-scale opening motion (hinge-bend motion), whereby domain a1 moved away from the a2b1b2 core of Nrp2 (video data not shown). The opening motion described here occurred in several discrete "steps," whereby stable intermediates formed as the simulation trajectory moved from the "closed" to "open" state of Nrp2 (Appendix Figures 3.2A, C-F). The aforementioned intermediate states showed a stepwise breaking of interactions between domain a1 and the a2b1b2 core, followed by increased flexibility at the hinge region via disruption of intra-hinge hydrogen bonds (Appendix Figures 3.2C-F). Together, these results demonstrate that domain a1 of Nrp2 can be spontaneously displaced from the a2b1b2 core in a stepwise fashion via a hinge-bend motion.

To further quantify the extent of the a1 domain opening motion, the structural file for the "open" state of *apo* Nrp2 was superimposed on the starting structural model of *apo* Nrp2 (Figure 3.2). A pair of spatially close residues was selected from the a1 domain (Q53) and b2 domain (Y458) in the starting structure, and the distance between the two residues was measured to be 14 Å in the "closed" state and increased to 63 Å in the "open" state (Figure 3.2). To further quantify the angle of rotation upon a1 opening, the DynDom program (116, 117) was employed which showed that the a1 domain rotation takes place along a hinge formed by residues of the a1a2 loop. The a1 domain is therefore rotated along the hinge by an angle of 89° compared to its starting orientation (Figure 3.2). In addition to the domain rotation about the axis defined in Figure 3.2, domain a1 also "tilts" about an additional axis by approximately 15 degrees as Nrp2 transitions from its closed to open state (Appendix Figures 3.5C, D). These observations illustrate, in atomistic detail, the movements required to displace domain a1 from the a2b1b2 core, which is essential to the formation of the Nrp2-HCMV pentamer complex. However, in order to uncover further details of how such a domain rearrangement could be triggered in nature, we examined the molecular determinants required for HCMV binding in our Nrp2 simulations.

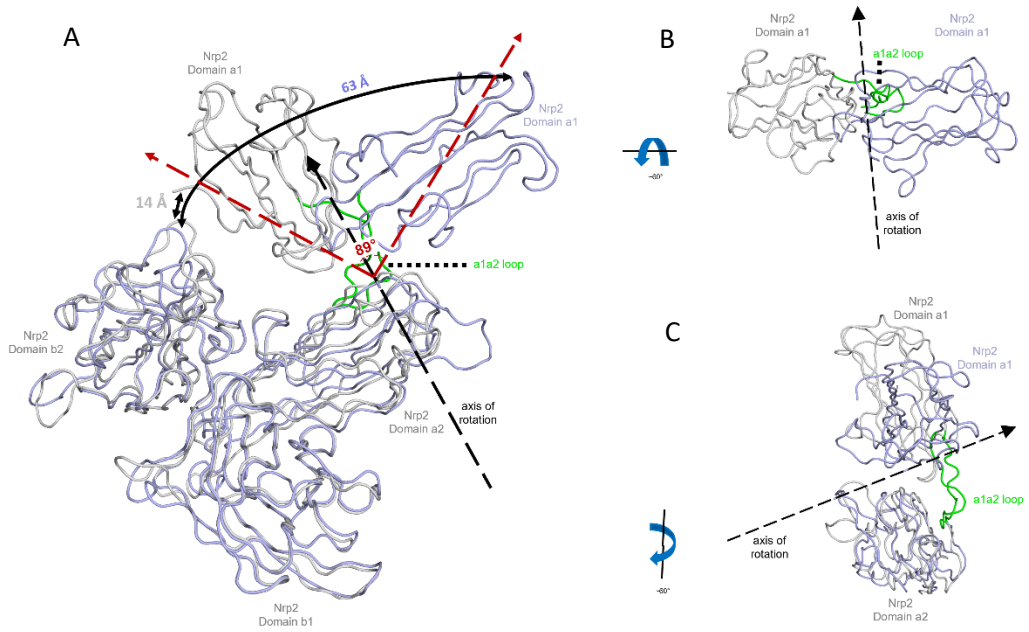


Figure 3.2 – Visualization of the Nrp2 hinge-bending motion whereby domain a1 "opens" away from the a1b1b2 core along the a1a2 loop. A. The open structure for *apo* Nrp2 Replicate 3 (purple) is obtained from the MD trajectory frame at 300 ns and is superimposed on the starting structure (grey). The distance between Q53 and Y458 of the a1 domain and b2 domain, respectively, is measured in the two structures. The black solid arrows show the distances between the two residues color coded for the close structure (14 Å) and the open structure (63 Å). For the a1 domain opening motion to occur, the a1a2 loop acts as the hinge and forms the axis of rotation as shown by the black dashed arrow. The brick-colored arrows show the angle between the a1 domains of the closed and open structures by taking the difference in the spatial position of Q53 around the hinge in the two structures. B. Top-down view of the hinge-bending opening motion of the a1 domain around the axis of rotation formed by the a1a2 loop, and C. Side view of the hinge-bending opening motion.

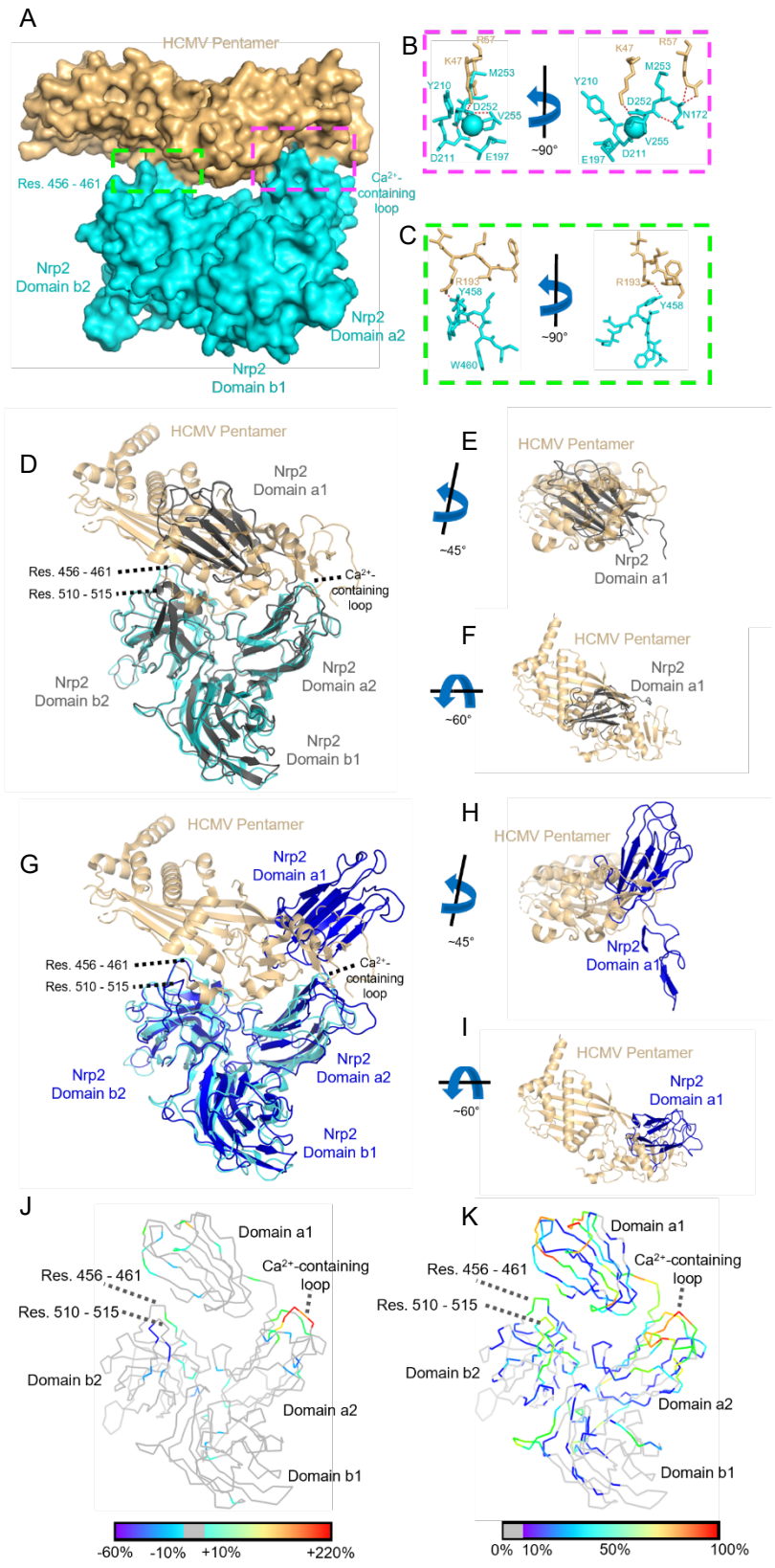
3.5.2 Ca^{2+} binding alters molecular determinants required for HCMV pentamer binding in Nrp2

After examining the flexibility of Nrp2 in both its Ca^{2+}-bound and *apo* states, we sought to further examine the molecular determinants required for HCMV pentamer binding to Nrp2 to gain insight into the dynamic properties of these regions and how they may contribute to complex formation with the HCMV pentamer. Ca^{2+} binds Nrp2 through a combination of side chain and backbone interactions with amino acids within a loop region of domain a2 (see Section 3.2.2, Figure 3.1.A). A recent study has shown that Ca^{2+} is required for Nrp2 binding by the HCMV pentamer, and that the Nrp2 Ca^{2+}-containing loop in domain a2 along with a loop in domain b2 (Residues 456-

461) form a sizable portion of the binding interface ((105), Figure 3.3.A-C). However, in order for the HCMV pentamer-Nrp2 complex to form, the a1 domain of Nrp2 must be displaced from the a2b1b2 core to avoid steric clashes (Figure 3.3.D-F). The displacement of domain a1 in our open-state Nrp2 MD simulations was therefore examined with respect to the orientation of the HCMV pentamer in the previously published Nrp2-HCMV pentamer complex (105). Upon overlaying the open state of Nrp2 with the Nrp2-HCMV pentamer complex, it was noted that the open state of Nrp2 clearly circumvented a number of the aforementioned steric clashes by displacing the a1 domain towards the N-terminus of HCMV protein UL128 (Figure 3.3.G-I). With this observation in mind, it was clear that the a1 domain could be displaced via the opening motion observed in our Nrp2 simulations in order to allow the formation of the HCMV pentamer-Nrp2 complex. However, several questions remained with respect to the mechanism by which the HCMV pentamer gains access to Nrp2 with domain a1 displaced. As a starting point, we asked how does the presence of Ca^{2+} impact the orientation of amino acids in Nrp2 recognized by the HCMV pentamer? Knowledge of these molecular determinants of Nrp2 in solution were therefore important in understanding how Nrp2 is recognized by the HCMV pentamer, and how the HCMV pentamer alters the conformational landscape of Nrp2 upon binding. To this end, we examined amino acid dihedral dynamics in Nrp2 to both understand the impact of Ca^{2+} on the dynamic properties of the system and further characterize dynamics changes that may serve as molecular determinants for Nrp2 recognition by the HCMV pentamer.

To examine backbone dihedral angle dynamics, Ramachandran plots were generated for each amino acid position in Nrp2 for the duration of the MD simulations. At several amino acid positions, marked changes in backbone amino acid dihedral dynamics were evident when comparing the Nrp2-Ca^{2+} to *apo* Nrp2 (Appendix Figure 3.6). To interpret these changes in the aforementioned Ramachandran plots, two main analysis metrics were adopted. First, the number of unique backbone dihedral angles sampled over the course of the MD simulations were examined (i.e., quantification of the number of dihedral microstates sampled by each amino acid position in Nrp2-Ca^{2+} vs. *apo* Nrp2). Dihedral microstates were defined as 2° × 2° bins on a Ramachandran plot, and bins that had non-zero occupancy were counted to determine the number of microstates

sampled (see Section 3.2.3). Several amino acids proximal and distal to the bound Ca^{2+} reported



[Figure on previous page]

Figure 3.3 – Molecular determinants of Nrp2 required for HCMV pentamer binding. A. Surface representation of Nrp2 (a2b1b2) in complex with the HCMV pentamer (PDB 7M22, (105)), HCMV pentamer is shown in light orange, and Nrp2 in cyan. B. Stick representation of hydrogen bonding network in the Ca^{2+}-containing loop of Nrp2 with the HCMV pentamer. C. Stick representation of Nrp2 Residues 456-461 in complex with the HCMV pentamer. D. Several regions of Nrp2 were unresolved in PDB 7M22, and to highlight these missing features our Nrp2 homology model in the "closed" state (grey) was aligned to the previously deposited structure. Steric clashes are observed between the HCMV pentamer and domain a1/ Res. 510-515 of Nrp2 upon alignment. E. Side view of steric clashes between the HCMV pentamer and domain a1 of Nrp2. F. Top-down view of steric clashes between the HCMV pentamer and domain a1 of Nrp2. G. Overlay of the Nrp2-HCMV pentamer complex (PDB 7M22, (105)), with the "open" state of Nrp2 identified via our MD simulations. H. Side view of the HCMV pentamer and displaced domain a1 of Nrp2. I. Top-down view of HCMV pentamer and displaced domain a1 of Nrp2. J. A colour-coded ribbon representation of percentage change in dihedral microstates upon removal of Ca^{2+} from the Nrp2 structure. Amino acids with an increase in dihedral microstates upon removal of Ca^{2+} (those that explored additional dihedral conformations) are shown in warm colours, whereas amino acids that explored fewer dihedral microstates upon removal of Ca^{2+} are shown in cool colours. Amino acids with a magnitude of change in dihedral microstates <10% are shown in grey (n = 3 simulations for Ca^{2+}-bound and *apo* simulations, averaged values of 3 replicates shown). K. Percentage non-overlapping dihedral angle populations in Ca^{2+}-bound vs. *apo* simulations is shown for each amino acid position. For example, a percentage of 100% indicates that there is no overlap between the Φ/Ψ populations observed in the Ca^{2+}-bound simulations and the Φ/Ψ populations observed in the *apo* simulations (see Appendix Figure 3.6 for example, n = 3 simulations for Ca^{2+}-bound and *apo* simulations, averaged values of 3 replicates shown, figure generated using Pymol).

drastic changes in the number of dihedral microstates sampled upon removal of the Ca^{2+} ion, including amino acids directly involved in forming interaction interfaces with the HCMV pentamer (Figure 3.3.J). We therefore conclude that Ca^{2+} binding shifts the dynamics of the Nrp2 system by causing some amino acids to exhibit either enhanced or reduced conformational flexibility with respect to the number of dihedral microstates sampled. Together, these results suggest that Ca^{2+} binding has global impacts on the conformational freedom which may play a vital role in modulating the dynamic properties of Nrp2 required for HCMV pentamer recognition.

To further characterize the differences in dihedral angle populations in the Ca^{2+}-bound vs. *apo* states of Nrp2, previous workflows developed by our group were amended to determine to what extent dihedral angle populations overlapped in the Nrp2-Ca^{2+} vs. *apo* Nrp2 simulations (120, 121). Several amino acid positions proximal and distal to the bound Ca^{2+} had different dihedral angle populations upon Ca^{2+} removal (Figure 3.3.K). The aforementioned changes in amino acid dihedral angle dynamics were most evident at inter-domain interfaces within Nrp2, as well as regions forming interaction surfaces with the HCMV pentamer. With this data in mind, the amino

acid dihedral angles of several key residues in our MD simulations were compared to the dihedral angles observed in the Nrp2-HCMV pentamer complex (PDB 7M22, (105)). With Ca^{2+} bound, several amino acid positions displayed dihedral populations shifted towards dihedral angles observed in the Nrp2-HCMV pentamer complex (Appendix Figure 3.6.A-D). However, some amino acid positions such as V255 and Y210 adopted dihedral conformations drastically different than dihedral angles observed in the HCMV pentamer-Nrp2 complex (Appendix Figure 3.6.E-H, Figure 3.4). Additionally, previous studies have shown that V255 is directly involved in forming an interaction interface between the HCMV pentamer and Nrp2 via a hydrogen-bonding network (Figure 3.4.A). However, in our Nrp2-Ca^{2+} simulations, V255 is not involved in hydrogen bonding interactions in the Ca^{2+}-containing loop region of Nrp2 (Figure 3.4.B, C). In the case of the HCMV pentamer-Nrp2 complex, V255 is involved in a hydrogen bonding interaction with D252, which in turn is stabilized by the hydrogen bonding interactions between K47 (of UL128 in the HCMV pentamer) and D252. We therefore propose that some amino acid positions in Nrp2 utilize an "induced fit" model for HCMV pentamer binding, whereby the HCMV pentamer must reorient specific binding site backbone dihedrals in order to form the Nrp2-HCMV pentamer complex. In summary, Ca^{2+} binding results in global alterations in the number of microstates explored by several amino acid positions, as well as alterations in the backbone dihedral population distribution at regions critical for the binding and recognition of Nrp2 by the HCMV pentamer. As well, we also note that the large-scale opening motion observed in our *apo* Nrp2 simulations represents a path that can be used to mitigate steric clashes during the formation of complexes involving Nrp2 and the HCMV pentamer.

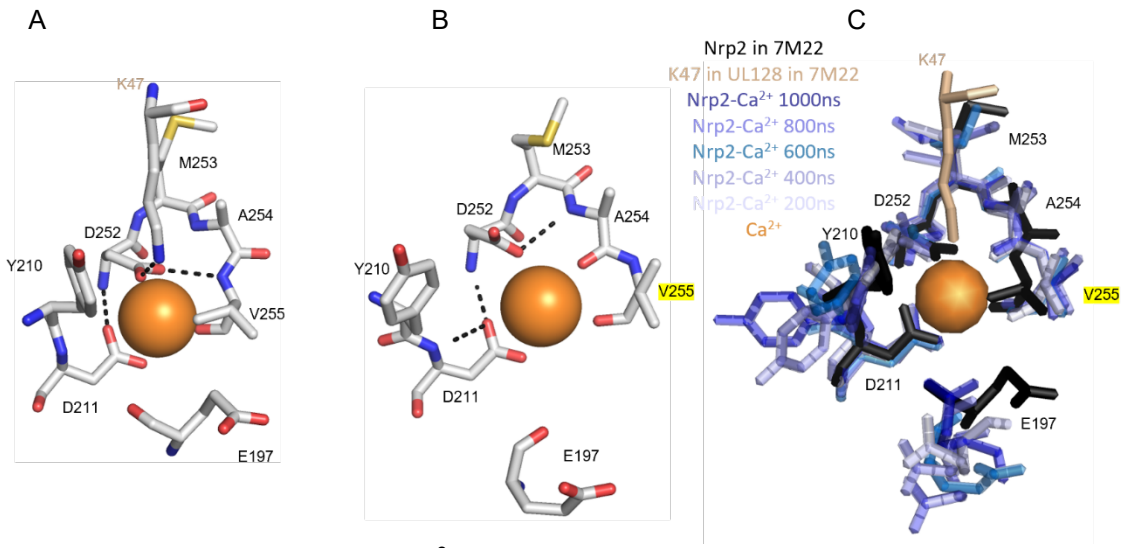


Figure 3.4 – Amino acids of the Ca^{2+}-containing loop in Nrp2 adopt altered conformations and hydrogen bonding networks when bound by the HCMV pentamer. A. Stick representation of the hydrogen bonding of the Ca^{2+}-containing loop in Nrp2 when bound by the HCMV pentamer (PDB 7M22). In the Nrp2-HCMV pentamer complex, K47 (of UL128 in the HCMV pentamer) stabilizes a hydrogen bonding network involving itself, and D252/V255 in Nrp2. This hydrogen bonding network displaces V255 into a unique conformation not observed in our simulations. B. A representative hydrogen bonding network of the Ca^{2+}-containing loop in Nrp2-Ca^{2+} from our MD simulations (snapshot at 600 ns) showing V255 is not involved in hydrogen bonding interactions in the Ca^{2+}-containing loop region of Nrp2. C. Overlay of snapshots at every 200 ns in our Nrp2-Ca^{2+} MD simulations (blue shades) aligned with the Nrp2-HCMV pentamer complex (Nrp2: black, HCMV pentamer: wheat). V255 in the Nrp2-HCMV pentamer complex displays a distinct and unique conformation when compared to our Nrp2-Ca^{2+} simulation data (Figure generated using Pymol).

3.5.3 Enhanced conformational flexibility of *apo* Nrp2 triggers observation of a1 domain opening:

After examining how Ca^{2+} impacts the local dynamics of the residues proximal and distal to the Ca^{2+} binding site, we further investigated how Ca^{2+} binding impacted the local dynamics of the residues beyond the Ca^{2+} binding site, including residues from all four domains. We proposed that absence of Ca^{2+} allows for enhanced conformational flexibility of the residues, consequently enabling the exploration of large-scale dynamics such as the a1 domain opening. To characterize the aforementioned dynamics and the correlated movements between specific residues and/or regions, we performed an empirical comparison between the residue cross-correlations in the Nrp2-Ca^{2+} and *apo* Nrp2 systems. Analysis of the residue cross-correlation heatmaps for the Nrp2-Ca^{2+} replicates show varying intensities of positive cross-correlations among residues within individual domains, evident by observing a high positive cross-correlations as “squares” along the diagonal from bottom left to top right (Figure 3.5.A-C). It is specifically notable that the a2 domain, where the

Ca^{2+} ion is located, shows the least positive cross-correlation tendency among all the four domains. This likely indicates that the intra-domain communication and dynamics of the a2 domain may be restricted in the presence of a bound Ca^{2+} ion. It is also interesting to note that the domains a1 and b1 show a negative cross-correlation of ≤ -0.6 (Figure 3.5), even though these two domains are situated farthest away from each other. This suggests that cross-correlation between amino acids is not dependent on their physical proximity, and perhaps propagates as a signal through the global structure of Nrp2, irrespective of the distance between the residues.

The *apo* Nrp2 replicates show positive cross-correlations among residues within individual domains, similar to the Nrp2-Ca^{2+} replicates. However, compared to the a2 domains of *apo* Nrp2 replicates that do not explore the domain a1 opening motion (Figure 3.5.D, E), the replicate in which the domain a1 opening motion is observed (Figure 3.5.F) shows a distinctly high positive cross-correlation of ≥0.7 within the a2 domain (yellow box). Furthermore, *apo* Nrp2 when exploring the described opening motion demonstrates cross-correlation values of ≥0.6 between the domains b1 and b2 (black box), indicating a high positive cross-correlation among the two domains in comparison to the other replicates. Overall, a high positive cross-correlation between the residues of domains a2b1b2 in *apo* Nrp2 upon opening may perhaps be an indication of the existence of a tighter a2b1b2 core as suggested previously (91), where this core functions together to trigger the a1 domain opening motion. Furthermore, *apo* Nrp2 simulations which do not sample the opening motion lack a high positive cross-correlation between the a2b1b2 domains, which is possibly the reason why the a1 domain opening motion is not observed in these two replicates. Apart from a strong positive cross-correlation, we also observe a strong negative cross-correlation between domains a1 and b2 in the opening *apo* Nrp2 simulation (Figure 3.5.F, green box). Negative cross-correlation may be indicative of a repulsive communication between the two domains, where the two domains move away from each other, further assisting the opening of the a1 domain.

To summarize, with the Ca^{2+} ion bound to the Ca^{2+} containing loop, both the local and global dynamics of the Nrp2-Ca^{2+} system are restricted, as evidenced from low cross-correlation values for the residues of the a2 domain and non-appearance of any positive or negative cross-correlation between residues of any other domains, respectively. However, in the *apo* Nrp2 system the a1

domain opening is achieved when the a2b1b2 domains function together as a core versus the a1 domain, evident from a2b1b2 high-positive cross-correlation. This motion is further assisted by repulsive interaction between the a1 and b2 domain, evident from the high negative cross-correlation between the a1 and b2 domains. Correlated dynamics between the a2b1b2 domains is therefore required in order to trigger the a1 domain opening motion reported in this study.

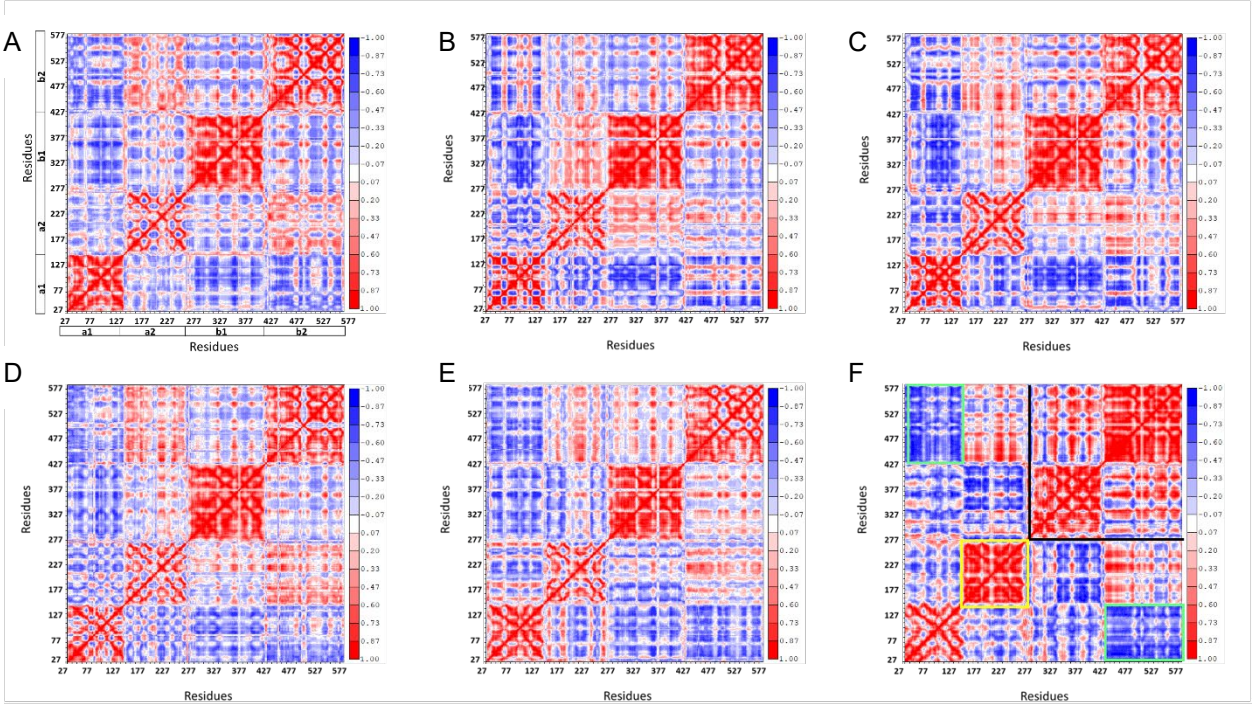


Figure 3.5 – Principal component analysis-based residue cross-correlation heatmaps. A-C. Residue cross-correlation heatmaps for the Nrp2-Ca^{2+} system replicates. Panel A also includes a 2D representation of the different domains along the two axes to help identify which domain the given residues belong to. Residues belonging to the four domains show different intensities of positive cross-correlations among themselves, evident from the squares along the diagonal from bottom left to top right. Of the four domains, the a2 domain, which houses the Ca^{2+} ion, shows the least positive cross-correlation. D-F. Residue cross-correlation heatmaps for *apo* Nrp2 system replicates. Panel F shows a high positive cross-correlation among the residues belonging to domain a2, shown in yellow box. Additionally, a high positive cross-correlation of ≥0.6 can be seen for most residues of domain b1 and domain b2, shown in black box in panel F. A high negative cross-correlation between the residues of domain a1 and domain b2 can be seen in the green box in panel F.

3.6 Discussion

Together, the results of our study suggest two possible mechanisms by which the dynamics of the conformational landscape of Nrp2 can be exploited by the HCMV pentamer for viral entry (Figure 3.6). First, the “Spontaneous opening” mechanism (Figure 3.6.A-D), whereby the domain a1 opening motion of Nrp2 spontaneously occurs *in vivo*, followed by HCMV pentamer binding and recognition of the “open” Nrp2 complex. We postulate that such a mechanism is feasible as we have observed in our molecular dynamics simulations spontaneous opening of domain a1 in Nrp2 exposing the majority of the surface required to prevent steric clashes in the HCMV pentamer-Nrp2 complex. In the spontaneous opening mechanism, the dynamic and flexible nature of the a1 domain coupled with coordinated repulsion of the a1 domain by the a2b1b2 core enables sampling of a large-scale opening motion. With the a1 domain partially displaced, the HCMV pentamer can recognize the partially formed interaction interface and stabilize the remaining amino acids into their preferred conformations (i.e., induced fit). A second possible mechanism for the formation of the HCMV pentamer-Nrp2 complex is the “pushed-button” mechanism (Figure 3.6), whereby the HCMV pentamer interacts with a “closed”-state Nrp2 and elicits a conformational change in Nrp2 to displace domain a1. A “pushed-button” mechanism would involve interaction of the HCMV pentamer (or a component of the pentamer) with a “button” on Nrp2 in order to alter its inherent dynamics and facilitate the domain a1 opening motion. In the closed state of Nrp2, several areas involved in forming an interaction interface with the HCMV pentamer are solvent exposed (Appendix Figure 3.7). It is therefore possible that the HCMV pentamer (or a component of the pentamer) could bind and elicit a global conformational change in Nrp2. We speculate that the Ca^{2+}-containing loop in Nrp2 is a likely candidate for the “button” region of Nrp2. Upon introduction of K47 from UL128 of the HCMV pentamer (“the finger”) the hydrogen bonding network of the Ca^{2+}-containing loop in Nrp2 is pushed into an altered conformation. Specifically, K47 enables the formation of a hydrogen bond network between K47, D252, and V255 which is not observed in the absence of the HCMV pentamer (Figure 3.4). We therefore propose that this altered hydrogen bonding network could propagate a signal in Nrp2, triggering the domain a1 opening motion.

However, further research is required to characterize signal transmission from the “button-pushing” step to the domain a1 opening motion.

The chance occurrence of the a1 domain opening motion in our simulations is an attribute of the stochastic nature of MD simulations and witnessing such an opening motion on a 1 μs timescale is a rare and exciting result (e.g., as discussed in (91, 105)). Since Nrp2 and Nrp1 share 44% sequence similarity (88), similar correlated dynamics likely can be utilized by other viruses that target Nrp1 to gain cell entry (125-130). As well, according to previous studies, the interface between the a1 domain and the a2b1b2 core of Nrp2 is non-conserved (91). Together with previous findings describing the properties of Nrp2 binding by the HCMV pentamer (105, 124), the mechanistic details of the domain a1 opening motion reported in this study represent foundational knowledge critical for downstream antiviral therapy design for the Neuropilins (e.g. antiviral therapies where the coordinated dynamics can be targeted and disrupted to prevent the exploitation of the a1 domain opening by the viral proteins).

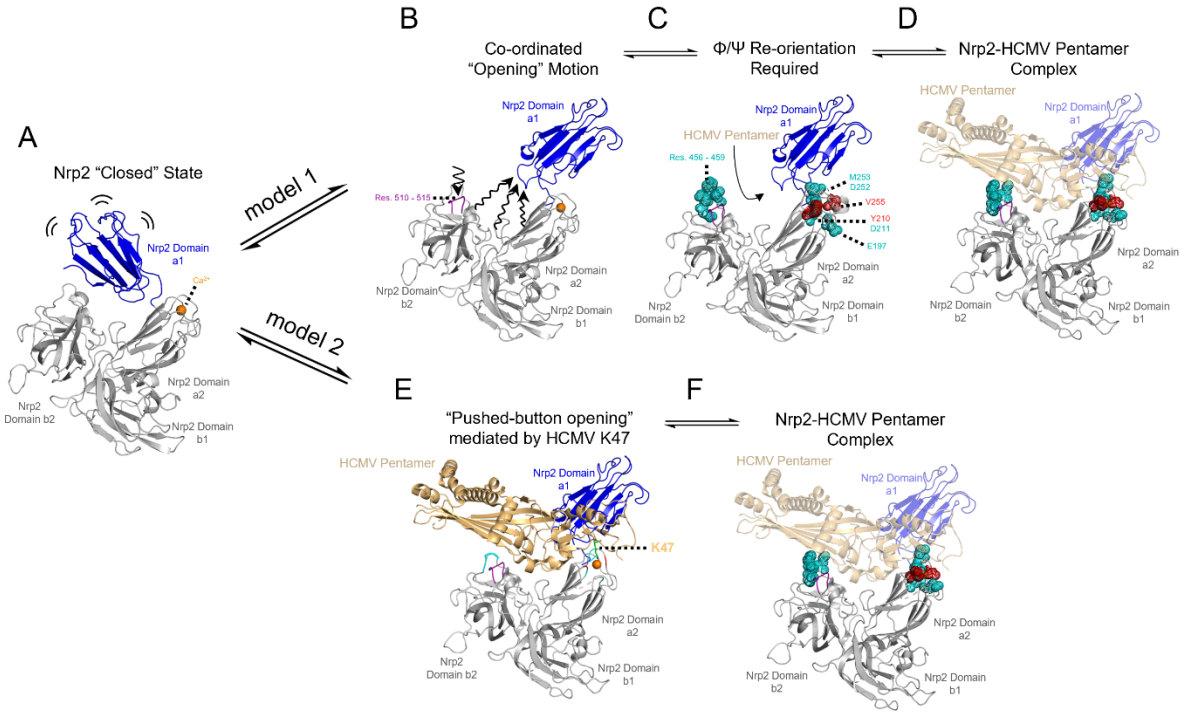


Figure 3.6 – Model of the Nrp2 conformational landscape demonstrating the "spontaneous opening" model versus a "pushed-button" model to allow binding of the HCMV pentamer. A. Cartoon representation of Nrp2 shown in a "closed" state, whereby domain a1 (85) displays inherent local flexibility. B. A co-ordinated motion by a2b1b2 repels and triggers displacement of a1. A flexible extension (Res. 510-515, purple) is also displaced outward allowing accessibility to regions that will interface with the HCMV pentamer. C. Prior to binding of the HCMV pentamer, several amino acid dihedrals involved in forming an interaction interface with the HCMV pentamer (PDB 7M22) are oriented in the conformation observed in the Nrp2-HCMV pentamer complex (cyan spheres), whereas others are oriented in different conformations (red spheres). Re-orientation of some amino acid dihedrals is therefore required (induced fit) to form the Nrp2-HCMV pentamer complex. D. Cartoon representation of Nrp2 domain a1 in the "open" conformation observed in our study (which must be further displaced) is overlaid onto a previously published Nrp2-HCMV pentamer complex (PDB 7M22). The HCMV pentamer is shown in light orange, and as shown would require further displacement of domain a1 in order to bind the Ca^{2+}-containing loop of Nrp2. E. The HCMV pentamer interacts with a "closed" state Nrp2 and elicits a conformational change in Nrp2 via a "pushed-button" mechanism mediated by the K47 residue to displace domain a1. F. Cartoon representation of Nrp2 domain a1 in the "open" conformation observed in our study overlaid onto a previously published Nrp2-HCMV pentamer complex (PDB 7M22).

CHAPTER 4: MOLECULAR DYNAMICS GUIDED RATIONAL REDUCTION OF AMINO ACID ALPHABET REVEALS UNDERLYING PROTEIN DESIGN PRINCIPLES

4.1 Preface

This chapter is an early draft of a format-neutral manuscript being prepared for publication. In this chapter, we report our rational reduced alphabet protein (RAP) design pipeline, an integration of molecular dynamics simulations coupled with downstream analysis techniques, employed to design RAP variants for proteins with distinct structures and functions using a two-pronged approach. The research was conceptualized by Hans-Joachim Wieden and me. I developed the methodology and software, performed script writing, data curation and processing, formal analysis, visualization, and figure preparation. The manuscript is written by Hans-Joachim Wieden and me.

4.2 Abstract

Proteins are made up of a common set of 20 amino acids known as the standard amino acid alphabet (AAA). Conventional protein design approaches for applications in biofuel production, pharmaceuticals and gene therapy generally utilize a design space limited to this standard AAA, thus restricting the incorporation of unnatural amino acids and the novel chemical functions that can be harnessed from them. Reducing the standard AAA is an approach that could free up the codon space for this and accelerate computational-based protein design. A functional protein written with a reduced alphabet will cost less to produce *in vivo* and *ex vivo*. Previously reported reduced alphabet protein (RAP) variants have shown little to no activity relative to their wild-type counterparts. We hypothesize that this is due to the over-reliance on substitution rules based on the physico-chemical properties of amino acids, neglecting the importance of protein dynamics on structure and function. The aim of our research is to develop a generalizable computational approach to design RAPs to facilitate efficient forward-engineering of proteins. We combine a set of *in silico* techniques to investigate and validate the dynamic and structural properties of the designed RAPs. In our two-pronged approach, we utilize both the distinct conservation scores of residues within a protein and the physico-chemical properties of amino acids for designing the

reduced alphabet variants (RA-variants) of three proteins with distinct structures and functions. In the long run, the developed novel computational framework will enable development of a generalizable computational approach to design RAPs, thus facilitating forward-engineering of proteins with applications in understanding and modifying disease-causing protein variants and widening the accessibility of individualized therapies and personalized medications.

4.3 Introduction

The primary structure of proteins consists of a linear sequence of amino acids, and the variety of amino acids in each protein is referred to as the amino acid alphabet (AAA) of the respective protein. The canonical AAA of most proteins usually encompasses twenty genetically encoded amino acids, referred to as the standard AAA. The amino acid composition of proteins differs in terms of the variety and population of individual amino acids included in them, where proteins may naturally comprise only a subset of the standard set. Such proteins that naturally consist of an alphabet smaller than the 20 AAA can be called small alphabet proteins (SAPs). SAPs are widely found in nature and have been proposed to be the only proteins present on early Earth, thus giving rise to theories that early life began with a 10 AAA which later expanded (131) to the current standard alphabet. The existence of SAPs in nature has inspired engineering and design of protein variants with smaller AAAs, and since that involves reducing the size of the AAA, the resultant proteins with a reduced AAA can be termed reduced alphabet proteins (RAPs). A previous study has shown that the prebiotic alphabet has near-optimal encoding of modern single domain folds and that designing RAPs using the prebiotic alphabet can create stable thermophilic structures (33) (47).

Aside from understanding natural and ancestral protein folding, rationally artificially reducing the alphabet size of proteins has several other advantages. In conventional protein design, the sequence space of a wild-type protein is probed and reorganised to generate multiple variants, which are then screened to identify highly optimal mutants that meet the criteria set by the designer. For every amino acid in the part of the protein that is being re-engineered, there are 20 possible substitutions so that the search space grows polynomially fast, potentially with an exponent as

large as 20 if we want to try all possibilities, the more amino acids we consider for substitution. While this leaves ample room for natural evolution to provide diverse protein structures and functions, the entire sequence space is not explored in nature, and is also practically impossible to explore using current laboratory processes (45, 132). Reducing the size of the AAA would reduce the sequence space to a subset of the total possible variants for a standard AAA. The 20 AAA is somewhat redundant in terms of physicochemical properties of the amino acids. For example, more than one amino acid has aliphatic side chains, similarly there are multiple amino acids with aromatic side chains, etc. We propose that this redundancy can be used as the starting point for reducing the amino acid alphabet, which will provide a basis for understanding general protein design rules for different classes of proteins. Another advantage of a reduced alphabet is that, once the non-essential amino acids are removed, their associated codons, which become unassigned, could then be used for introducing non-canonical amino acids into the target protein's AAA to introduce novel functional capabilities.

Previous research towards reduced alphabet protein design have focused on generating variants that retain wild-type protein folding after alphabet reduction (133, 134). For example, demonstration of a stable four α-helix bundle protein with a 5 AAA established that a RAP could possess a stable structure (135). Although maintaining structural integrity of a protein is necessary, retaining structural features in itself does not ensure protein functionality. Later studies focused on preserving both structure and function while designing RAPs. One such work was on the archaeal enzyme chorismate mutase, which showed that functional proteins can be designed from an alphabet size as small as 9 amino acids (35), however, the results were achieved by employing random mutagenesis strategies, an unpredictable and difficult to transfer strategy for biotechnology-relevant protein design. Additionally, although the resultant RAP enzyme folded into a structure similar to the wild-type protein, it retained only one-third of the wild-type protein activity. Together, all these studies indicate that preserving the structure is not enough and does not ensure functionality in a RAP. We hypothesize this can be attributed to protein dynamics, a fundamental property of the proteins. Proteins are dynamic entities with ever-changing shape and conformation, which allows them to interact with various binding partners and perform all sorts of functions.

Therefore, we predict that it is important to preserve both the structure and the dynamics of a protein to improve the chances of designing a functional RAP, unlike prior studies where researchers only examine protein folding predictions. We postulate that protein dynamics serves as the connecting link between the structure and function of protein, responsible for conveying a variety of signals and communications throughout the protein's surface, and hence, tweaking the structure of the protein would affect the dynamics and ultimately the function. This argument is suggested by the findings of the previously mentioned RAP design studies that demonstrated partial or complete loss of function in the resultant RAPs (34, 35, 136). The RAP design approach described in this work serves as the next level of protein molecular design where we examine the underlying dynamics of proteins with different structures and functions, as these dynamic motions are critical to their respective functions.

In this work, we have constructed and reported a systematic framework for designing RA-variants of proteins with different structures and functions. We employ molecular dynamics simulations coupled with computational analysis techniques to study protein dynamics, a fundamental property of proteins which is commonly overlooked in protein engineering studies. Our test system includes three model proteins with distinct structures and functions. The first model protein, chorismate mutase (CM), is an all α-helical enzyme from *Methanocaldococcus jannaschii*, responsible for catalyzing the chemical reaction for the conversion of chorismate to prephenate in the shikimate pathway (137). The second model protein is Initiation factor 1 (IF1), a β-pleated protein from *Escherichia coli*, an essential component of prokaryotic protein synthesis (138, 139). The third model protein included in our study is the 30S ribosomal protein S10 (rpS10), an α+β protein from *E. coli*, involved in the binding of tRNA to the ribosome (140). Our RAP design strategy involves performing amino acid substitutions based on the two most fundamental properties of amino acids, the physicochemical properties, and their conservation trends across related species. The dynamics of reduced alphabet chorismate mutase (RA-CM), reduced alphabet IF1 (RA-IF1) and reduced alphabet rpS10 (RA-rpS10) variants have been assessed using a set of computational analysis techniques. Depending on how the variants performed compared to the wild-type protein,

the variants were scored and ranked accordingly. The AAA sizes of the RA-variants designed in this work range from 17 AAA to as small as 8 AAA.

Our results show that small changes in the amino acid alphabet can have large implications on both the structure and dynamics of the proteins. Furthermore, a correlation between the secondary structure content and the amino acid substitution strategy is observed when designing RA-variants for a protein. In the case of the α-helical protein chorismate mutase, the conservation-based RA-CM variants perform better at preserving the wild-type dynamics. On the other hand, in the case of β-pleated IF1, physicochemistry-based RA-IF1 variants performed better at retaining wild-type dynamics. However, in the case of α+β rpS10, both conservation and physicochemistry-based RA-rpS10 variants showed comparable trends in preserving wild-type protein dynamics. Furthermore, structure prediction results of the 8 AAA variants using AlphaFold shows that even with such a small AAA size, the RA-variants designed using our approach preserved a secondary structure nearly identical to the wild-type protein. These findings demonstrate that our rational RAP design approach can generate reduced alphabet variants with wild-type-like dynamics and structure. Moreover, coupled with conservation of structure and dynamics, the RAPs generated using this approach have increased possibilities of preserving and displaying wild-type functions. Testing and extension of our RAP design pipeline on proteins from other functional and structural classes will help our understanding of the protein design rules in nature, thereby guiding the creation of a generalizable RAP design system. Ultimately our RAP design pipeline can be used as a basis for the forward-engineering of proteins, particularly in designing *de novo* proteins with novel or improved functions with applications in personalized medicine, e.g., for altering disease-causing proteins in individualized therapies.

4.4 Methods

4.4.1 Reduced alphabet design: substitution strategy: To generate RA-variants for the three proteins (CM, IF1, and rpS10), we first performed literature studies to identify amino acids critical for maintaining the structure or/and function of the proteins. The idea was to avoid substituting such important amino acids, at least in the initial substitution rounds, to prevent loss of structure or

function associated with these residues. Thereafter, the conservation pattern of the amino acids was identified using the Consurf server (described below) and the least conserved amino acids were selected first for substitution and *vice versa*. Our amino acid substitution approach utilized two different types of substitution strategies, first based on physicochemical properties of amino acids (generates chemistry-based variants) and the second based on conservation trends of amino acids (generates conservation-based variants). The chemistry-based substitutions were performed using a modified BLOSUM 62 (141, 142) matrix. This matrix was used to define amino acid groups based on their physicochemical properties including the charges and secondary structure propensities of the amino acids, thus allowing a reduction of the alphabet size while preserving amino acid diversity in the RA-variants generated (Appendix Figure 4.1). For the conservation-based substitutions, a global evolutionary conservation profile of the amino acids for the model proteins was constructed using the ConSurf server (143-145). ConSurf generated a multiple sequence alignment used to generate a sequence logo chart using the WebLogo server (146, 147) that displayed the conservation trends of amino acids. The least conserved amino acids were removed first and *vice versa*. The best substituent for an amino acid to be removed was determined by examining the sequence logo and selecting the second most prevalent amino acid at that position, thus guiding the conservation-based substitution (Appendix Figures 4.2-4.4). Overall, twenty two RA-variants were generated for each of the three model proteins, based on either physicochemistry or conservation rules. The generated RA-variants had alphabet sizes ranging from 17 AAA to as small as 8 AAA (Figure 4.1). Subsequently, homology modeling using SWISS-MODEL (148, 149) was performed to generate 3D-folded structures of the RA-variants, where PDB structures of the parent or wild-type (wt) proteins were used as the template. The quality of the generated 3D structures was assessed using Ramachandran plots (150), to visualize if the backbone dihedral angle combinations (Φ-Ψ) of amino acid residues appeared in the energetically allowed regions of the Ramachandran plot. (Appendix Figures 4.5-4.7, Appendix Table 4.1). In addition, the unfolded linear models of the RA-variants were also generated using the tleap program (151) available in the AMBER MD simulation package (114). Each step of AAA reduction utilized a two-pronged approach, where specific amino acid(s) were removed from the parent protein.

Therefore, the nomenclature of the RA-variants essentially displays the branch number and the substitution strategy involved to design the variant. For example, the 16A variant of chorismate mutase involves removing alanine and glutamine in the first scheme, and removing cysteine and phenylalanine in the second scheme, making the two prongs/branches (Figure 4.1). For each branch, both physicochemistry and conservation-based substitutions were performed to replace the removed amino acids. Therefore, in total, each AAA size resulted in four different variants named as 16_1_chem, 16_1_cons, 16_2_chem, and 16_2_cons, where the initial number denotes the AAA size, followed by the branch to which the RA-variant belongs, and lastly the substitution strategy used. A few RA-variants have also been designed using uni-pronged approach where only one set of physicochemistry and conservation-based substitutions were performed to generate two variants, such as IF1-17 AAA variants and 8 AAA variants for all three model proteins (Figure 4.1).

4.4.2 Molecular Dynamics simulations: Protocols mentioned in section 3.4.3 of Chapter 3 were used to prepare the respective molecular dynamics simulation systems for the RA-variants for the three proteins using the AMBER simulation package (114). After performing charge neutralization, energy minimization and equilibration of the systems using the protocols from section 3.4.3, 1 μs MD production simulations were performed on three replicates of each RA-variant using a time step of 2 fs and the Particle Mesh Ewald Molecular Dynamics (pmemd) module of the AMBER 16 package (114). The resulting MD data was combined to generate trajectories using the cpptraj program (115). The cpptraj program was also used to calculate the root mean square deviations (RMSD) and root mean square fluctuations (RMSF) for the wild-type proteins and their RA-variants. RMSD indicates the overall stability of the protein throughout the simulation time compared to the starting structure, whereas the RMSF identifies the residues or regions of high flexibility.

Figure 4.1 – Reduced alphabet design for the three model proteins. A. Chorismate mutase from *M. jannaschii*, B. IF1 from *E. coli*, and C. rpS10 from *E. coli*. Starting from the wild-type alphabets, each step of alphabet reduction involved removal of selected amino acids shown in pink ovals. The removed amino acids were substituted by other amino acids based on either physicochemical similarities or conservation trends. Reduced alphabet variants designed using physicochemical similarities were called chemistry-based variants, shown in blue boxes, and the variants designed using conservation trends in proteins were called conservation-based variants, shown in green boxes. The alphabet sizes of the variants are shown on the left.

4.4.3 *In silico* assessment of RA-variants

4.4.3.1 Cα covariance analysis and protein network construction: The propensity for two amino acids to move in the same direction is termed as the Cα covariance and is described by Equation 3.

$$c_{ij} = \frac{\langle (x_{i,t} - \bar{x}_i) \bullet (x_{j,t} - \bar{x}_j) \rangle}{\left(\langle (x_{i,t} - \bar{x}_i) \rangle^2 \langle (x_{j,t} - \bar{x}_j) \rangle^2 \right)^{1/2}} \quad \text{(Equation 3)}$$

Here c_{ij} is the normalized covariance between amino acids i and j, $x_{i,t}$ and $x_{j,t}$ are the Cartesian coordinates of the Cα of amino acids i and j at frame t. $\bar{x_i}$ and $\bar{x_j}$ are the time-averaged Cα Cartesian coordinates of amino acids i and j, and ⟨ ⟩ indicate the time-averaging of the quantities inside the brackets. The Cα covariance was calculated in the Carma software (152) using an in-house approach previously developed by our group (153, 154). The normalized covariance calculated from each frame of the MD trajectory was used to plot a histogram where a pair of amino acids were defined to have a large covariance if they deviated more than three standard deviations from the mean of the histogram. All protein networks were constructed using Gephi-0.9.1 (155), where each node represents an amino acid. Edges were drawn between amino acids that demonstrate a large covariance and maintain a distance of ≤4.5 Å between their Cα's for 75% of the simulation time, indicating that the residues are in close proximity for a majority of the simulation time. The size of each node is determined by its Betweenness Centrality (*Bx(n)*) value (156) which measures the number of times a node lies on the shortest path between other nodes. Nodes with a high betweenness centrality are interesting because they control the information flow of the respective communication paths. These nodes can represent important signalling pathways within a protein and can form targets for drug discovery. The *Bx(n)* value is calculated using Equation 4, where σ_{st} is the number of shortest paths from node s to t and $\sigma_{st}(n)$ is the number of shortest paths from node s to t that pass through node n.

$$Bx(n) = \sum_{\substack{s \neq n \\ t \neq \{n,s\}}} \frac{\sigma_{st}(n)}{\sigma_{st}} \quad \text{(Equation 4)}$$

4.4.3.2 Principal component analysis: Principal component analysis (PCA) (157) was employed to detect correlations (positive or negative) between the Cα carbons due to the residue motions in the

RA-variants and compare them to the wt protein. PCA was measured using the protein dynamics (ProDy) interface available in the normal mode wizard option under the Extension tab on VMD. The parameter topology file and the trajectory file were provided as the input. The generated output included separate residue cross-correlation heatmaps for wt proteins and RA-variants. A scree plot (158) was also generated which shows different motions sampled by the protein system and the amount of time the system spends sampling each motion.

4.4.3.3 Structure based modeling of protein folding: The protein folding pathways were studied using the SMOG2 (159) software package, installed and used with the GROMACS simulation suite. A symmetric matrix defining the interactions between residues, known as a contact map, was defined for the folded structures. Subsequently, linear polypeptide chains were allowed to fold where the contact maps of the folded structure served as a guide to the folding pathway. One hundred folding simulations were performed for each wt protein and the RA-variant. The generated output included averaged residue contact maps for the wt proteins and the RA-variants.

4.4.4 RA-variants scoring and ranking: After assessing the dynamic properties of the RA-variants, a scoring strategy was employed to create a hierarchy of the variants (Figure 4.2). All the variants were assigned a starting score of zero. For every type of computational analysis performed, a quantitative assessment was employed to compare the dynamic properties of the RA-variant triplicates to the wt, where if the variant's behaviour resembled the wt, the score of the variant increases by one, and vice versa (Appendix Tables 4.2 - 4.4). For example: RMSD values of ≤± 1Å qualified as similar, whereas for RMSF, the flexibilities of individual residues, particularly the structurally and/or functionally important residues, were compared to that of the wt to identify the variants with similar RMSF trends. For protein structure networks, the overall shape of the network and the betweenness centrality *(Bx(n))* values were used to identify variants similar to the wt. For PCA comparison, trends of strong positive and negative cross-correlation between residues and the different motions sampled by the variants were investigated to identify variants similar to the wt. Lastly, for SMOG folding, the folding pathway adopted by a polypeptide chain and the number of contacts formed between the residues was compared to identify variants similar to the

wt. At the end, the total scores of the variants at each substitution step were used for ranking. The top two scoring variants for each alphabet size were taken forward to the next step of alphabet reduction process (Appendix Tables 4.2 - 4.4).

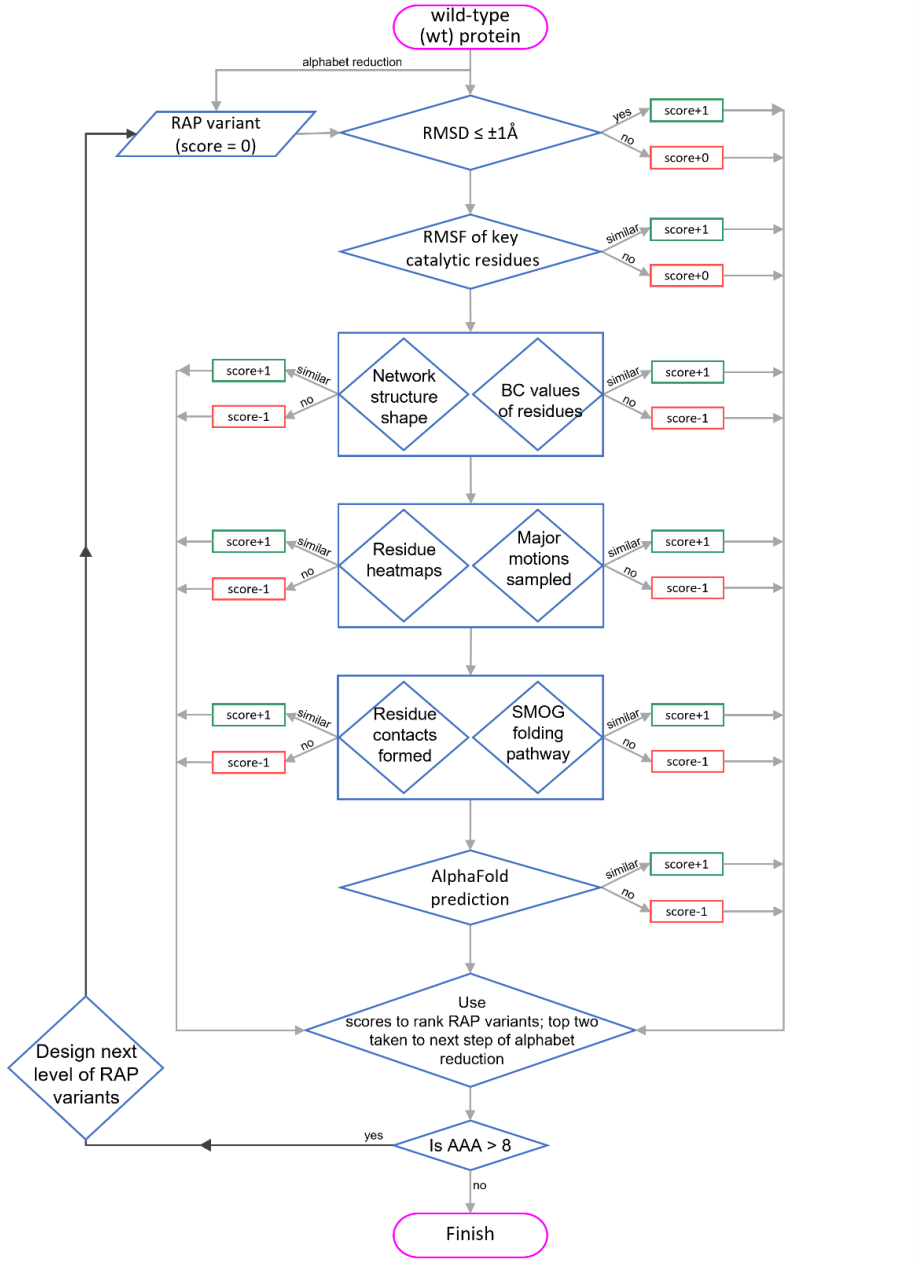


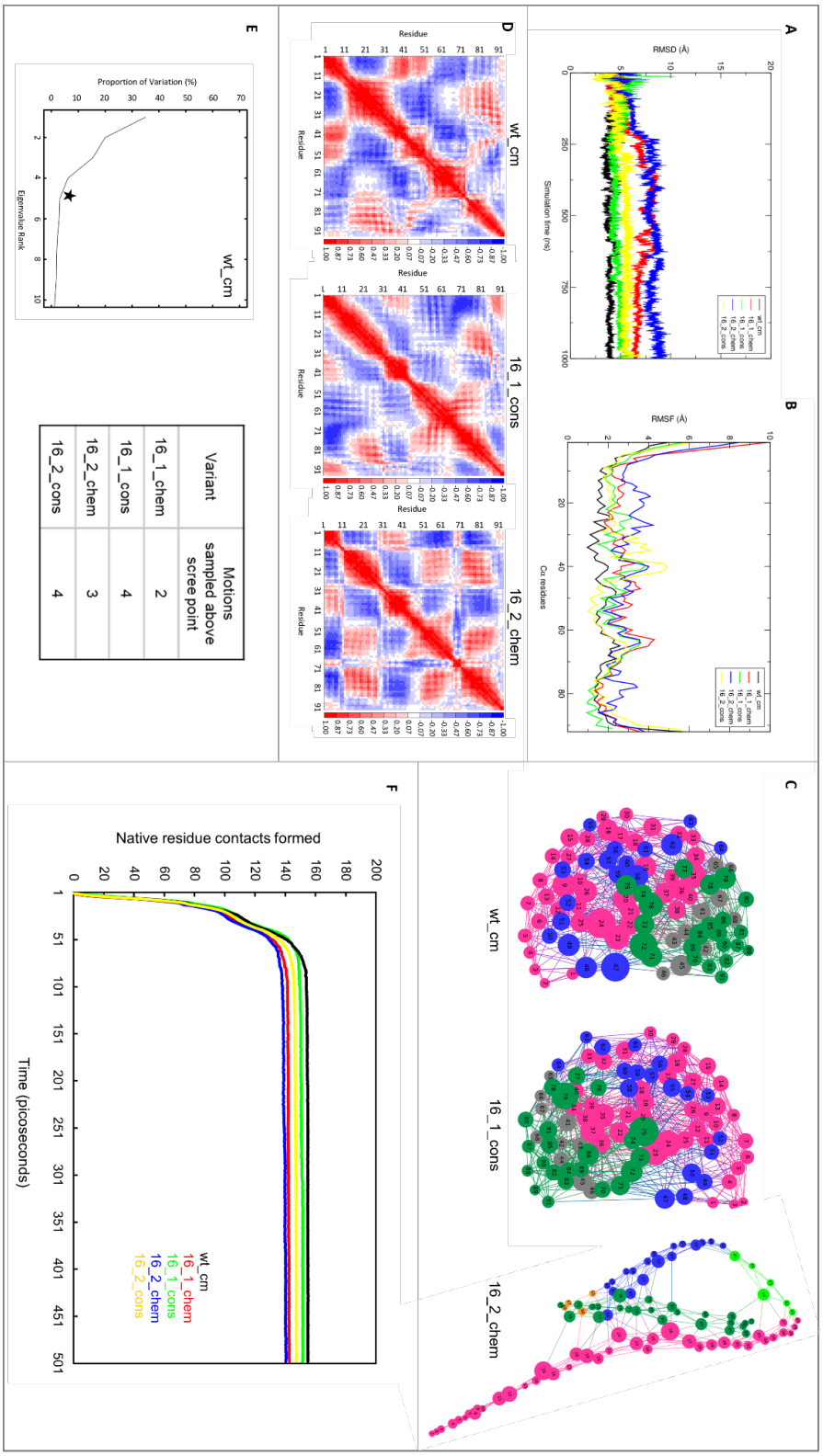
Figure 4.2 – Reduced alphabet variants scoring and ranking. Molecular dynamic properties of the reduced alphabet variants are compared to that of the wild-type (wt) protein using different computational analysis techniques, shown in blue diamonds (see Appendix Tables 4.2 - 4.4). A similarity with the wt protein gives a positive score, shown in green rectangles, whereas differences from wt-behaviour are penalized with deduction or no increment in scores, shown in red rectangles. The final scores of all variants for a given alphabet size are used to rank the variants. Top two scoring variants are carried forward for the next step of alphabet reduction.

4.5 Results

4.5.1 Conservation based variants outperform chemistry-based variants in α-helical chorismate mutase: Our first model protein chorismate mutase, an enzyme from *M. jannaschii*, is a primarily α-helical protein. 1 μs MD trajectories were examined to determine the impact of AAA reduction on the overall dynamics of RA-CM variants in comparison to the wild-type chorismate mutase (wt-cm). The RMSD and RMSF analysis of the RA-CM variants served as the first tier of selection, where we compared the overall flexibilities of the RA-CM variants to the wt-cm. The RMSD time courses of the 16A variants show that both the conservation-based variants have RMSD values similar to the wt-cm, whereas both chemistry-based variants demonstrated higher RMSD values compared to the wt-cm (Figure 4.3.A). A quantitative comparison of the RMSF values of residues, particularly the structurally and/or functionally important residues was performed to identify the effect of substitutions on flexibilities of residues (data not shown) in variants compared to the wt. We observed that the residue flexibilities of the conservation-based variants were similar to the wt-cm and much lower than the chemistry-based variants (Figure 4.3.B, Appendix Table 4.2).

Amino acid substitutions affect protein network dynamics: Long- and short-range interactions between individual amino acids are an important aspect of protein dynamics, responsible for transferring information across the protein network. Since proteins are highly dynamic entities, introducing too many unsuitable amino acids substitutions can lead to disruption of the overall protein network, causing the shape of the network to change. Similarly, an amino acid with a high betweenness centrality (*Bx(n)*) value in a communication path of a protein controls the flow of information within the protein (160), implying that replacing such an amino acid with an unsuitable substituent may reduce its *Bx(n)* value and thereby the flow of information in the protein network. Therefore, we investigated the protein networks of the wt-cm and the RA-CM variants for the preservation of edges and nodes in the networks, and the *Bx(n)* values of the key residues. Comparison of the wt-cm network structure along with the 16A network structures demonstrated that the shape of the networks for three of the four 16A variants was similar to the wt-cm (Figure 4.3.C, Appendix Table 4.2). However, the shape of the network structure of the 16_2_chem variant was completely different compared to the wt-cm network, demonstrating that unsuitable amino acid

substitutions can impact the overall shape of the protein network (Figure 4.3.C). To further understand the impact of substitutions on individual amino acid dynamics, we selected the amino



Variant	Motions sampled above scree point
16_1_chem	2
16_1_cons	4
16_2_chem	3
16_2_cons	4

[Figure on previous page]

Figure 4.3 – *In silico* assessment of RA-CM variants. A-B. RMSD and RMSF results for wt-cm and 16A variants. Both conservation-based variants show trends similar to the wt-cm, whereas the chemistry-based variants show higher RMSDs and RMSFs. C. Representative results for network structures of wt-cm followed by results of selected variants which demonstrate a noticeable similarity or difference with respect to the wt-cm, in this case, 16A_1_cons, and 16_2_cons variants, respectively. Residues are shown as circular nodes and labeled according to the residue numbering. Residues are coloured based on the domain they belong to, and the size of the node represents the *Bx(n)* value of the residue. D. PCA derived residue cross-correlation heatmaps of wt_cm, followed by results of selected variants which demonstrate a noticeable similarity or difference with respect to the wt-cm, in this case, 16_1_cons, and 16_2_chem variants, respectively. High positive cross-correlations are shown with higher intensity of red and high negative cross-correlations are shown with higher intensity of blue, whereas white shows no correlation. E. PCA derived scree plot for wt-cm. The x-axis shows the top ten motions sampled, and the y-axis shows the time (as a percentage) that the system spends sampling each motion. The star symbol shows the scree point. The sampling times for 16A variants are shown in the table. F. SMOG derived folding pathway analysis of wt-cm and 16A variants. The x-axis shows the time in picoseconds and the y-axis shows the number of native residue contacts formed. The folding pathway adopted by all 16A variants is similar to the wt-cm.

acids with the top ten *Bx(n)* values in the wt-cm network and compared to the 16A variants' network (Appendix Table 4.5). Our results show that six of ten residues with the highest *Bx(n)* values in the wt-cm were also among the top ten residues for both conservation-based variants. The 16_1_chem variant preserved only four of the top ten wt-cm *Bx(n)* residues, whereas the 16_2_chem variant preserved only three of the high *Bx(n)* residues from the wt-cm. Our findings therefore suggest that amino acid substitutions can affect the overall shape of the protein network, and assessment of network shape can be used to categorize and rank the dynamic properties of the RA-variants designed using different substitution strategies. Additionally, it was observed that both conservation-based 16A variants outperform the chemistry-based 16A variants in preserving the high *Bx(n)* values of the highly central amino acids in the wt-cm, thereby establishing that *Bx(n)* values of the amino acids proves to be an important technique to rank the RA-variants. Subsequently, we extended the use of network structure analysis to rank other RA-variants with alphabet sizes ranging from 14A to 8A, the results for which have been compiled in Appendix Table 4.2.

Conservation based variants preserve wt-cm dynamics: Several studies have established the role of motions such as side chain rotations in active sites, backbone motions during protein folding, major domain motions, etc. as governing factors of protein function (161, 162). Therefore,

we analyzed the effect of amino acid substitutions on the motions sampled by the wt-cm versus the RA-CM variants using PCA generated residue cross-correlation heatmaps and scree plots. Upon comparing the residue cross-correlation heatmaps, we observed that the overall residue cross-correlation trends for the 16_1_chem, 16_1_cons and 16_2_cons variants were similar to the residue cross-correlation trends of the wt-cm. However, the 16_2_chem variant shows increased positive cross-correlations between several residues compared to the wt-cm heatmap, perhaps indicating that the amino acids in the 16_2_chem variant experience increased interactions due to higher flexibility, as witnessed by high RMSD and RMSF values of the 16_2_chem variant. A key step while studying protein motions is to quantify the different motions sampled by the target proteins. To that end, we employed the PCA derived scree plots which provides a 2D representation of the major motions sampled by a system and the relative sampling times for each motion (158, 163). A scree plot identifies the dominant motion(s), identified by their eigenvalue ranks and their respective sampling times, and differentiates them from the remaining sampling noise shown by a “kink” in the plot, referred to as the scree point. Therefore, a scree plot indicates how many major motions are sampled by any system for a given time, indicating the tendency of the system to explore the conformational landscape, a measure of system’s flexibility. In the case of wt-cm, we observed that four of the top ten motions sampled by the wt-cm were above the scree point. Quantitatively, the four motions make up 75 percent of the sampling time which means that the wt-cm spends 75 percent of the simulation time sampling four different motions (Figure 4.3.E). However, sampling times of different motions for the 16A variants shows a different trend. The chemistry-based variants sample two and three motions above the scree point, for 16_1_chem and 16_2_chem variant respectively, appearing to be more rigid compared to the wt-cm. However, the conservation-based variants sample four motions above the scree point, suggesting that the conservation-based variants have a flexibility similar to the wt-cm, where they sample four different motions for the majority of the simulation time. Subsequent analysis of the other RA-CM variants showed similar results where most of the conservation-based variants showed similar flexibility to the wt-cm whereas most of the chemistry-based variants appeared more rigid (Appendix Table 4.2).

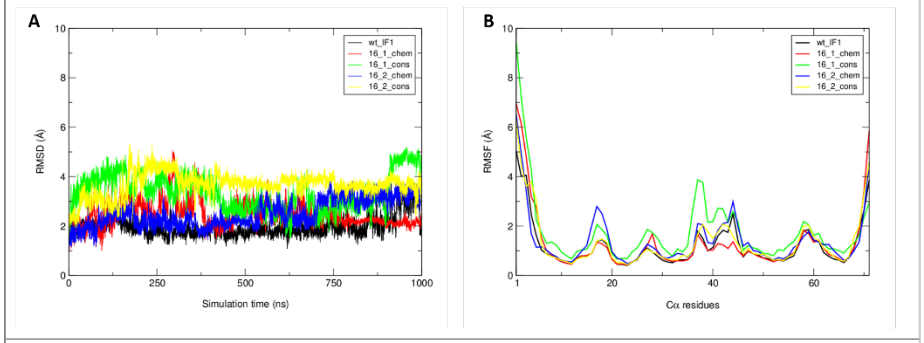
Conservation based variants strongly preserve residue contacts: Protein folding in nature may involve formation of more than one transition state during the course of folding. Computational analysis of folding of a linear polypeptide chain into the 3D folded structure describes the folding pathway it adopts and the number of native contacts it forms during the folding process. Adopting a folding pathway similar to the parent protein and preserving the native contacts between residues increases the variant's chances to be functionally active. We therefore wanted to examine the folding pathway with respect to the number of native contacts formed by each variant compared to wt-cm. To do so we used the SMOG folding analysis which identifies the folding pathway adopted by a polypeptide chain to fold into the target folded structure and the number of contacts formed between the residues. In case of the wt-cm, the polypeptide chain folds completely in the first 100 ps and forms 155 residue contacts (Figure 4.3.F). Although all the 16A variants adopt a folding pathway similar to the wt-cm and fold completely within 100 ps, the total number of residue contacts formed differs among the variants. With 151 and 148 contacts formed by 16_1_cons and 16_2_cons variants respectively, the conservation-based variants form more residue contacts compared to the chemistry-based contacts, 142 and 140 contacts in 16_1_chem and 16_2_chem variants, respectively. With a higher number of native contacts preserved compared to the chemistry-based variants, it can be suggested that the conservation-based variants are more likely to preserve the properties of wt-cm. Consequently, we extended the folding analysis to the other RA-CM variants and identified that the conservation-based variants preserve more contacts than the chemistry-based variants (Appendix Table 4.2).

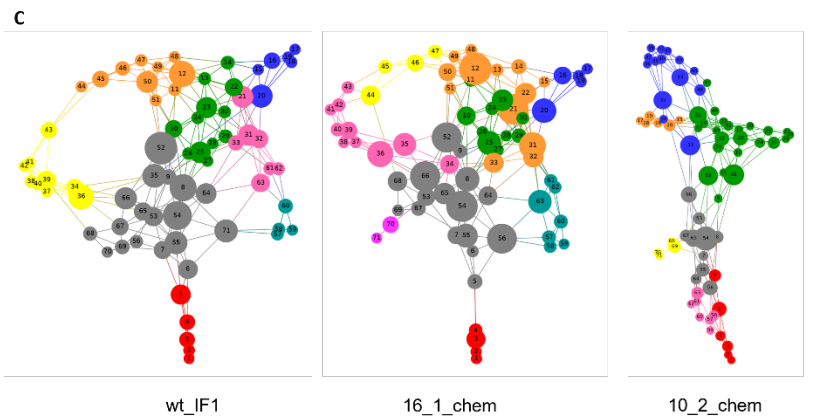
To summarize, analysis of the dynamic properties of the RA-CM variants using diverse computational analysis techniques revealed that changes in amino acid alphabet of a protein affect its dynamics. Of the two strategies employed for amino acid substitutions, chemistry and conservation, the conservation-based RA-CM variants outperform the chemistry-based variants and score higher on the scoring matrix at every step of alphabet reduction: eight out of the ten times, conservation-based variants were taken forward to the next level of alphabet reduction (Appendix Table 4.2).

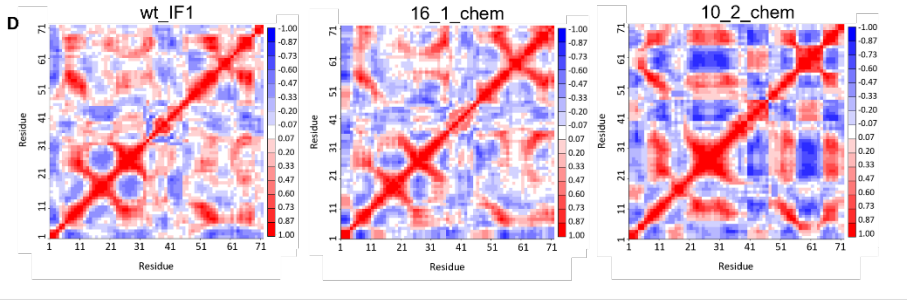
4.5.2 Chemistry based variants outperform conservation-based variants in β-pleated IF1:

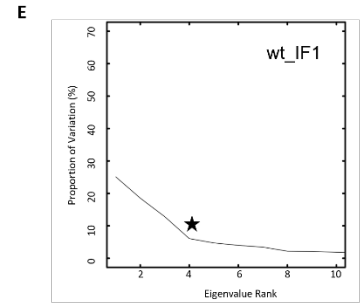
IF1, a protein essential in *E. coli* protein synthesis, constituted our second model protein, due to its small size and a predominantly β-pleated secondary structure. To understand the impact of amino acid substitutions on the overall dynamics, we studied 1μs MD trajectories of the wild-type IF1 (wt-IF1) and the reduced alphabet IF1 (RA-IF1) variants. The RMSD graphs of the 16A variants reveal that both chemistry-based variants have RMSD values similar to the wt-cm (Figure 4.4.A). The RMSDs of both conservation-based variants however were significantly higher than the wt-IF1. Upon further investigation of the RMSF values, we observed a similar trend (Figure 4.4.B), where the overall RMSF of the 16_1_chem variant was similar to the wt-IF1, and the RMSF of the 16_2_chem was lower than the wt-IF1. The RMSFs of both conservation-based variants were higher than the wt-IF1 and the chemistry-based variants. Comparable results were obtained for most of the RA-IF1 variants with alphabet sizes range from 14A to as small as 8A (Appendix Table 4.3). Based on these findings, we speculated that the chemistry-based variants perhaps preserve wt-IF1 dynamics better than the conservation-based variants and therefore we further examined the dynamic properties of the RA-IF1 variants by investigating their network structures.

To obtain a detailed understanding of the differences between the RA-IF1 variants and the wt-IF1, we analyzed and compared the network structure shapes and the *Bx(n)* values of the residues of the 16A IF1 variants to that of the wt-IF1. Interestingly the shapes of all four 16A variants networks were similar to the shape of the wt-IF1 network (Figure 4.4.C). However, the *Bx(n)* values of the residues segregates the variants that are similar to the wt-IF1 from the others. Of the top ten residues with highest *Bx(n)* values in the wt-IF1, seven and six of these residues appear in the top ten *Bx(n)* scorers of the 16_1_chem variant and the 16_2_chem variant, respectively. Both conservation-based variants preserve only four of the top ten central residues of the wt-IF1 (Appendix Table 4.6). The results for other RA-IF1 variants have been compiled into a table (Appendix Table 4.3). To summarize, most RA-IF1 variants preserved the overall shape of the network. However, the differences in how a signal travels through the network could be observed by changes in the nodes/edges of the networks, particularly, the identities of residues with

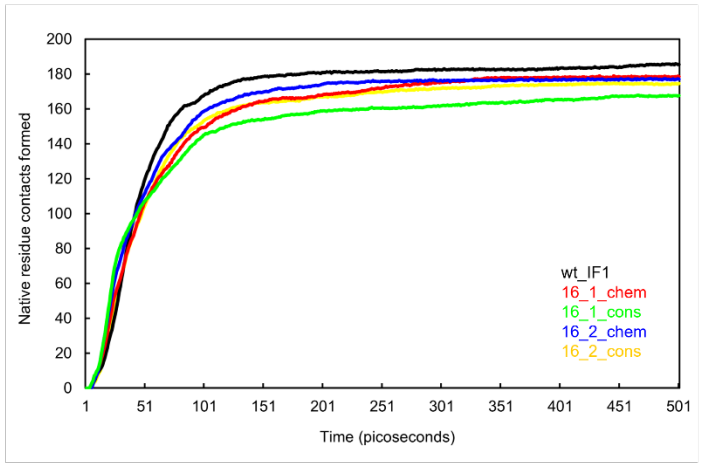








Variant	Motions sampled above scree point
16_1_chem	4
16_1_cons	3
16_2_chem	4
16_2_cons	3



[Figure on previous page]

Figure 4.4 – *In silico* assessment of RA-IF1 variants. A-B. RMSD and RMSF results for wt-IF1 and 16A variants. Both chemistry-based variants show trends similar to the wt-IF1, whereas the conservation-based variants show higher RMSDs and RMSFs. C. Representative results for network structure of wt-IF1 followed by results of selected variants which demonstrate a noticeable similarity or difference with respect to wt-IF1, in this case, 16A_1_chem, and 10_2_chem variants, respectively. Residues are shown as circular nodes and labeled according to the residue numbering. Residues are coloured based on the domain they belong to, and the size of the node represents the *Bx(n)* value of the residue. D. PCA derived residue cross-correlation heatmaps of wt_IF1, followed by results of selected variants which demonstrate a noticeable similarity or difference with respect to wt-IF1, in this case, 16_1_chem, and 10_2_chem variants, respectively. High positive cross-correlations are shown with higher intensity of red and high negative cross-correlations are shown with higher intensity of blue, whereas white shows no correlation. E. PCA derived scree plot for wt-IF1. The x-axis shows the top ten motions sampled, and the y-axis shows the time (as a percentage) that the system spends sampling each motion. The star symbol shows the scree point. The sampling times for 16A variants are shown in the table. F. SMOG derived folding pathway analysis of wt-IF1 and 16A variants. The x-axis shows the time in picoseconds and the y-axis shows the number of native residue contacts formed. The folding pathway adopted by the 16_1_cons variant is different from that of the wt-IF1, whereas the other variants adopt a folding pathway similar to wt-IF1.

high *Bx(n)* values, thereby establishing the importance of comparison of the *Bx(n)* values of amino acids to rank the variants.

The backbone motions and amino acid side chain motions are an integral part of protein dynamics, and therefore it was critical to investigate the motions sampled by the RA-IF1 variants in comparison to the wt-IF1. A comparison of residue cross-correlation heatmaps showed that apart from subtle differences, the residue cross-correlation trends in all 16A IF1 variants were similar to those for wt-IF1 (Figure 4.4.D). We further examined the various motions sampled by wt-IF1 and compared them to the motions sampled by the 16A IF1 variants (Figure 4.4.E) using scree plots. The wt-IF1 samples four motions above the scree point, where the system spends 60 percent of simulation time sampling these four motions. Among the 16A variants, both chemistry-based variants demonstrate nearly identical behaviour to the wt-IF1 and to each other, where both systems sample four motions above the scree point, spending 63 percent of simulation time sampling the four motions. The conservation-based variants however, sample only three motions above the scree point and spend roughly 60 percent of the time sampling only the three motions. Together these findings suggest that the chemistry-based 16A IF1 variants preserve motions and sampling times similar to wt-IF1, thus demonstrating that chemistry-based variants perform better at preserving wt-IF1-like dynamics, compared to the conservation-based variants which

demonstrate a loss of dynamics (Appendix Table 4.3). Together the results for the 16 variants and other RA-IF1 variants show that the chemistry-based variants perform better at preserving wt-like dynamics in IF1 (Appendix Table 4.3).

Lastly, to study the *in silico* folding pathway adopted by the 16A IF1 variants and the native residue contacts preserved in the resulting folded states compared to the wt-IF1, we employed SMOG folding analysis coupled with the study of residue contacts. Our results show that both chemistry-based 16A variants adopt folding pathways similar to wt-IF1 and preserve almost all native residue contacts (Figure 4.4.F). Compared to the 185 residue contacts formed in the wt-IF1, the 16_1_chem variant preserves 177 contacts and the 16_2_chem variant preserves 176 contacts. Conversely, the 16_1_cons variant not only adopts a different folding pathway compared to wt-IF1 but also preserves the least number of native residue contacts (167) among all the 16A variants. Perhaps employing a folding pathway different from wt-IF1 restricts the 16_1_cons variant from folding completely, as evidenced by the lesser number of contacts formed. The 16_2_cons variant scores the second lowest in preserving 174 residue contacts, although it takes the same folding pathway as adopted by wt-IF1. Results for the folding pathways and contacts formed during folding of the other RA-IF1 variants are summarized in Appendix Table 4.3.

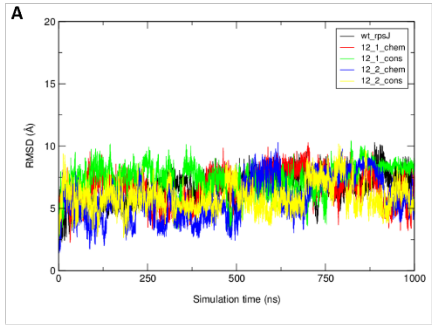
In conclusion, among all the RA-IF1 variants, the chemistry-based variants outperformed the conservation-based variants and scored higher on the scoring matrix at every step of alphabet reduction. In total, eight out of the nine times, chemistry-based variants performed better and were taken forward to the next level of alphabet reduction for IF1 (Appendix Table 4.3).

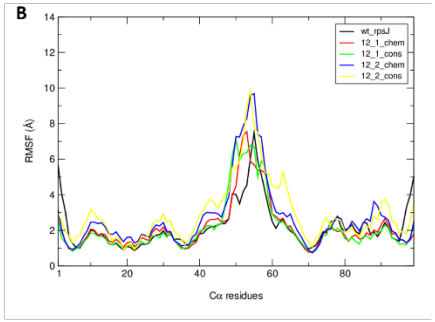
4.5.3 Chemistry- and conservation-based variants perform equally in α+β rpS10: After investigating the effect of amino acid substitution strategies on purely α-helical (chorismate mutase) or β-pleated (IF1) protein systems, we were interested in comparing these to a protein with a mix of α+β secondary structure. To this end, we selected the 30S ribosomal protein S10 (rpS10), an α+β structured protein from *E. coli* as the third model protein in this work. Taking a route similar to the previous protein systems, we first studied the impact of amino acid substitution on the overall dynamics of the reduced alphabet rpS10 (RA-rpS10) variants in comparison to the wt-rpS10

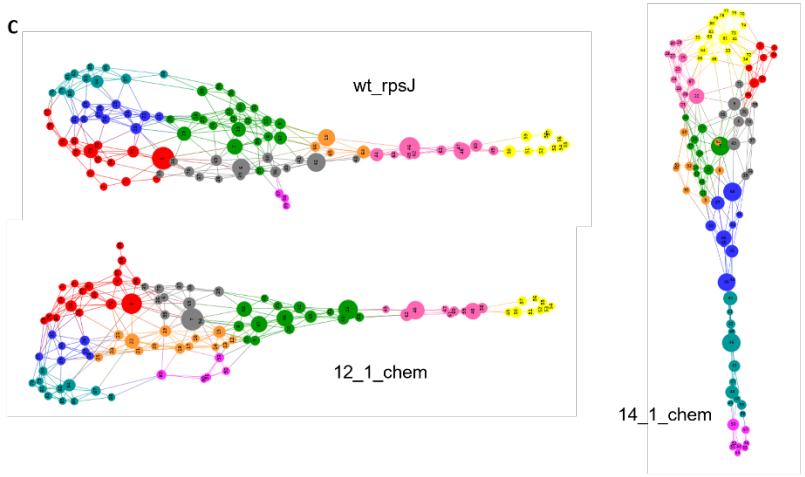
dynamics. We analyzed 1μs trajectories of the variants and compared their RMSD and RMSF trends to the wt-rpS10. The overall RMSD trends of all 12A variants were similar to wt-rpS10 (Figure 4.5.A). Quantitatively, the RMSD of the wt-rpS10 was 5.9 Å, whereas the RMSDs of the 12A variants were 5.8 Å, 6.6 Å, 6.7 Å, and 7.4 Å, for 12_1_chem, 12_1_cons, 12_2_chem, and 12_2_cons variants, respectively. A comparable trend was observed for the RMSF values of the 12A variants (Figure 4.5.B), where the RMSFs of the 12A variants were 2.2 Å, 2.2 Å, 2.1 Å, and 2.6 Å, compared to the RMSF of 2.3 Å for the wt-rpS10. The extended loop region is the most flexible region in the wt-rpS10 as well as all the RA-rpS10 variants. Since the function of the extended loop is to protrude into the 30S subunit and interact with other ribosomal proteins and the 16S rRNA, preserving the dynamics of the extended loop served as an important criterion for ranking the RA-rpS10 variants. Subsequent analysis of RMSD and RMSF values of the other RA-rpS10 variants also showed comparable behaviours of both chemistry and conservation-based variants (Appendix Table 4.4).

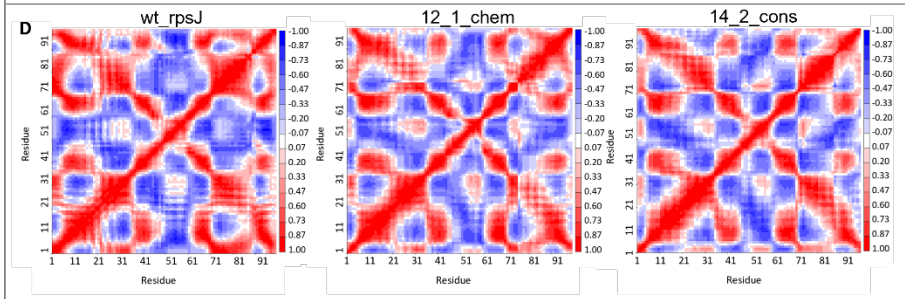
The network structures and amino acid *Bx(n)* values of the 12A variants were compared with those of the wt-rpS10. The overall network shapes (Figure 4.5.C) of all 12A variants were similar to the wt-rpS10 network shape. Additionally, of the top ten residues with the highest *Bx(n)* values, all 12A variants preserved at least five residues in their respective networks (Appendix Table 4.7). Interestingly, four residues from the extended loop qualify into the top ten residues with the highest *Bx(n)* values in the wt-rpS10 and are also preserved by the 12A rpS10 variants. We also analyzed the network properties of other RA-rpS10 variants and found that both chemistry and conservation-based variants show similar network structures and *Bx(n)* values of residues at each step of alphabet reduction (Appendix Table 4.4).

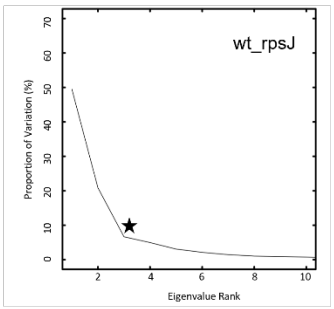
We further investigated the motions sampled by the RA-rpS10 variants and the cross-correlations between the residues. A close inspection of the heatmaps demonstrates a near identical behaviour among the 12A variants and with the wt-rpS10 (Figure 4.5.D), a similarity which is further noticed in the sampling time of different motions in the 12A variants. The scree plot of the wt-rpS10 samples three motions above the scree point, for a total of 75 percent of simulation time.



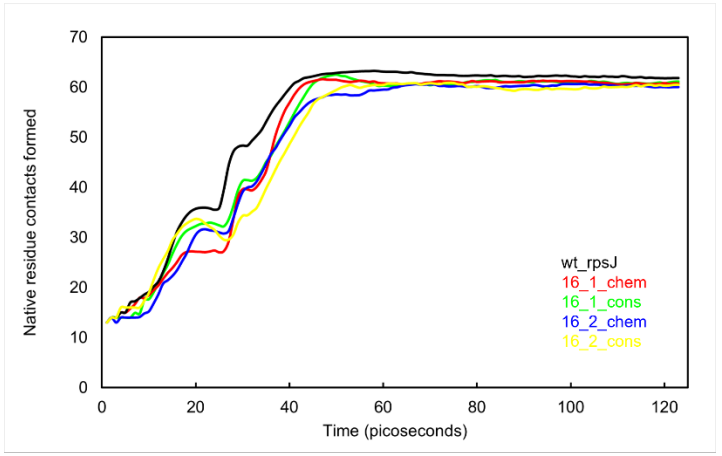








Variant	Motions sampled above scree point
12_1_chem	3
12_1_cons	3
12_2_chem	3
12_2_cons	3



[Figure on previous page]

Figure 4.5 – *In silico* assessment of RA-rpS10 variants. A-B. RMSD and RMSF results for wt-rpS10 and 12A variants. Both chemistry and conservation-based variants show similar RMSDs and RMSFs as those of wt-rpS10. C. Representative results for network structure of wt-rpS10 followed by results of selected variants which demonstrate a noticeable similarity or difference with respect to wt-rpS10, in this case, 12_1_chem, and 14_1_chem variants, respectively. Residues are shown as circular nodes and labeled according to the residue numbering. Residues are coloured based on the domain they belong to, and the size of the node represents the *Bx(n)* value of the residue. D. PCA derived residue cross-correlation heatmaps of wt_rpS10 followed by results of selected variants which demonstrate a noticeable similarity or difference with respect to wt-rpS10, in this case, 12_1_chem, and 14_1_chem variants, respectively. High positive cross-correlations are shown with higher intensity of red and high negative cross-correlations are shown with higher intensity of blue, whereas white shows no correlation. E. PCA derived scree plot for wt-rpS10. The x-axis shows the top ten motions sampled, and the y-axis shows the time (as a percentage) that the system spends sampling each motion. The star symbol shows the scree point. The sampling times for 12A variants are shown in the table. F. SMOG derived folding pathway analysis of wt-rpS10 and 12A variants. The x-axis shows the time in picoseconds and the y-axis shows the number of native residue contacts formed. The folding pathway adopted by all the 12A variant are different from that of wt-rpS10, although the 12A variants preserve almost all of the native residue contacts after folding.

The 12A variants also sample three motions above the scree point, for 65 percent, 72 percent, 63 percent, and 70 percent of simulation times for 12_1_chem, 12_1_cons, 12_2_chem, and 12_2_cons variants, respectively (Figure 4.5.E). Similar results were observed for all other RA-rpS10 variants, and the data has been compiled into a table (Appendix Table 4.4). Altogether, the PCA results for 12A variants and other RA-rpS10 variants established that the RA-rpS10 variants designed using both conservation and chemistry-based substitutions behave similarly with respect to wt-rpS10.

To study another critical property of protein dynamics, which is the folding of the polypeptide chain into the final 3D folded structure, we employed SMOG analysis and studied the folding pathway taken by the linear chains of the 12A rpS10 variants, compared to that of the wt-rpS10, along with the number of residue contacts formed (Figure 4.5.F). Results reveal that all of the 12A variants adopted folding pathways different from the wt-rpS10. However, it is surprising that even after going through different folding pathways, the variants were able to preserve almost all the residue contacts. For a total of 62 contacts formed in wt-rpS10, the 12A variants were able to form 60, 61, 59, and 60 contacts in 12_1_chem, 12_1_cons, 12_2_chem, and 12_2_cons variants, respectively. Similar results were obtained for other RA-rpS10 variants, where we

witnessed the linear chains of variants taking different folding pathways, but eventually, they all converged to a similar structure and preserved almost all native residue contacts (Appendix Table 4.4).

In summary, both chemistry and conservation-based RA-rpS10 variants performed equally in preserving molecular dynamics of the wt-rpS10 at each step of alphabet reduction. The scores obtained by the variants were extremely close to each other, thereby making the ranking process a bit challenging. Nevertheless, based on the scores of the variants, five out of the ten times, chemistry-based variants were taken forward to the next level of alphabet reduction, whereas conservation-based variants were selected the other five times. Together, these results underline the importance of the RAP scoring and ranking strategy designed and implemented in this work towards identifying the sometimes small differences in the dynamic properties of the RA-variants. Such differences can be used to rank the RA-variants, thereby generating a hierarchy of the best RA-variants which closely resemble the wt-protein dynamics, thus having higher chances of preserving the wt-protein functions, an advantage not offered by other RAP design approaches reported previously.

4.6 Discussion

Previous attempts to design reduced alphabet proteins using strategies such as random mutagenesis (136) or binary pattern module of polar and non-polar amino acids (35) have resulted in a loss of function even while completely preserving the structure. We attribute this failure to the lack of screening for preservation of the dynamic properties of the proteins. Additionally, reliance on stringent amino acid substitution criteria such as mentioned above can also be the cause for hampering the preservation of parent protein properties. In this work, we employ a combination of MD simulations coupled with downstream computational analysis to investigate how altering the primary structure of the protein affects its dynamics. Targeting the two most important properties of amino acids, physicochemistry and conservation, we employ a substitution principle with a broader basis. Testing our pipeline on three different proteins, we show for the first time that RAP design seems to vary with secondary structural composition of proteins. Our results show that in a

protein with dominant α-helical secondary structure composition, such as chorismate mutase, the conservation-based reduced alphabet variants outperform chemistry-based variants in preserving native protein dynamics (Figure 4.6.A). However, in a protein with dominantly β-pleated composition, such as IF1, the chemistry-based variants preserve native dynamics better than the conservation-based variants (Figure 4.6.B). Lastly, our finding for an α+β mix composition of a protein, such as rpS10, show that both chemistry and conservation-based variants perform equally in maintaining native protein dynamics (Figure 4.6.C). Together these findings suggest that the secondary structure composition of a protein plays a key role in determining which substitution strategy would be preferred for designing reduced alphabet variants for the protein, while preserving dynamics of the parent protein.

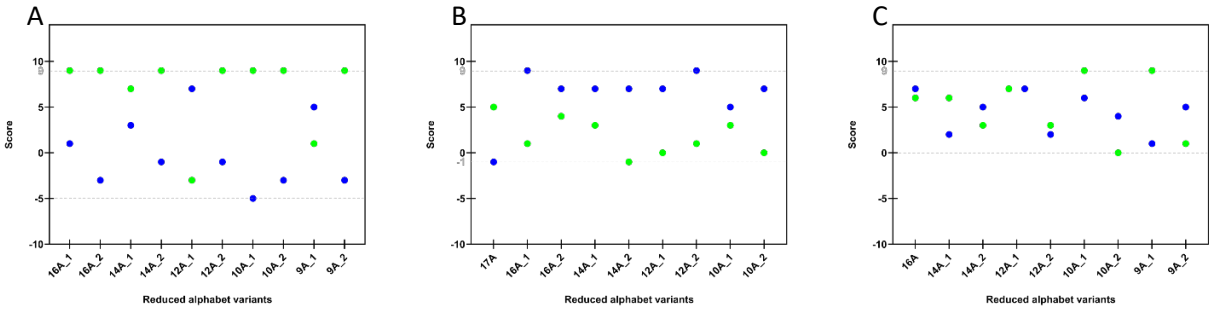


Figure 4.6 – Final scores of RA-variants of the three model proteins. A. Chorismate mutase, B. Initiation Factor 1 (IF1), C. 30S ribosomal protein 10 (rpS10). The x-axis shows the AAA sizes, and the y-axis shows the final scores of the variants based on their performance in individual computational analyses. The highest and lowest scores obtained by the RA-variants are denoted by dashed grey lines. Consistently with the color scheme shown in the RAP design flowchart (Figure 4.1), blue dots represent the chemistry-based variants and green dots represent the conservation-based variants.

To emphasize the significance of preserving native dynamics for designing functional RAPs, we employed a set of techniques to study different aspects of protein dynamics. Our findings demonstrate that reduced alphabet variants based on different substitution strategies behave differently compared to the parent protein. Analysis of the RMSD and RMSF of the MD trajectories of the RA-variants provided insight into the overall behaviour of the variants, and therefore was used as the first tier of assessing the RA-variants. A visual analysis of the MD trajectories of RA-variants such as the 9_2_chem variant of chorismate mutase (Appendix Table 4.2) and 14_2_cons variants of IF1 (Appendix Table 4.3) reveal that the disruption of secondary structure of the protein

is coupled with high RMSD and RMSF values. These findings prove that insertion of unsuitable amino acids can severely impact both the structure and dynamics of the RA-variants. Subsequent analysis of the protein networks, motions, and folding of the RA-variants served as more stringent criteria to differentiate between the dynamic behaviour of the variants compared to the parent proteins. The difference in network shape of the 10_2_chem variant of IF1 (Figure 4.4.C), and major differences in the residue cross-correlation heatmaps (Figure 4.4.D), further showcase how slight differences in amino acid composition (e.g., replacing phenylalanine with tyrosine at two positions in a 91-residue chorismate mutase, a 2% change in the amino acid composition of the RAP variant versus the wt-cm) can have large-scale impact on protein dynamics. Collectively, the computational techniques employed in this work to study protein dynamics served as sequential tiers for scoring the RA-variants and thereby facilitating the identification of the best RA-variants for each protein. Together, our findings confirm that altering the amino acid composition of proteins can have large-scale implications on protein dynamics, which can be studied using a combination of protein dynamic analysis techniques. Finally, we employed AlphaFold (164) to predict the structures of the 8A variants designed for the three model proteins. The structures predicted by AlphaFold were nearly identical to the structures of the parent proteins (Appendix Figure 4.8), with RMSD of 1 Å between the predicted structures and the parent protein structures. Therefore, it can be concluded that our RAP design pipeline can generate variants that preserve both the structure and dynamics of the parent protein, providing confidence and reliability to our RAP design pipeline. To summarize these findings in the context of the relationship between the structure, dynamics and function of proteins, it can be established that reduced alphabet proteins that preserve the structure and dynamics of the parent protein are perhaps more likely to preserve function as well, compared to the variants that preserve only structure or dynamics.

The biggest drawbacks of the previous works on RAP design were the lack of rational substitution criteria such as combination of polar and nonpolar residues, random mutagenesis, etc., however, we utilized a more rational and unbiased approach where the least conserved amino acids were substituted first. This approach allowed for reducing the AAA size while preserving the key amino acids which are conserved due to their important role in maintaining either the structure

or function of the protein. Following this approach, we were able to reduce the AAA of our model proteins to 8 AAA, the smallest AAA designed in this work. Starting from the 18 AAA for all three model proteins, 8 AAA variants include only 44% of the initial alphabet size, which is a drastic drop in AAA size. These results suggest that as long as a rational AAA substitution strategy is employed along with an attempt to preserve key structural or functional residues, the AAA of a protein can be reduced to a great extent without losing structure or function. Furthermore, as a measure of the efficiency and robustness of the RAP design pipeline, having several scoring criteria such as RMSD, RMSF, network structures, *Bx(n)* values of residues, folding pathways, residue cross-correlations, etc., served as a comprehensive testing matrix of the RAP variants, thereby increasing our confidence in the identification of the best RAP. In the future, this RAP design pipeline can be expanded for testing other proteins with diverse structures and functions for a more thorough testing of our substitution strategies and providing feedback to our RAP design pipeline to make it more robust for a wide range of proteins.

With a goal of the rational design of reduced alphabet proteins, in this work we have showcased a rational RAP design pipeline that can efficiently generate reduced alphabet variants which preserve both structure and dynamics of the parent protein *in silico*. Highlights of our RAP design pipeline include performing amino acid substitution based on more than one rationale, in-depth evaluation of protein dynamics using multiple analyses, and robust scoring and ranking criteria enabling the identification of the best RAP variants. Furthermore, preliminary *in vivo* testing has been performed for 12A-rpS10 variants. *E. coli* knockdown strains of rpS10 were transformed with plasmids carrying RAP-rpS10 coding sequences. Downstream analysis of the RAP-strains shows growth phenotype of the E. coli cells containing the 12A-rpS10 variant (data not shown). However, additional testing of the 12A-rpS10 and other RAP variants needs to be performed to validate our findings. Additionally, preliminary protein overexpression and purification studies have been performed on multiple RA-IF1 variants. The coding sequences of the wt and the RA-IF1 variants were cloned into the pET28b plasmid using restriction enzymes Nde1 and EcoR1. The plasmids were then transformed into DH5α competent cells, followed by overexpression of the proteins in BL21(DE3) cells where protein production was induced using IPTG. Subsequently, the

wt and the RA-IF1 variants were purified using nickel affinity chromatography and were tested for solubility where all tested proteins precipitated out of the solution (data not shown). Therefore, future directions could involve optimizing protein expression, cell lysis, and purification conditions to promote protein solubility.

One of the main arguments for pursuing reduced alphabet protein design is the quest to understand the origin of life. While this is important, reducing protein alphabet complexity will also provide a basis for understanding general protein design rules for different classes of proteins, offering several biotechnological implications. It can provide a basis for the forward-engineering of proteins on a reduced alphabet scaffold leading to the generation of functional *de novo* proteins (165-167) such as enzymes with novel or enhanced catalytic activities. A reduced alphabet protein that functions optimally will allow full rationality in protein design and will open up the sequence space for the incorporation of non-native amino acids. The design of an *E. coli* strain that runs on reduced alphabet proteins, although a far-reaching objective, could enable significant strain orthogonality, preventing crosstalk with natural cellular systems and eliminating the challenges of using genetically modified organisms in bioproduction and other bio-applications.

Additionally, our RAP design pipeline can also be extended to eukaryotic proteins with biomedical and pharmaceutical applications. Proteins such as eIF4E (eukaryotic Translation Initiation Factor 4E) are essential for eukaryotic translation initiation but are also upregulated in multiple cancer types (168) in humans. Therefore, targeted substitution of amino acid residues that interact with oncogenic partners such as PI3K/AKT/mTOR can be performed to disrupt their interaction. Such a result can be achieved by either introducing amino acid substituents that would either modify the shape of the binding site (due to a different preferred secondary structure of the introduced amino acid) or would disrupt the interaction between the protein partners (due to different physicochemical properties of the side chain of the introduced amino acid), thereby abolishing the progression of the oncogenic signaling downwards. Using RAPs for such an application is particularly useful because mass substitutions of amino acids can allow a single RAP to abolish interactions with multiple partners thus eliminating the tedious task of repeated identification and mutation of specific residues in target proteins. However, prior to that, ways for

targeted delivery of RAPs, in this case RA-eIF4E needs to be determined. Since eIF4E plays an essential role in eukaryotic translation initiation, downstream experimental testing can be performed using either *in vitro* transcription/translation kits or by using *in vivo* survival phenotype assays performed in yeast. Overall, the expansion of our test set of proteins (currently consisting of archaeal and bacterial proteins) to include eukaryotic proteins will help us to understand protein design rules from all the different domains of life, thus offering an understanding of the evolution of proteins while guiding protein engineering studies in various life forms.

With *ex vivo* systems (e.g., PURE system), RAPs might provide the springboard for the generation of self-replicating *in vitro* transcription/translation kits. Self-replicating protein expression kits that have RAP-enzyme components might have increased activity as amino acid flux will be directed from self-replication/maintenance to maximal product yield. In fact, this is applicable to *in vivo* systems as well; the reduced energy requirement for cellular metabolism in a 'reduced complexity strain' can allow increased recombinant product yield. Such tools (reduced alphabet *in vivo* and *ex vivo* systems) can then be used for tasks such as bioremediation, point-of-need vaccine production and even for establishing life on remote planets (since reduced complexity protein modules for *ex vivo* systems would ensure reduced payload from Earth).

CHAPTER 5: SUMMARY AND CONCLUSIONS

Proteins, composed of long amino acid chains, are the most abundant biomolecules in a cell and are at the forefront of the majority of cellular functions. In the present world where protein engineering lies at the epicentre of bioengineering, it is essential to understand the protein design rules that exist in nature to enable modification of proteins and unlock novel structure and functions. Moreover, given the overlap in physicochemical properties of the standard amino acids, one or more of such possibly redundant amino acids can be removed to artificially design a reduced amino acid alphabet and give rise to reduced alphabet proteins (RAPs). Although several researchers have attempted to design RAPs in the past, most have resulted in partial or complete loss of function in the resultant RAPs, without a clear understanding of what caused the loss of activity. This research thesis was designed to understand the protein design rules that exist in nature to be utilized for efficient forward engineering of proteins, particularly to design RAPs. This chapter serves to conclude the thesis by summarizing the key findings with respect to the research aims and will also discuss the value and contribution of this thesis in the field of protein engineering. This chapter will also review the limitations of this research along with proposing possible opportunities for future research.

5.1 Evolution of the Standard Amino Acid Alphabet and Identification of Protein Design Principles

The AAA size and composition varies for different proteins, thereby sparking curiosity whether the current AAA is the only alphabet to have existed on planet Earth since the beginning of time or whether it is a product of gradual evolution from a smaller subset known as the prebiotic amino acids (29, 30, 36). Moreover, the difference in the amino acid composition in proteins with different functions suggests the existence of underlying protein design principles where amino acids are selected differentially for various functions. Even after several decades of extensive research, little is known about the evolution of the standard amino acid set, or the underlying protein design principles.

In Chapter 2, an extensive bioinformatics analysis was performed to shed light on the complexity of the AAA and nature's protein design principles. Results indicated that small alphabet proteins (with less than 20 amino acids in their AAA) commonly exist in nature with an alphabet size as small as a 3 AAA. With respect to smaller alphabets, we observed a trend of amino acid exclusion where some amino acids are removed from the AAA more often, whereas certain amino acids are tightly preserved. Prebiotic amino acids (believed to be the first set of amino acids to be found on early Earth), make up majority of the highly preserved amino acids in proteins from diverse domains of life and with distinct functions. Additionally, for the first time, our results demonstrated the existence of a direct correlation between the protein sequence length and AAA size, implying that the increase in protein sequence length led to incorporation of additional amino acids into the AAA, or vice versa. Altogether, these findings strongly indicate that the modern-day standard AAA has gradually evolved from an initial smaller subset comprised mostly of prebiotic amino acids, in agreement with previous theories including the RNA world hypothesis (68). With respect to the protein design principles utilized in nature, our findings revealed that the amino acid composition of a protein is customized according to the function, and cellular localization of a protein.

Cumulatively, these findings can serve as a foundation to facilitate and benefit the field of protein engineering. With the understanding of the AAA evolution trends along with the design rules essential for allowing diverse functions, the AAA complexity of a protein can be altered to synthesize novel protein variants in a near-natural manner. In chapter 4, the design principles have been utilized to rationally reduce the AAA size of multiple proteins to generate RAP variants. It is crucial to highlight that the findings of this study were limited by the availability of protein sequence entries in the UniProt/Swiss-Prot database. Our Swiss-Prot derived dataset comprising of roughly half a million protein sequence entries was used as the input to our bioinformatics pipeline. Given the size of our dataset, one of the future directions is to expand on this initial dataset as more sequences become available, to further establish the accuracy of our bioinformatics pipeline as well as the results. Additionally, an inherent bias exists with respect to the representation of the different domains of life on the UniProt database, where the number of sequence entries deposited for bacterial and eukaryotic proteins is much higher than the number of entries deposited for

archaeal and viral proteins. The bias in the representation of the domains of life may result in a bias in the results, particularly when making interpretations for protein design rules for each of the domains of life. This bias can be eliminated by incorporating additional archaeal and viral proteins into the UniProt database. Furthermore, this work can be easily expanded to investigate specific properties of proteins such as pathological roles, various post-translational modifications, proteins with diverse domains and motifs, and proteins with different coding sequence variants.

5.2 Structural Dynamics of Nrp2 Reveals Motions Required for HCMV Proteins Binding

Nrp2 serves as an essential cell surface receptor and plays a central role in a wide range of physiological and signaling processes. Viruses such as HCMV and SARS-CoV-2 exploit Nrp2 to gain entry into the host cells. Several previous studies have suggested that a major conformational change is required, particularly of the a1 domain of Nrp2, for the viral proteins to bind. However, the details pertaining to the conformational landscape of Nrp2 prior to and during the viral protein binding still remain elusive, since little is known about the structural dynamics and conformational flexibility of Nrp2, particularly with respect to the presence and absence of the Ca^{2+} ion in the Ca^{2+}-binding site of the a2 domain of Nrp2. Chapter 3 employed in-house developed molecular dynamics guided computational assessment of the structural dynamics of Nrp2, and how it is exploited by HCMV to gain entry into the host cells. Additionally, Chapter 3 served as a biological benchmark for the validation of our computational protein dynamics analysis methodology which is aimed at studying different aspects of biomolecular dynamics which are often neglected while performing structural or functional studies. The dynamic properties of a protein can include distinct motions adopted by the protein backbone or side chain in simulated environments, backbone torsion angles, communication between different residues in the protein, folding of the linear polypeptide chain into the folded three-dimensional structure of the protein. The biomolecular dynamics analysis methodology developed and benchmarked in Chapter 3 has been subsequently implemented to examine the structural dynamics of reduced alphabet protein variants designed in Chapter 4.

For the first time, our results reveal that a large-scale conformational change takes place in the absence of Ca^{2+} ion, where the a1 domain samples an opening motion with respect to the

core domains (a2b1b2) of Nrp2, which is likely exploited by the viral proteins to bind to the HCMV pentamer binding sites. This opening motion is further triggered when the core domains strongly repel the a1 domain in a highly coordinated fashion as demonstrated by the residue cross-correlation results. Furthermore, we demonstrate that the backbone torsion angles of residues in the Ca^{2+} binding site are altered, in the absence of the Ca^{2+} ion, thereby suggesting that Ca^{2+} binding reorients and restricts the backbone of specific residues into conformations that capriciously favour HCMV pentamer binding. Lastly, based on our findings, we propose a model suggesting the possible mechanism pertaining to how the structural dynamics and inherent conformational landscape of Nrp2 is exploited by the HCMV pentamer proteins to gain entry into the host cells. The mechanistic details of Nrp2 dynamics identified in Chapter 3 can prove beneficial for the development of antiviral therapies for Nrp2.

Altogether, Chapter 3 provides unprecedented insight into the mechanistic details of the conformational landscape of Nrp2 using our in-house developed biomolecular dynamics analysis approach. Although we observed the a1 domain opening only in the absence of Ca^{2+}, we speculate it is equally possible for the Nrp2-Ca^{2+} holo-protein to sample the domain a1 opening. Therefore, future directions include performing extended simulations of the Nrp2-Ca^{2+} system to investigate whether the a1 domain opening motion is witnessed in the presence of Ca^{2+}. Additional replicates of both *apo* Nrp2 and Nrp2-Ca^{2+} systems can be subjected to MD simulation coupled with downstream analysis to increase the sample size and to witness the opening motion in additional replicates. Furthermore, in view of the recent studies showing the interaction between the HCMV pentamer protein and Nrp2 (91, 105), the next step includes the MD simulation and analysis of the HCMV pentamer-Nrp2 complex to identify the mechanistic details of interaction between the two proteins. As a long-term goal, downstream investigation of the HCMV pentamer-Nrp2 complex can be performed in the presence of potential drugs that disrupt the interaction between the two, thereby facilitating the development of an HCMV vaccine targeting the pentamer complex. On the experimental side, future studies will involve investigating the *in vivo* mechanism of the receptor operation, since the biological role of the a1 domain opening motion still remains unknown.

5.3 Rational Design of Reduced Alphabet Proteins

In conventional protein design, the natural sequence space of a protein comprising the twenty standard amino acids is probed and reorganised to generate multiple variants. However, the standard amino acid alphabet (AAA) is not only redundant in terms of physicochemical properties of the twenty amino acids, but also restricts exploration of novel structures and functions that can be unlocked using unnatural amino acids. Reducing the AAA size of existing proteins is an approach where the overlap between the standard amino acids can be used to design reduced alphabet proteins (RAPs) with several downstream applications. On one hand, RAPs can identify the 'bare minimum' number of amino acids required to sustain a functional protein, whereas on the other hand, the "freed-up spaces" after removing overlapping amino acids can be utilized to insert unnatural amino acids to assign novel properties to desired proteins. Several previous studies have attempted to design RAPs with little to no success, which we speculate to have happened because the focus was only on retaining the structure of the parent protein. Since proteins are dynamic entities, the structural dynamics serve as the connecting link between the structure and function of the protein, and thus a failure of preserving the dynamics may lead to partial or total loss of function. Chapter 4 reports a generalizable and systematic RAP design framework which utilizes two fundamental properties of amino acids to generate RAP variants: physicochemical properties and conservation trends. We utilized the biomolecular dynamics analysis methodology benchmarked in Chapter 3 to design RAP variants and to identify the best variants at each step of alphabet reduction. Our RAP design framework has been tested on three different model protein systems with distinct structures and functions to identify the design principles utilized in nature which can be subsequently applied to achieve forward engineering of proteins.

For the three model protein systems represented by chorismate mutase (α-helical archaeal enzyme), IF1 (β-pleated translation initiation factor) and rpS10 (α+β ribosomal protein), our results show that altering the amino acid composition significantly impacts the structure and dynamics of the RAP variants. Additionally, for the first time, we reveal that the secondary structure content of a protein plays a determining role as to which substitution strategy (physicochemistry-based vs. conservation-based) can generate RAP variants that can closely resemble parent protein

dynamics. Conservation-based RAP variants outperform the chemistry-based variants in α-helical chorismate mutase, whereas chemistry-based variants performed better at preserving parent protein dynamics in β-pleated IF1. In the case of α+β rpS10, both conservation- and chemistry-based RAP variants demonstrate similar efficiency in conserving parent protein dynamics. Together these findings reveal that there is a direct relation between the secondary structural organization of a protein and the substitution strategy of RAP design, thereby generating RAP variants with increased likelihood of preserving parent protein dynamics. Furthermore, the combination of protein dynamics analysis techniques used in this work teased out the miniscule differences in the structural dynamic properties of the different RAP variants, which facilitated the ranking of the RAP variants and assisted with selecting the best ones going forward. Lastly, as a proof of concept that our RAP design pipeline works well in preserving the similarity in structure and dynamics of the RAP variants relative to the parent protein, we employed AlphaFold, a revolutionary protein structure prediction server, to predict the structure of our 8 AAA variants. Our results show that the 8 AAA variants for all three model systems were nearly identical to the parent protein structures with an RMSD of ≤1Å.

Altogether our findings demonstrate that our RAP design framework can identify the protein design principles employed in nature and can use them to significantly reduce the complexity of the AAA of proteins without compromising the structural and dynamic properties of the system. Such RAP variants may show higher likelihood of preserving native function as opposed to RAP variants that preserve only the structure. As a future-direction, the biochemical properties of the three model proteins can be tested experimentally to validate our findings. Additionally, this RAP design framework can be tested on other proteins with more complex structures and functions to make this pipeline more generalizable and robust.

5.4 Final Remarks

The overarching goal of this thesis was to investigate the amino acid alphabets of natural proteins in order to identify the principles of protein design in nature with an aim of utilizing those principles for the rational reduction of the AAAs of proteins to design RAPs.

This thesis provides valuable insight into the protein design principles in nature with pioneering evidence that the AAA size is directly proportional to the length of proteins and that the amino acid compositions are customized to suit the function and cellular location of proteins. Furthermore, highlighting the importance of protein dynamics, this thesis develops and benchmarks a protein dynamics analysis toolkit where MD simulations coupled with downstream computational investigation techniques reveals unforeseen motions in Nrp2, offering applications in development of antiviral therapies. Lastly, using the above-mentioned pipeline, this thesis demonstrates that the AAA of proteins can be significantly reduced without compromising protein structure or dynamics, allowing generation of RAP variants with increased likelihood of preserving parent protein function. Altogether this thesis extensively studies the amino acid alphabets of proteins to address some of the key unanswered questions in the field of protein evolution and protein engineering.

REFERENCES

1. ChEBI-EMBL-EBI, CHEBI:83813 - proteinogenic amino acid. (2015).

2. Y. Lu, S. Freeland, On the evolution of the standard amino-acid alphabet. *Genome Biol* **7**, 102 (2006).

3. A. Bock *et al.*, Selenocysteine: the 21st amino acid. *Mol Microbiol* **5**, 515-520 (1991).

4. J. A. Krzycki, The direct genetic encoding of pyrrolysine. *Curr Opin Microbiol* **8**, 706-712 (2005).

5. M. Rother, J. A. Krzycki, Selenocysteine, pyrrolysine, and the unique energy metabolism of methanogenic archaea. *Archaea* **2010** (2010).

6. Y. Zhang, P. V. Baranov, J. F. Atkins, V. N. Gladyshev, Pyrrolysine and selenocysteine use dissimilar decoding strategies. *J Biol Chem* **280**, 20740-20751 (2005).

7. F. T. Senguen, T. M. Doran, E. A. Anderson, B. L. Nilsson, Clarifying the influence of core amino acid hydrophobicity, secondary structure propensity, and molecular volume on amyloid-beta 16-22 self-assembly. *Mol Biosyst* **7**, 497-510 (2011).

8. L. G. Wade, Organic Chemistry 7th Edition, Chapter 24 - Amino Acids, Peptides, and Proteins. 1153-1199.

9. A. B.-C. Interest, A Brief Guide to the Twenty Common Amino Acids. (2014).

10. G. Litwack, "Human Biochemistry". (2018), https://doi.org/10.1016/C2009-0-63992-1 chap. Protein Biosynthesis.

11. R. D. Sussmuth, A. Mainz, Nonribosomal Peptide Synthesis-Principles and Prospects. *Angew Chem Int Ed Engl* **56**, 3770-3821 (2017).

12. V. H. B. Serrao *et al.*, The Specific Elongation Factor to Selenocysteine Incorporation in Escherichia coli: Unique tRNA(Sec) Recognition and its Interactions. *J Mol Biol* **433**, 167279 (2021).

13. N. Fischer *et al.*, The pathway to GTPase activation of elongation factor SelB on the ribosome. *Nature* **540**, 80-85 (2016).

14. Y. Qian, R. Zhang, X. Jiang, G. Wu, The constraints between amino acids influence the unequal distribution of codons and protein sequence evolution. *R Soc Open Sci* **8**, 201852 (2021).

15. D. G. Longstaff *et al.*, A natural genetic code expansion cassette enables transmissible biosynthesis and genetic encoding of pyrrolysine. *Proc Natl Acad Sci U S A* **104**, 1021-1026 (2007).

16. A. Ambrogelly, S. Palioura, D. Soll, Natural expansion of the genetic code. *Nat Chem Biol* **3**, 29-35 (2007).

17. K. Kuriyama, M. Hirouchi, [Structure and function of gamma-aminobutyric acid (GABA) receptor: current state and prospectives]. *Nihon Yakurigaku Zasshi* **94**, 7-15 (1989).

18. L. A. C. Antonin Ginguay, "Encyclopedia of Biological Chemistry (Third Edition)" in Encyclopedia of Biological Chemistry (Third Edition). (2021), chap. Amino Acids | Amino Acid Metabolism☆.

19. T. C. Reynolds JA, Nature's Robots: A History of Proteins (Oxford Paperbacks). . *Oxford University Press*, 15 (2003).

20. L. Brody (2022) Levels of protein organization-National Human Genome Research Institute. in *Protein*, pp About Genomics / Educational Resources / Talking Glossary of Genomic and Genetic Terms / Protein. https://www.genome.gov/genetics-glossary/Protein. Date Accessed: September 29, 2022.

21. G. K. Philip, S. J. Freeland, Did evolution select a nonrandom "alphabet" of amino acids? *Astrobiology* **11**, 235-240 (2011).

22. E. Munoz, M. W. Deem, Amino acid alphabet size in protein evolution experiments: better to search a small library thoroughly or a large library sparsely? *Protein Eng Des Sel* **21**, 311-317 (2008).

23. H. Y. Zhang, Exploring the evolution of standard amino-acid alphabet: when genomics meets thermodynamics. *Biochem Biophys Res Commun* **359**, 403-405 (2007).

24. I. P. Shabalkin, P. I. Shabalkin, A. S. Iagubov, [Evolution of the genetic alphabet and amino acid code]. *Zh Evol Biokhim Fiziol* **39**, 488-494 (2003).

25. A. Fernandez, Lower limit to the size of the primeval amino acid alphabet. *Z Naturforsch C J Biosci* **59**, 151-152 (2004).

26. C. Mayer-Bacon, S. J. Freeland, A broader context for understanding amino acid alphabet optimality. *J Theor Biol* **520**, 110661 (2021).

27. A. D. Solis, Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins. *Proteins* **83**, 2198-2216 (2015).

28. A. S. Burton, J. C. Stern, J. E. Elsila, D. P. Glavin, J. P. Dworkin, Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. *Chem Soc Rev* **41**, 5459-5472 (2012).

29. S. Pizzarello, A. L. Weber, Prebiotic amino acids as asymmetric catalysts. *Science* **303**, 1151 (2004).

30. P. van der Gulik, S. Massar, D. Gilis, H. Buhrman, M. Rooman, The first peptides: the evolutionary transition between prebiotic amino acids and early proteins. *J Theor Biol* **261**, 531-539 (2009).

31. J. S. Italia *et al.*, An orthogonalized platform for genetic code expansion in both bacteria and eukaryotes. *Nat Chem Biol* **13**, 446-450 (2017).

32. J. T. Wong, Evolution of the genetic code. *Microbiol Sci* **5**, 174-181 (1988).

33. M. Kimura, S. Akanuma, Reconstruction and Characterization of Thermally Stable and Catalytically Active Proteins Comprising an Alphabet of ~ 13 Amino Acids. *J Mol Evol* **88**, 372-381 (2020).

34. D. S. Riddle *et al.*, Functional rapidly folding proteins from simplified amino acid sequences. *Nat Struct Biol* **4**, 805-809 (1997).

35. K. U. Walter, K. Vamvaca, D. Hilvert, An active enzyme constructed from a 9-amino acid alphabet. *J Biol Chem* **280**, 37742-37746 (2005).

36. R. Shibue *et al.*, Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci Rep* **8**, 1227 (2018).

37. A. Kawahara-Kobayashi, M. Hitotsuyanagi, K. Amikura, D. Kiga, Experimental evolution of a green fluorescent protein composed of 19 unique amino acids without tryptophan. *Orig Life Evol Biosph* **44**, 75-86 (2014).

38. M. R. Waterman, C. M. Jenkins, I. Pikuleva, Genetically engineered bacterial cells and applications. *Toxicol Lett* **82-83**, 807-813 (1995).

39. A. Dumas, L. Lercher, C. D. Spicer, B. G. Davis, Designing logical codon reassignment - Expanding the chemistry in biology. *Chem Sci* **6**, 50-69 (2015).

40. H. Neumann, Rewiring translation - Genetic code expansion and its applications. *FEBS Lett* **586**, 2057-2064 (2012).

41. J. M. Chalker, G. J. Bernardes, B. G. Davis, A "tag-and-modify" approach to site-selective protein modification. *Acc Chem Res* **44**, 730-741 (2011).

42. C. R. Hall *et al.*, Site-Specific Protein Dynamics Probed by Ultrafast Infrared Spectroscopy of a Noncanonical Amino Acid. *J Phys Chem B* **123**, 9592-9597 (2019).

43. A. J. Bannister, T. Kouzarides, Regulation of chromatin by histone modifications. *Cell Res* **21**, 381-395 (2011).

44. U. H. Shah, R. Toneatti, S. A. Gaitonde, J. M. Shin, J. Gonzalez-Maeso, Site-Specific Incorporation of Genetically Encoded Photo-Crosslinkers Locates the Heteromeric Interface of a GPCR Complex in Living Cells. *Cell Chem Biol* **27**, 1308-1317 e1304 (2020).

45. M. T. Reetz, D. Kahakeaw, R. Lohmer, Addressing the numbers problem in directed evolution. *Chembiochem* **9**, 1797-1804 (2008).

46. E. T. Parker *et al.*, A plausible simultaneous synthesis of amino acids and simple peptides on the primordial Earth. *Angew Chem Int Ed Engl* **53**, 8132-8136 (2014).

47. A. D. Solis, Reduced alphabet of prebiotic amino acids optimally encodes the conformational space of diverse extant protein folds. *BMC Evol Biol* **19**, 158 (2019).

48. E. V. Koonin, A. S. Novozhilov, Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99-111 (2009).

49. V. Kubyshkin, N. Budisa, The Alanine World Model for the Development of the Amino Acid Repertoire in Protein Biosynthesis. *Int J Mol Sci* **20** (2019).

50. V. Kubyshkin, N. Budisa, Anticipating alien cells with alternative genetic codes: away from the alanine world! *Curr Opin Biotechnol* **60**, 242-249 (2019).

51. H. Saito, The RNA world 'hypothesis'. *Nat Rev Mol Cell Biol* 10.1038/s41580-022-00514-6 (2022).

52. M. Neveu, H. J. Kim, S. A. Benner, The "strong" RNA world hypothesis: fifty years old. *Astrobiology* **13**, 391-403 (2013).

53. W. Ma, What Does "the RNA World" Mean to "the Origin of Life"? *Life (Basel)* **7** (2017).

54. A. Pressman, C. Blanco, I. A. Chen, The RNA World as a Model System to Study the Origin of Life. *Curr Biol* **25**, R953-963 (2015).

55. N. Lahav, The RNA-world and co-evolution hypotheses and the origin of life: implications, research strategies and perspectives. *Orig Life Evol Biosph* **23**, 329-344 (1993).

56. L. E. Orgel, F. H. Crick, Anticipating an RNA world. Some past speculations on the origin of life: where are they today? *FASEB J* **7**, 238-239 (1993).

57. L. E. Orgel, Evolution of the genetic apparatus. *J Mol Biol* **38**, 381-393 (1968).

58. L. E. Orgel, Evolution of the genetic apparatus: a review. *Cold Spring Harb Symp Quant Biol* **52**, 9-16 (1987).

59. F. H. Crick, The origin of the genetic code. *J Mol Biol* **38**, 367-379 (1968).

60. F. H. Crick, S. Brenner, A. Klug, G. Pieczenik, A speculation on the origin of protein synthesis. *Orig Life* **7**, 389-397 (1976).

61. M. M. C. David L. Nelson, *Lehninger Principles of Biochemistry* (W.H. Freeman, New York, NY, ed. 7, 2017).

62. F. Wang *et al.*, A systematic survey of mini-proteins in bacteria and archaea. *PLoS One* **3**, e4027 (2008).

63. L. Wall, *Programming Perl*. L. Mike, Ed. (O'Reilly \& Associates, Inc., 2000), pp. 850.

64. R. Anderson, Basics of Bash. *Methods Mol Biol* **2443**, 161-180 (2022).

65. M. Cooper (2014) Advanced bash-scripting guide: An in-depth exploration of the art of shell scripting.

66. T. UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699 (2018).

67. E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, A. Bairoch, UniProtKB/Swiss-Prot. *Methods Mol Biol* **406**, 89-112 (2007).

68. L. E. Orgel, Some consequences of the RNA world hypothesis. *Orig Life Evol Biosph* **33**, 211-218 (2003).

69. C. R. Woese, G. E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**, 5088-5090 (1977).

70. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A* **87**, 4576-4579 (1990).

71. M. Boyer, M. A. Madoui, G. Gimenez, B. La Scola, D. Raoult, Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. *PLoS One* **5**, e15530 (2010).

72. S. Bhattacharyya, U. Varshney, Evolution of initiator tRNAs and selection of methionine as the initiating amino acid. *RNA Biol* **13**, 810-819 (2016).

73. T. Meinnel, Y. Mechulam, S. Blanquet, Methionine as translation start signal: a review of the enzymes of the pathway in Escherichia coli. *Biochimie* **75**, 1061-1075 (1993).

74. R. Chattapadhyay, H. Pelka, L. H. Schulman, Initiation of in vivo protein synthesis with non-methionine amino acids. *Biochemistry* **29**, 4263-4268 (1990).

75. S. Li, N. V. Kumar, U. Varshney, U. L. RajBhandary, Important role of the amino acid attached to tRNA in formylation and in initiation of protein synthesis in Escherichia coli. *J Biol Chem* **271**, 1022-1028 (1996).

76. L. Pallanck, L. H. Schulman, Anticodon-dependent aminoacylation of a noncognate tRNA with isoleucine, valine, and phenylalanine in vivo. *Proc Natl Acad Sci U S A* **88**, 3872-3876 (1991).

77. M. G. Kearse, J. E. Wilusz, Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev* **31**, 1717-1731 (2017).

78. R. P. Bywater, Why twenty amino acid residue types suffice(d) to support all living systems. *PLoS One* **13**, e0204883 (2018).

79. C. A. Cotton *et al.*, Underground isoleucine biosynthesis pathways in E. coli. *Elife* **9** (2020).

80. L. Jiang *et al.*, Abiotic synthesis of amino acids and self-crystallization under prebiotic conditions. *Sci Rep* **4**, 6769 (2014).

81. J. Yuan *et al.*, Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. *FEBS Lett* **584**, 342-349 (2010).

82. N. P. Lukashenko, [Expanding genetic code: amino acids 21 and 22--selenocysteine and pyrrolysine]. *Genetika* **46**, 1013-1032 (2010).

83. Y. Zhang, V. N. Gladyshev, High content of proteins containing 21st and 22nd amino acids, selenocysteine and pyrrolysine, in a symbiotic deltaproteobacterium of gutless worm Olavius algarvensis. *Nucleic Acids Res* **35**, 4952-4963 (2007).

84. L. Lin, R. J. Pinker, N. R. Kallenbach, Alpha-helix stability and the native state of myoglobin. *Biochemistry* **32**, 12638-12643 (1993).

85. T. O. Utset *et al.*, Modified anti-CD3 therapy in psoriatic arthritis: a phase I/II clinical trial. *J Rheumatol* **29**, 1907-1913 (2002).

86. A. L. Kolodkin *et al.*, Neuropilin is a semaphorin III receptor. *Cell* **90**, 753-762 (1997).

87. H. Chen, A. Chedotal, Z. He, C. S. Goodman, M. Tessier-Lavigne, Neuropilin-2, a novel member of the neuropilin family, is a high affinity receptor for the semaphorins Sema E and Sema IV but not Sema III. *Neuron* **19**, 547-559 (1997).

88. C. Pellet-Many, P. Frankel, H. Jia, I. Zachary, Neuropilins: structure, function and role in disease. *Biochem J* **411**, 211-226 (2008).

89. Q. Schwarz, C. Ruhrberg, Neuropilin, you gotta let me know: should I stay or should I go? *Cell Adh Migr* **4**, 61-66 (2010).

90. L. S. Gammill, C. Gonzalez, M. Bronner-Fraser, Neuropilin 2/semaphorin 3F signaling is essential for cranial neural crest migration and trigeminal ganglion condensation. *Dev Neurobiol* **67**, 47-56 (2007).

91. B. A. Appleton *et al.*, Structural studies of neuropilin/antibody complexes provide insights into semaphorin and VEGF binding. *EMBO J* **26**, 4902-4912 (2007).

92. G. Gerna, A. Kabanova, D. Lilleri, Human Cytomegalovirus Cell Tropism and Host Cell Receptors. *Vaccines (Basel)* **7** (2019).

93. N. Martinez-Martin *et al.*, An Unbiased Screen for Human Cytomegalovirus Identifies Neuropilin-2 as a Central Viral Receptor. *Cell* **174**, 1158-1171 e1119 (2018).

94. M. Raaben *et al.*, NRP2 and CD63 Are Host Factors for Lujo Virus Cell Entry. *Cell Host Microbe* **22**, 688-696 e685 (2017).

95. S. Niland, J. A. Eble, Neuropilins in the Context of Tumor Vasculature. *Int J Mol Sci* **20** (2019).

96. N. Kofler, M. Simons, The expanding role of neuropilin: regulation of transforming growth factor-beta and platelet-derived growth factor signaling in the vasculature. *Curr Opin Hematol* **23**, 260-267 (2016).

97. C. Pellet-Many *et al.*, Neuropilins 1 and 2 mediate neointimal hyperplasia and re-endothelialization following arterial injury. *Cardiovasc Res* **108**, 288-298 (2015).

98. J. L. Harman, J. Sayers, C. Chapman, C. Pellet-Many, Emerging Roles for Neuropilin-2 in Cardiovascular Disease. *Int J Mol Sci* **21** (2020).

99. F. Nakamura, Y. Goshima, Structural and functional relation of neuropilins. *Adv Exp Med Biol* **515**, 55-69 (2002).

100. M. Rossignol, M. L. Gagnon, M. Klagsbrun, Genomic organization of human neuropilin-1 and neuropilin-2 genes: identification and distribution of splice variants and soluble isoforms. *Genomics* **70**, 211-222 (2000).

101. P. Bork, Complement components C1r/C1s, bone morphogenic protein 1 and Xenopus laevis developmentally regulated protein UVS.2 share common repeats. *FEBS Lett* **282**, 9-12 (1991).

102. P. Bork, G. Beckmann, The CUB domain. A widespread module in developmentally regulated proteins. *J Mol Biol* **231**, 539-545 (1993).

103. L. A. Gregory, N. M. Thielens, G. J. Arlaud, J. C. Fontecilla-Camps, C. Gaboriaud, X-ray structure of the Ca2+-binding interaction domain of C1s. Insights into the assembly of the C1 complex of complement. *J Biol Chem* **278**, 32157-32164 (2003).

104. S. Takagi *et al.*, The A5 antigen, a candidate for the neuronal recognition molecule, has homologies to complement components and coagulation factors. *Neuron* **7**, 295-307 (1991).

105. D. Wrapp *et al.*, Structural basis for HCMV Pentamer recognition by neuropilin 2 and neutralizing antibodies. *Sci Adv* **8**, eabm2546 (2022).

106. T. U. Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2021).

107. A. Waterhouse, Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46**, W296-W303 (2018).

108. J. W. H.M. Berman, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, RCSB-PDB. *Nucleic Acid Research* **28**, 235-242 (2000).

109. C. R. G N Ramachandran, V Sasisekharan, Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**, 95-99 (1963).

110. A. Waterhouse, Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. , SWISS-MODEL Workspace/ GMQE. *Nucleic Acid Research* **46**, 296-303 (2018).

111. M. Bertoni, F. Kiefer, M. Biasini, L. Bordoli, T. Schwede, Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep* **7**, 10480 (2017).

112. L. Schrödinger, & DeLano, W., The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. (2020).

113. H. M. A. D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman, AMBER: calcium parameters -AmberTools. (2021).

114. D. A. Case *et al.*, The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668-1688 (2005).

115. D. R. Roe, T. E. Cheatham, 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084-3095 (2013).

116. S. Hayward, R. A. Lee, Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J Mol Graph Model* **21**, 181-183 (2002).

117. R. A. Lee, M. Razaz, S. Hayward, The DynDom database of protein domain motions. *Bioinformatics* **19**, 1290-1291 (2003).

118. K. Nguyen, P. C. Whitford, Steric interactions lead to collective tilting motion in the ribosome during mRNA-tRNA translocation. *Nat Commun* **7**, 10586 (2016).

119. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. *J Mol Graph* **14**, 33-38, 27-38 (1996).

120. D. D. Smith, D. Girodat, D. W. Abbott, H. J. Wieden, Construction of a highly selective and sensitive carbohydrate-detecting biosensor utilizing Computational Identification of Non-disruptive Conjugation sites (CINC) for flexible and streamlined biosensor design. *Biosens Bioelectron* **200**, 113899 (2022).

121. D. D. Smith, J. P. King, D. W. Abbott, H. J. Wieden, Development of a Real-Time Pectic Oligosaccharide-Detecting Biosensor Using the Rapid and Flexible Computational Identification of Non-Disruptive Conjugation Sites (CINC) Biosensor Design Platform. *Sensors (Basel)* **22** (2022).

122. E. Mercier, D. Girodat, H. J. Wieden, A conserved P-loop anchor limits the structural dynamics that mediate nucleotide dissociation in EF-Tu. *Sci Rep* **5**, 7677 (2015).

123. I. T. Jolliffe, *Principal component analysis for special types of data* (Springer, 2002).

124. M. Kschonsak *et al.*, Structural basis for HCMV Pentamer receptor recognition and antibody neutralization. *Sci Adv* **8**, eabm2536 (2022).

125. Y. Yao *et al.*, Hydroxychloroquine treatment on SARS-CoV-2 receptor ACE2, TMPRSS2 and NRP1 expression in human primary pterygium and conjunctival cells. *Exp Eye Res* **214**, 108864 (2022).

126. F. Humayun *et al.*, Abrogation of SARS-CoV-2 interaction with host (NRP1) neuropilin-1 receptor through high-affinity marine natural compounds to curtail the infectivity: A structural-dynamics data. *Comput Biol Med* **141**, 104714 (2022).

127. M. L. Saiz *et al.*, Epigenetic targeting of the ACE2 and NRP1 viral receptors limits SARS-CoV-2 infectivity. *Clin Epigenetics* **13**, 187 (2021).

128. A. A. El-Arabey, M. Abdalla, Transplacental transmission of SARS-CoV-2 infection via NRP1. *Travel Med Infect Dis* **40**, 101987 (2021).

129. I. Kyrou, H. S. Randeva, D. A. Spandidos, E. Karteris, Not only ACE2-the quest for additional host cell mediators of SARS-CoV-2 infection: Neuropilin-1 (NRP1) as a novel SARS-CoV-2 host cell entry mediator implicated in COVID-19. *Signal Transduct Target Ther* **6**, 21 (2021).

130. C. Hopkins, J. R. Lechien, S. Saussez, More that ACE2? NRP1 may play a central role in the underlying pathophysiological mechanism of olfactory dysfunction in COVID-19 and its association with enhanced survival. *Med Hypotheses* **146**, 110406 (2021).

131. E. N. Trifonov, The triplet code from first principles. *J Biomol Struct Dyn* **22**, 1-11 (2004).

132. C. Jackel, D. Hilvert, Biocatalysts by evolution. *Curr Opin Biotechnol* **21**, 753-759 (2010).

133. F. Nerattini *et al.*, Design of Protein-Protein Binding Sites Suggests a Rationale for Naturally Occurring Contact Areas. *J Chem Theory Comput* **15**, 1383-1392 (2019).

134. E. Guarnera, R. Pellarin, A. Caflisch, How does a simplified-sequence protein fold? *Biophys J* **97**, 1737-1746 (2009).

135. L. Regan, W. F. DeGrado, Characterization of a helical protein designed from first principles. *Science* **241**, 976-978 (1988).

136. S. Akanuma, T. Kigawa, S. Yokoyama, Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proc Natl Acad Sci U S A* **99**, 13549-13553 (2002).

137. G. MacBeath, P. Kast, D. Hilvert, A small, thermostable, and monofunctional chorismate mutase from the archaeon Methanococcus jannaschii. *Biochemistry* **37**, 10062-10073 (1998).

138. C. L. Pon, B. Wittmann-Liebold, C. Gualerzi, Structure--function relationships in Escherichia coli initiation factors. II. Elucidation of the primary structure of initiation factor IF-1. *FEBS Lett* **101**, 157-160 (1979).

139. P. Milon, C. Maracci, L. Filonava, C. O. Gualerzi, M. V. Rodnina, Real-time assembly landscape of bacterial 30S translation initiation complex. *Nat Struct Mol Biol* **19**, 609-615 (2012).

140. N. Riehl, P. Remy, J. P. Ebel, B. Ehresmann, Crosslinking of N-acetyl-phenylalanyl [s4U]tRNAPhe to protein S10 in the ribosomal P site. *Eur J Biochem* **128**, 427-433 (1982).

141. S. R. Eddy, Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechnol* **22**, 1035-1036 (2004).

142. M. P. Styczynski, K. L. Jensen, I. Rigoutsos, G. Stephanopoulos, BLOSUM62 miscalculations improve search performance. *Nat Biotechnol* **26**, 274-275 (2008).

143. A. Armon, D. Graur, N. Ben-Tal, ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* **307**, 447-463 (2001).

144. M. Landau *et al.*, ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* **33**, W299-302 (2005).

145. H. Ashkenazy, E. Erez, E. Martz, T. Pupko, N. Ben-Tal, ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529-533 (2010).

146. T. D. Schneider, R. M. Stephens, Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097-6100 (1990).

147. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).

148. A. Waterhouse *et al.*, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296-W303 (2018).

149. T. Schwede, J. Kopp, N. Guex, M. C. Peitsch, SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* **31**, 3381-3385 (2003).

150. G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95-99 (1963).

151. P. Li, Tutorial for the LEaP Program. (2016).

152. N. M. Glykos, Software news and updates. Carma: a molecular dynamics analysis program. *J Comput Chem* **27**, 1765-1768 (2006).

153. D. Girodat, E. Mercier, K. E. Gzyl, H. J. Wieden, Elongation Factor Tu's Nucleotide Binding Is Governed by a Thermodynamic Landscape Unique among Bacterial Translation Factors. *J Am Chem Soc* **141**, 10236-10246 (2019).

154. H. J. Wieden, E. Mercier, J. Gray, B. Steed, D. Yawney, A combined molecular dynamics and rapid kinetics approach to identify conserved three-dimensional communication networks in elongation factor Tu. *Biophys J* **99**, 3735-3743 (2010).

155. M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**, e98679 (2014).

156. J. Yoon, A. Blumer, K. Lee, An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics* **22**, 3106-3108 (2006).

157. I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* **374**, 20150202 (2016).

158. R. D. Ledesma, P. Valero-Mora, G. Macbeth, The Scree Test and the Number of Factors: a Dynamic Graphics Approach. *Span J Psychol* **18**, E11 (2015).

159. J. K. Noel *et al.*, SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput Biol* **12**, e1004794 (2016).

160. M. Burdukiewicz *et al.*, Amyloidogenic motifs revealed by n-gram analysis. *Sci Rep* **7**, 12961 (2017).

161. R. J. Petrella, M. Karplus, The energetics of off-rotamer protein side-chain conformations. *J Mol Biol* **312**, 1161-1175 (2001).

162. B. Vergani *et al.*, Backbone dynamics of Tet repressor alpha8intersectionalpha9 loop. *Biochemistry* **39**, 2759-2768 (2000).

163. R. B. Cattell, The Scree Test For The Number Of Factors. *Multivariate Behav Res* **1**, 245-276 (1966).

164. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

165. D. R. Hicks *et al.*, De novo design of protein homodimers containing tunable symmetric protein pockets. *Proc Natl Acad Sci U S A* **119**, e2113400119 (2022).

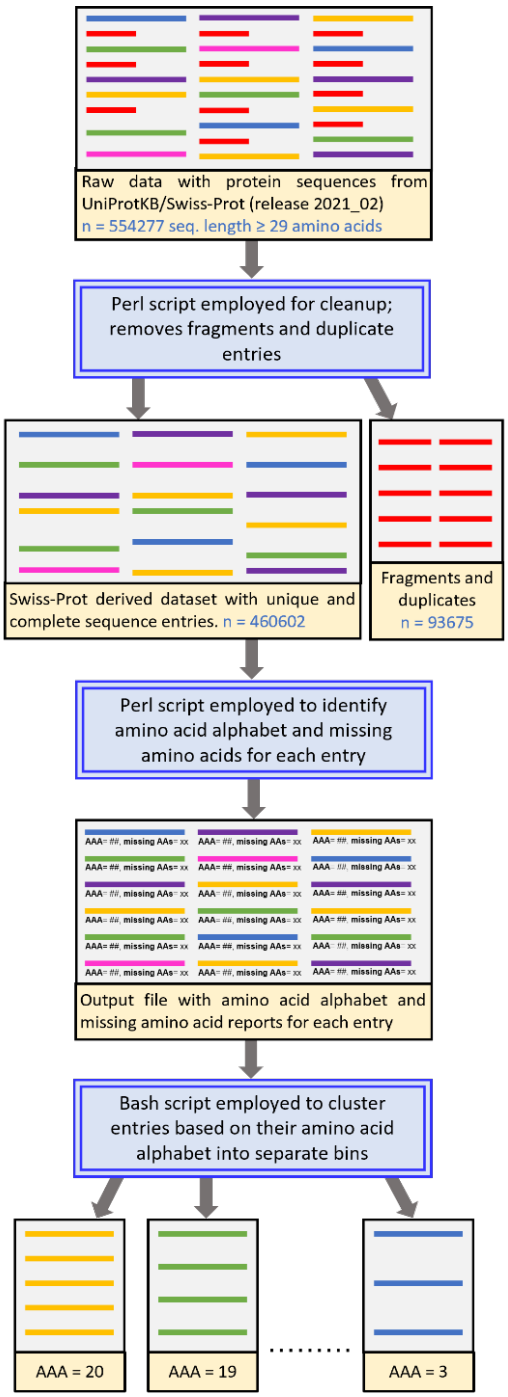
166. N. B. Woodall *et al.*, De novo design of tyrosine and serine kinase-driven protein switches. *Nat Struct Mol Biol* **28**, 762-770 (2021).

167. A. A. Vorobieva *et al.*, De novo design of transmembrane beta barrels. *Science* **371** (2021).

168. A. C. Hsieh, D. Ruggero, Targeting eukaryotic translation initiation factor 4E (eIF4E) in cancer. *Clin Cancer Res* **16**, 4914-4920 (2010).

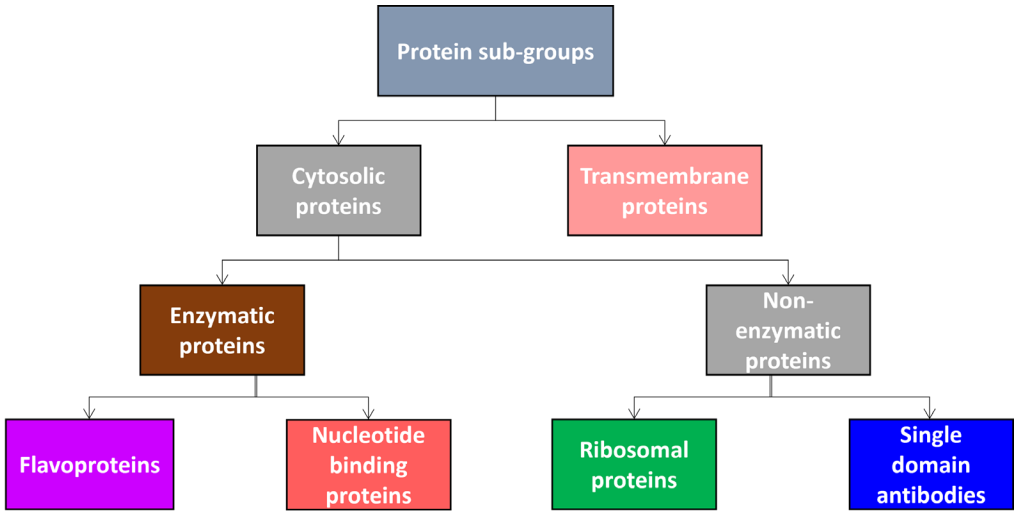
Appendix 1

Supplemental Material to Chapter 2



Appendix Figure 2.1 – Data mining and bioinformatic analysis pipeline. Protein sequence entries from the UniProtKB/Swiss-Prot database constitute the raw dataset. In-house developed Perl scripts are employed for data cleanup to remove fragments and duplicate entries from the raw dataset. Post cleanup, the final Swiss-Prot derived dataset is used for AAA analysis, consisting of counting the amino acid alphabet for each entry and binning them based on their alphabet size. Data standardization includes codon assignment correction and normalization for all amino acids.

Appendix Figure 2.2 – Distribution of prebiotic amino acids in small alphabet proteins. The amino acids have been color coded based on their side chain properties. Glycine is the first prebiotic amino acid to be included in the smallest alphabet proteins of size 3 AAA. Gradually other prebiotic amino acids are included as the AAA size increases. All ten prebiotic amino acids become incorporated while the alphabet expands from 3 AAA to 6 AAA, however all the prebiotic amino acids appear together for the first time in a 13 AAA protein.



Appendix Figure 2.3 – Hierarchical overlap between the different protein sub-groups. The first tier based on cytosolic vs. membrane localization of proteins. The second tier is based on enzymatic and non-enzymatic proteins.

Appendix Table 2.1 – Exclusion fraction (E_f) values of amino acids in proteins from the four domains of life- Archaea, Bacteria, Eukaryota and Virus. The amino acids (columns colored in grey) have been arranged in decreasing order of exclusion fraction trends observed for each domain of life. The E_f values have been colored using the coloring scheme from Table 2.2 (prebiotic amino acids are shown in orange).

Amino acid	Archaea (n=18442)	Amino acid	Bacteria (n=255063)	Amino acid	Eukaryota (n=171523)	Amino acid	Virus (n=15574)
W	0.53	W	0.58	W	0.47	W	0.46
C	0.25	C	0.25	C	0.14	C	0.17
H	0.08	H	0.053	H	0.098	H	0.097
Q	0.05	Y	0.036	Y	0.049	Q	0.040
F	0.019	F	0.017	M	0.038	Y	0.039
Y	0.019	Q	0.014	Q	0.037	F	0.034
N	0.019	N	0.014	D	0.027	E	0.028
D	0.012	D	0.010	F	0.025	N	0.023
E	0.0052	E	0.0071	E	0.025	D	0.022
P	0.0044	P	0.0068	N	0.024	K	0.022
T	0.0034	K	0.0061	K	0.018	P	0.019
K	0.0028	T	0.0023	I	0.012	G	0.011
S	0.0019	I	0.0022	P	0.010	A	0.0088
I	0.0018	R	0.0017	R	0.0066	I	0.0062
R	0.0012	G	0.0010	T	0.0057	R	0.0061
A	0.0012	S	0.00090	A	0.0056	M	0.0044
G	0.0010	A	0.00080	V	0.0051	T	0.0040
V	0.00075	M	0.00075	G	0.0041	V	0.0037
L	0.00060	V	0.00056	L	0.0027	S	0.0012
M	0.00011	L	0.00054	S	0.0012	L	0.0010

Appendix Table 2.2 – Exclusion fraction (E_f) values of methionine in proteins from the four domains of life.

Domain of life	Exclusion fraction value for methionine
Archaea	0.001
Bacteria	0.0007
Eukaryota	0.04
Virus	0.004

Appendix Table 2.3 – Exclusion fraction (E_f) values of amino acids in proteins from *Escherichia coli* and *Homo sapiens*. The amino acids have been arranged in decreasing order of E_f trends observed for amino acids in the overall Swiss-Prot derived dataset. The E_f values have been colored using the coloring scheme from Table 2.2 (prebiotic amino acids are shown in orange).

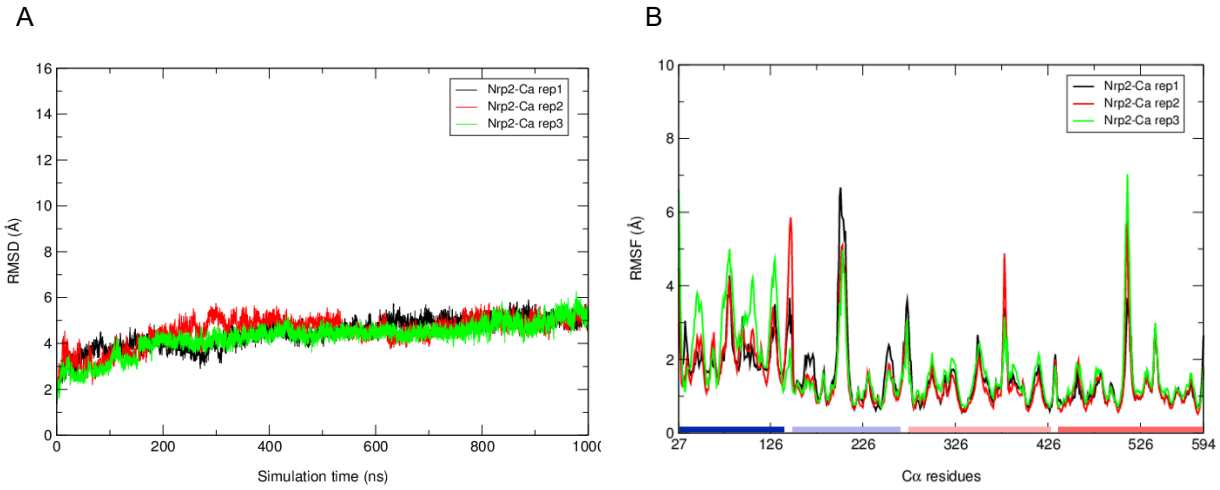
Amino acid	*E. coli* n=4371	*H. sapiens* n=20322
W	0.43	0.50
C	0.27	0.11
H	0.08	0.09
Y	0.05	0.08
Q	0.03	0.01
F	0.02	0.04
N	0.02	0.05
D	0.02	0.03
E	0.02	0.02
M	0.00	0.00
K	0.01	0.02
P	0.01	0.01
I	0.00	0.02
T	0.01	0.01
R	0.01	0.00
A	0.00	0.01
G	0.01	0.00
V	0.00	0.00
L	0.00	0.00
S	0.00	0.00

Appendix Table 2.4 – Small alphabet proteins (SAPs) population in different prokaryotic and eukaryotic genera. Left column indicates the AAA sizes and corresponding values demonstrate the total number of proteins that constitute the given AAA size.

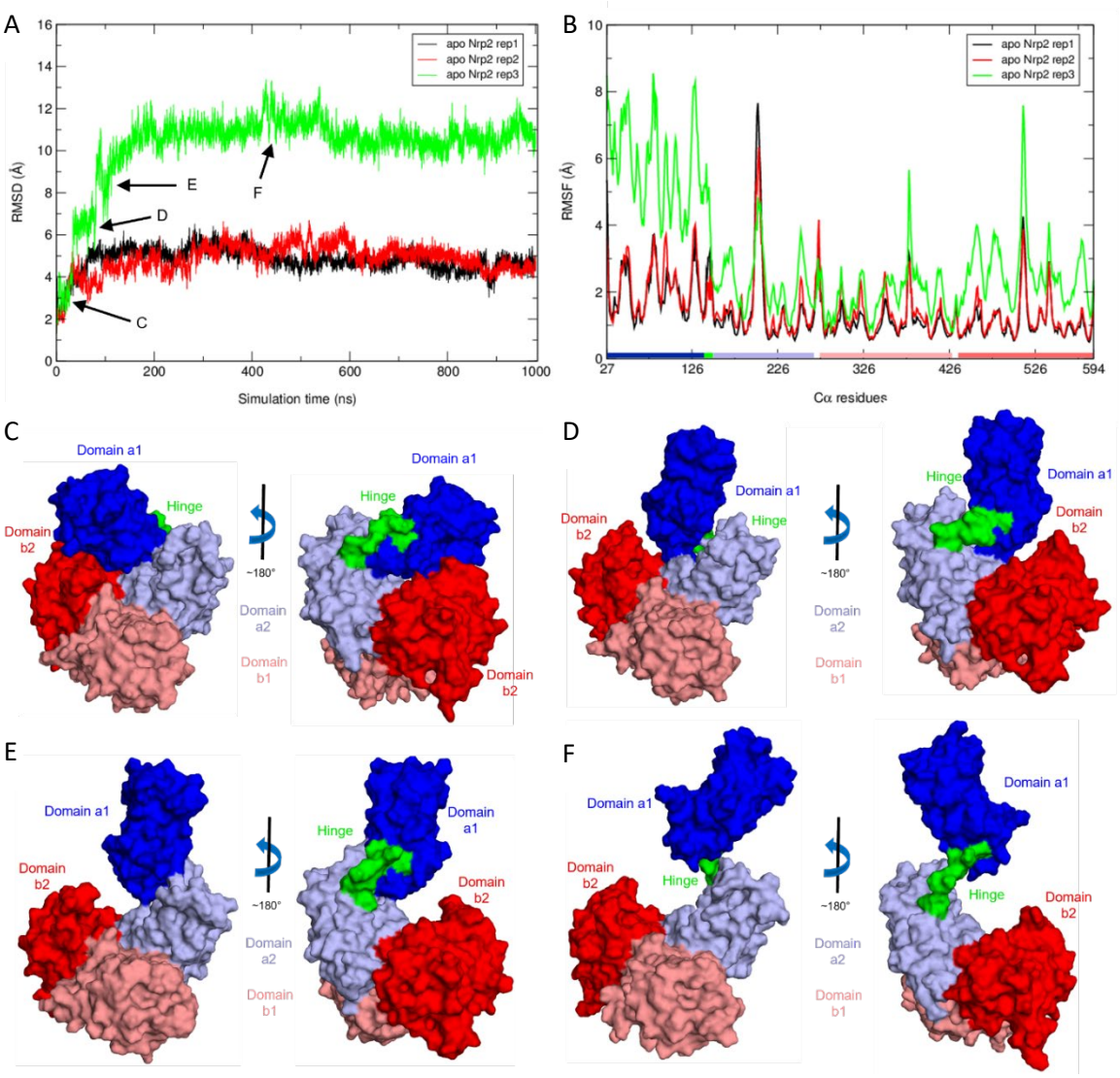
AAA	***E. coli*** n=4371	**Salmonella** n=5954	**Pseudomonas** n=9268	**Aeromonas** n=978	***H. sapiens*** n=20322	**Pan** n=678	**Rattus** n=7571	**Macaca** n=1401
20	3176	4873	6879	692	17917	585	6821	1197
19	776	831	1642	177	1578	47	524	121
18	238	148	505	73	464	29	160	56
17	98	63	168	25	158	11	37	12
16	43	24	37	7	78	1	14	5
15	13	5	26	2	30	3	6	0
14	13	4	2	2	35	0	6	3
13	4	5	2		25	0	1	3
12	7	0	7		18	0	1	1
11	2	1			9	1	0	2
10	1				6	1	0	0
9					1		1	0
8					0			1
7					2			
% SAPs	9.59	4.20	8.06	11.15	4.06	6.78	2.99	5.92

Appendix 2

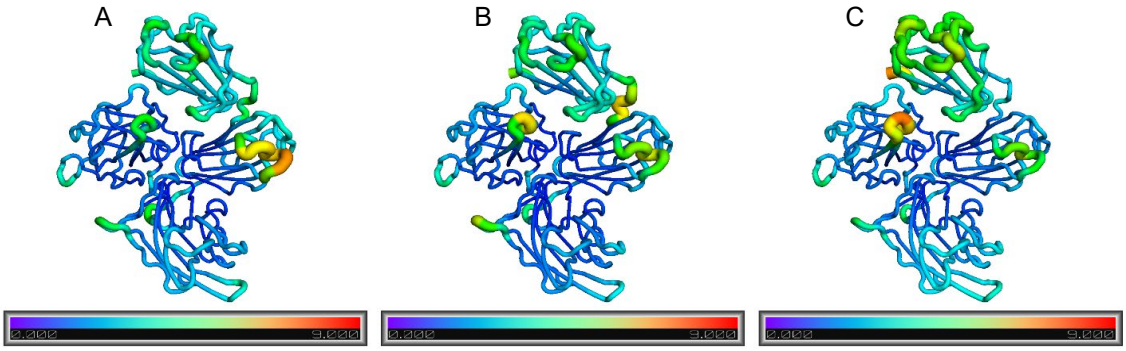
Supplemental Material to Chapter 3



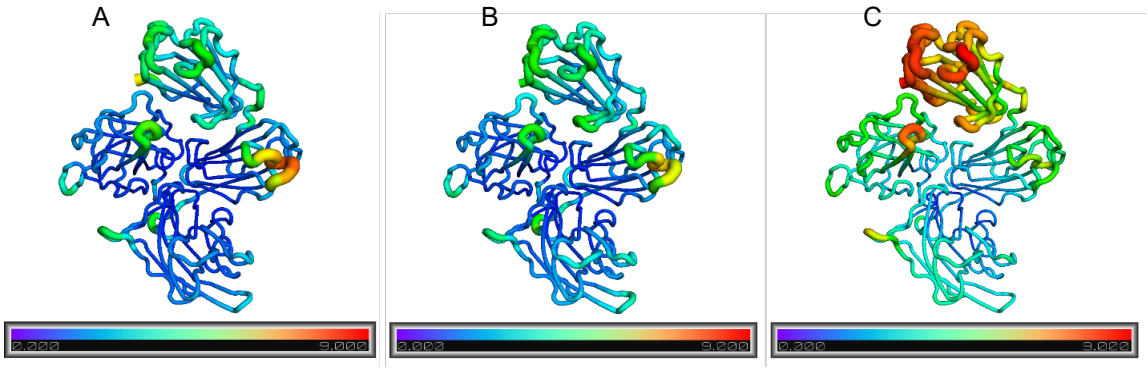
Appendix Figure 3.1 – RMSD and RMSF plots of the Nrp2-Ca^{2+} system replicates for 1 µs MD simulation. A. RMSD: Stable trajectories are observed for all replicates during the entire simulation period after a brief equilibration phase, with a mean RMSD of ~5 Å. B. RMSF: The color bars at the bottom denote the domains as described in Figure 3.1.A. All replicates show comparable RMSFs, where residues belonging to the inter-domain or intra-domain secondary structural loop elements demonstrate a comparatively higher flexibility (i.e., “spikes” in the RSMF plots).



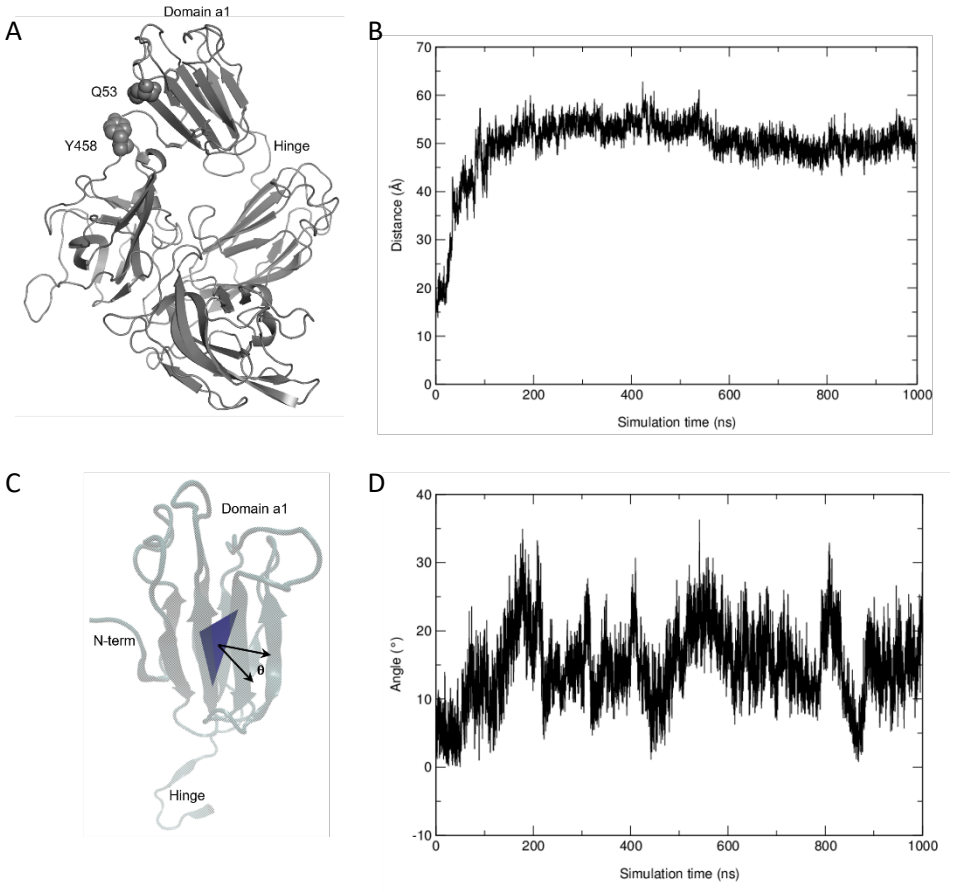
Appendix Figure 3.2 – RMSD and RMSF plots of the *apo* Nrp2 system replicates for 1 µs MD simulation. A. RMSD: The plots show higher RMSD during the initial equilibration relative to the starting structures and remain stable for the remainder of simulation. Replicate 3 shows a significantly higher RMSD than the other replicates due to a domain a1 opening motion. Along the trajectory of the observed opening of *apo* Nrp2, several intermediates which remain stable for >20ns are observed (indicated with arrows, letters correspond with the structures shown in panels below). B. RMSF: The color bars at the bottom denote the domains as described in Figure 3.1.A. C. Surface representation of *apo* Nrp2 15ns timepoint in simulation. Domain a1 forms interactions with both domain b2/a2, and the hinge region is compact. D. Surface representation of *apo* Nrp2 at 50ns timepoint in simulation. Domain a1 interactions with domain b2/a2 are broken compared to those observed at the 15ns timepoint. However, the hinge region remains compact (rigidified by intra-hinge H-bonds). E. Surface representation of *apo* Nrp2 simulation at 80ns timepoint in simulation. Domain a1 interacts with domain a2, and the hinge region remains rigidified. F. Surface representation of *apo* Nrp2 at 425ns timepoint in simulation. Domain a1 is displaced from the a1b1b2 core, and the hinge region is extended (intra-hinge H-bonds broken relative to previous panels).



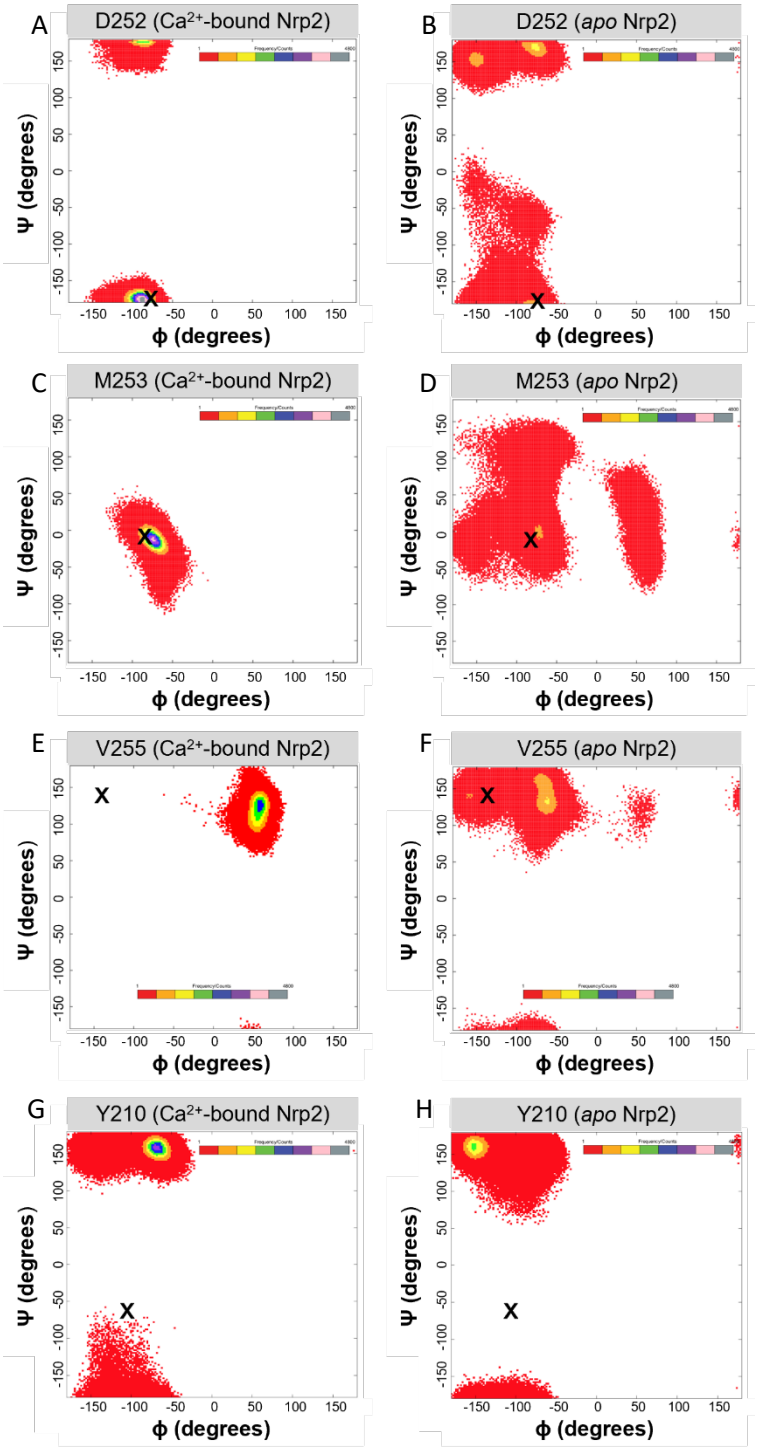
Appendix Figure 3.3 – Sausage plots displaying RMSF values reveal molecular dynamics "hotspots" of individual residues. Nrp2-Ca^{2+} replicates 1, 2, and 3 are shown in Panel A, B, and C, respectively. The RMSF values are plotted on the model structure of Nrp2-Ca^{2+} using Pymol. The scale (left to right) at the bottom represents the lowest RMSF of 0 Å in blue to highest RMSF of 9 Å in red.



Appendix Figure 3.4 – Sausage plots displaying RMSF values reveal molecular dynamics "hotspots" of individual residues. *apo* Nrp2 replicates 1, 2, and 3 are shown in panel A, B, and C, respectively. The RMSF values are plotted on the model structure of *apo* Nrp2 using Pymol. The scale (left to right) at the bottom represents the lowest RMSF of 0 Å in blue to highest RMSF of 9 Å in red. Replicate 3 explores the hinge-bending motion whereby the a1 domain "opens" away from the a2b1b2 core.

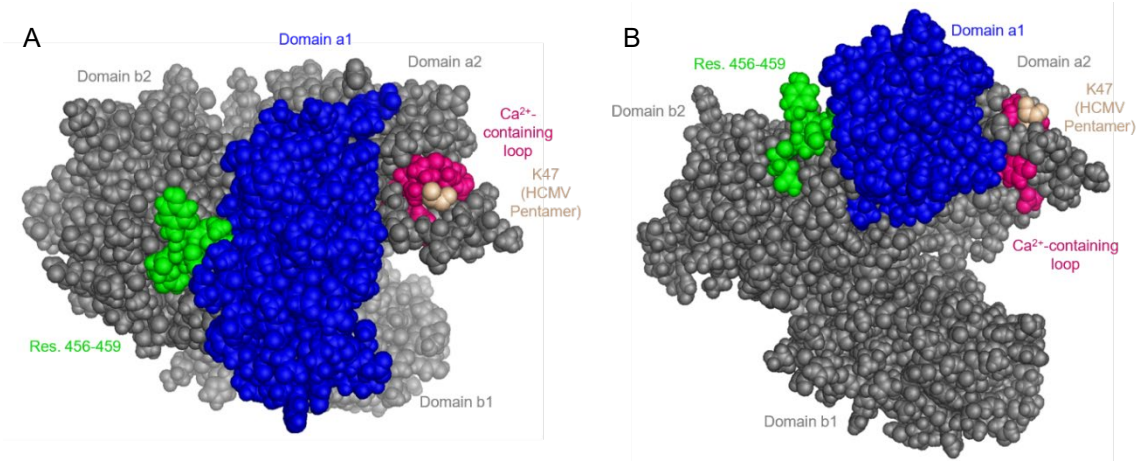


Appendix Figure 3.5 – Quantitative details about the a1 domain opening motion. A. Two spatially close amino acids from the a1 and b2 domains (Q53 and Y458 respectively) were selected to measure the distance domain a1 was displaced upon opening. B. Distance between Q53 and Y458 from b2 domain measured as a function of simulation time. C. The a1 domain "tilts" about an alternative axis in addition to the axis described in Figure 3.2. A plane was drawn through three amino acids (I88, Y121, F124) and a cross vector (perpendicular to this plane) was calculated for each frame for the duration of the simulation. The angle (θ) between each cross vector and the reference plane over time was determined, informing the degree to which the Nrp2 a1 domain "tilted" during the simulation. D. Plot of θ as a function of simulation time, demonstrating an initial "tilt" of approximately 15° as domain a1 opens, followed by oscillation as a1 remains in its open state.



Appendix Figure 3.6 – Representative Ramachandran plots for D252, M253, V255, and Y210 in Ca^{2+}-bound and *apo* Nrp2 simulations. Each Ramachandran plot depicts dihedral angles explored for the duration of the 1 μs MD simulation, where the frequency of a certain population being explored is expressed as heat maps. The black "x" on each plot reflects the dihedral angles observed in the Nrp2-HCMV pentamer complex PDB 7M22 (105). A. In the presence of bound Ca^{2+}, D252 exhibits a single population of dihedral angles which overlaps with the dihedrals observed in the Nrp2-HCMV pentamer complex. B. In the *apo* Nrp2 simulations, D252 adopts

multiple dihedral angle populations not observed in the Ca^{2+}-bound state of Nrp2 and has some population of dihedrals overlap with those observed in the Nrp2-HCMV pentamer complex. Compared to the Ca^{2+}-bound state, the *apo* state exhibits a 100% increase in the number of dihedral microstates (bins) occupied, and the percentage of non-overlapping Φ/Ψ angles is 80%. C. In the presence of bound Ca^{2+}, M253 exhibits a single population of dihedral angles which overlaps with the dihedrals observed in the Nrp2-HCMV pentamer complex. D. In the *apo* Nrp2 simulations, M253 adopts multiple dihedral angle populations not observed in the Ca^{2+}-bound state of Nrp2 and has some population of dihedrals overlap with those observed in the Nrp2-HCMV pentamer complex. Compared to the Ca^{2+}-bound state, the *apo* state exhibits a 104% increase in the number of dihedral microstates (bins) occupied, and the percentage of non-overlapping Φ/Ψ angles is 70%. E. In the presence of bound Ca^{2+}, V255 exhibits a single population of dihedral angles which does not overlap with the dihedrals observed in the Nrp2-HCMV pentamer complex. F. In the *apo* Nrp2 simulations, V255 adopts multiple dihedral angle populations not observed in the Ca^{2+}-bound state of Nrp2 and has some population of dihedrals overlap with those observed in the Nrp2-HCMV pentamer complex. Compared to the Ca^{2+}-bound state, the *apo* state exhibits a 102% increase in the number of dihedral microstates (bins) occupied, and the percentage of non-overlapping Φ/Ψ angles is 80%. G. Representative Ramachandran plot of Y210 in the Nrp2-Ca^{2+} simulation, where the amino acid position appears to be flexible and may sample the conformation observed in PDB 7M22 – however this is not the preferred conformation in our simulations. H. Representative Ramachandran plot of Y210 in the *apo* Nrp2 simulation.



Appendix Figure 3.7 – Sphere representation of the "closed" state of Nrp2 showing solvent accessibility of the interaction interface on Nrp2 accessible for binding. The Ca^{2+}-containing loop is shown in pink, the domain b2 loop (Residues 456-459) is shown in green, domain a1 is shown in blue, and the remaining regions of Nrp2 are shown in grey. K47 of protein UL128 in the HCMV pentamer is shown in tint and is superimposed into the cavity formed by the Ca^{2+}-containing loop (i.e., its binding site as shown in PDB 7M22). A. Top-down view of closed form of Nrp2 (snapshot at 200ns of Nrp2-Ca^{2+} simulation), and B. Side view of closed form of Nrp2.

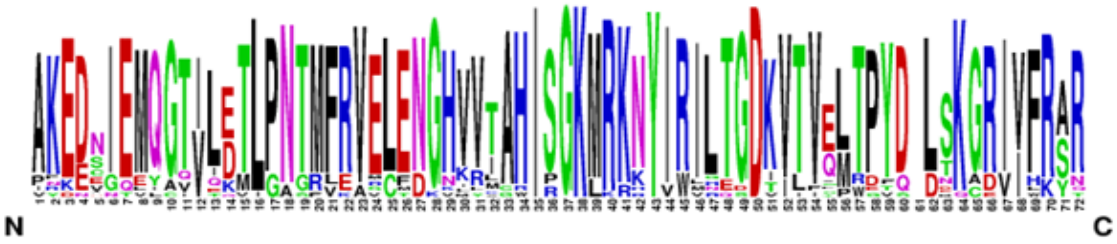
Appendix 3

Supplemental Material to Chapter 4

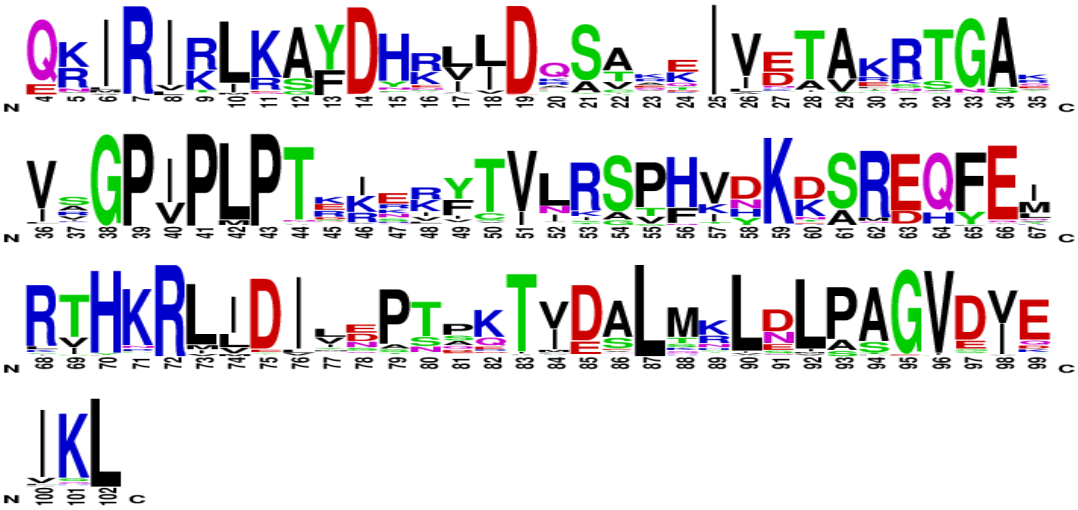
SUBS. MATRIX:	Aliphatic AILV	Aromatic FWY	Positive HKR	Negative DE	Nonpolar neutral GP	Sulphur containing CM	Polar neutral NQST	20 AMINO ACID ALPHABET
WILDTYPE CM:	AILV	F-Y	HKR	DE	GP	CM	NQS-	18A
16-ALPHABET-1:	-ILV	F-Y	HKR	DE	GP	CM	N-S-	RAP 16_1_CHEM, 16_1_CONS
16-ALPHABET-2:	AILV	--Y	HKR	DE	GP	-M	NQS-	RAP 16_2_CHEM, 16_2_CONS
14-ALPHABET-1:	-I-V	F-Y	HKR	DE	GP	CM	N---	RAP 14_1_CHEM, 14_1_CONS
14-ALPHABET-2:	AILV	--Y	HKR	-E	G-	-M	NQS-	RAP 14_2_CHEM, 14_2_CONS
12-ALPHABET-1:	-I-V	F--	HKR	DE	-P	CM	N---	RAP 12_1_CHEM, 12_1_CONS
12-ALPHABET-2:	AILV	--Y	-KR	-E	G-	-M	-QS-	RAP 12_2_CHEM, 12_2_CONS
10-ALPHABET-1:	-I-V	---	HKR	DE	-P	-M	N---	RAP 10_1_CHEM, 10_1_CONS
10-ALPHABET-2:	AILV	---	-KR	-E	--	-M	-QS-	RAP 10_2_CHEM, 10_2_CONS
9-ALPHABET-1:	-I-V	---	HKR	-E	-P	-M	N---	RAP 9_1_CHEM, 9_1_CONS
9-ALPHABET-2:	-ILV	---	-KR	-E	--	-M	-QS-	RAP 9_2_CHEM, 9_2_CONS
8-ALPHABET-1:	-ILV	---	-KR	-E	--	-M	N---	RAP 8_1_CHEM, 8_1_CONS

Appendix Figure 4.1 – Matrix employed to perform physicochemistry-based amino acid substitutions in chorismate mutase. Top panel shows the classification of amino acids based on their physicochemical properties. The amino acid alphabet of the wild-type chorismate mutase followed by the alphabets of the reduced alphabet variants is shown. Hyphens represent absent or removed amino acids. Similar matrices were employed to perform physicochemistry-based amino acid substitutions in the other two model proteins, IF1 and rpS10.

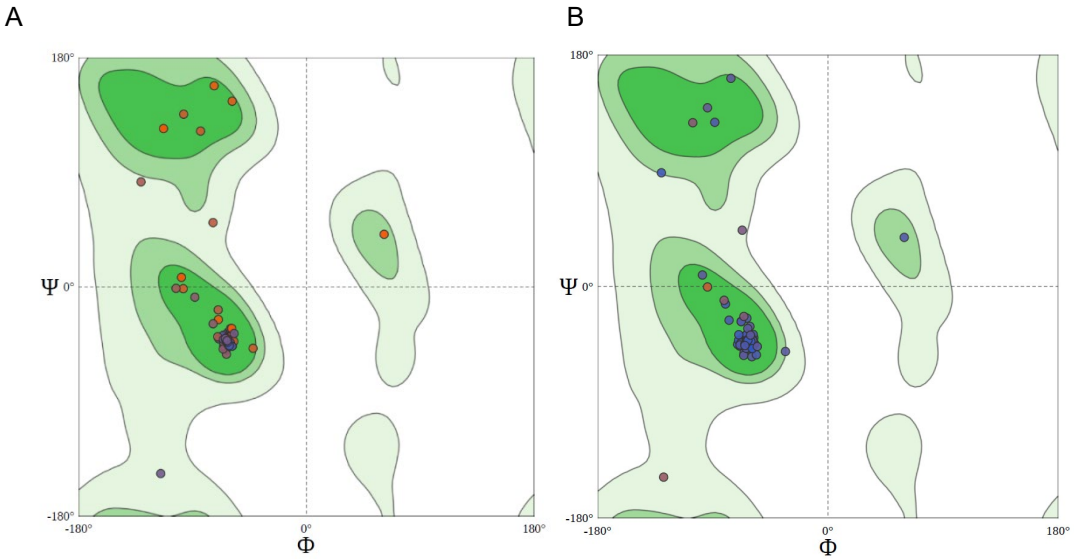
Appendix Figure 4.2 – WebLogo of the global evolutionary conservation profile of chorismate mutase. Weblogo is constructed using the multiple sequence alignment results of 150 chorismate mutase sequences generated by the ConSurf server (144, 147) for. Amino acids conservation trends at each position are identified by performing a multiple sequence alignment of homologous proteins from other organisms. For any given position, the amino acid on top shows the amino acid present at that position in the current protein sequence, followed by others in decreasing order of conservation trends observed in all other species and organisms.



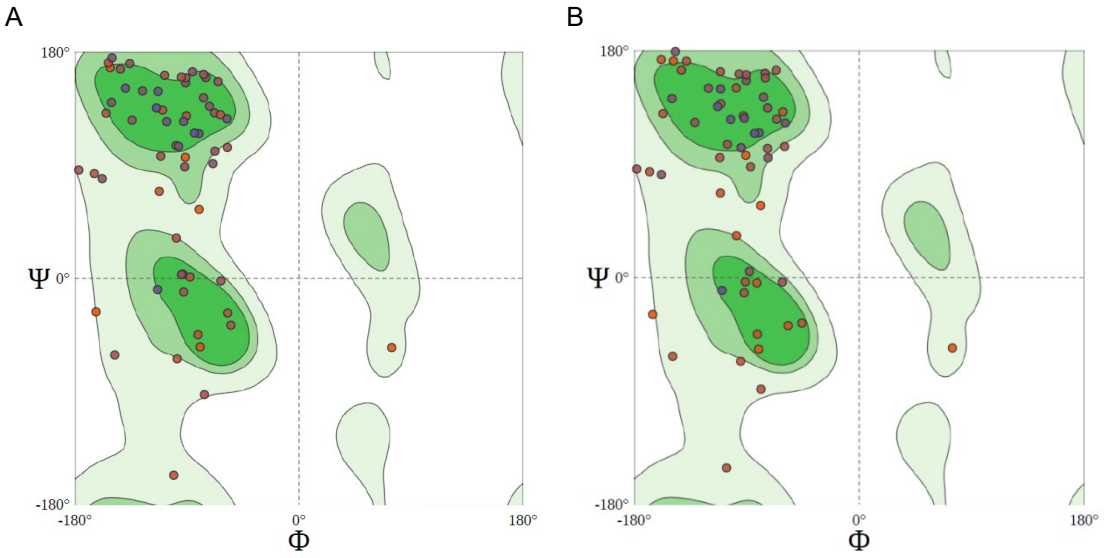
Appendix Figure 4.3 – WebLogo of the global evolutionary conservation profile of IF1. WebLogo is constructed using the multiple sequence alignment results of 150 IF1 sequences generated by the ConSurf server (144, 147). Amino acids conservation trends at each position are identified by performing a multiple sequence alignment of homologous proteins from other organisms. For any given position, the amino acid on top shows the amino acid present at that position in the current protein sequence, followed by others in decreasing order of conservation trends observed in all other species and organisms.



Appendix Figure 4.4 – WebLogo of the global evolutionary conservation profile of rpS10. Weblogo is constructed using the multiple sequence alignment results of 150 rps10 sequences generated by the ConSurf server (144, 147). Amino acids conservation trends at each position are identified by performing a multiple sequence alignment of homologous proteins from other organisms. For any given position, the amino acid on top shows the amino acid present at that position in the current protein sequence, followed by others in decreasing order of conservation trends observed in all other species and organisms.

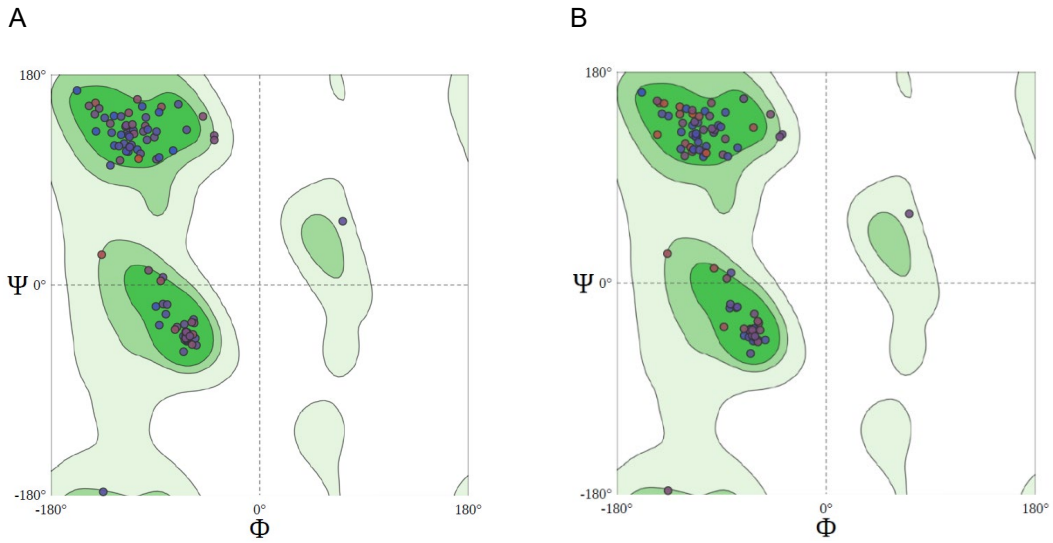


Appendix Figure 4.5 – Ramachandran plot analysis for the wt-CM and RA-CM variant. A. Ramachandran plot for the wild-type chorismate mutase. The x- and y-axis show the phi (Φ) and psi (Ψ) torsional angles respectively, and the individual amino acids are shown as circles. B. Ramachandran plot for the SWISS-MODEL generated structure for the 16_1_chem variant. The percentages of allowed Φ-Ψ dihedral angle combinations are 98% and 97% for the wt-CM and the 16_1_chem variant, respectively. The Ramachandran analysis results for other RA-CM variants are provided in Appendix Table 4.1 A.

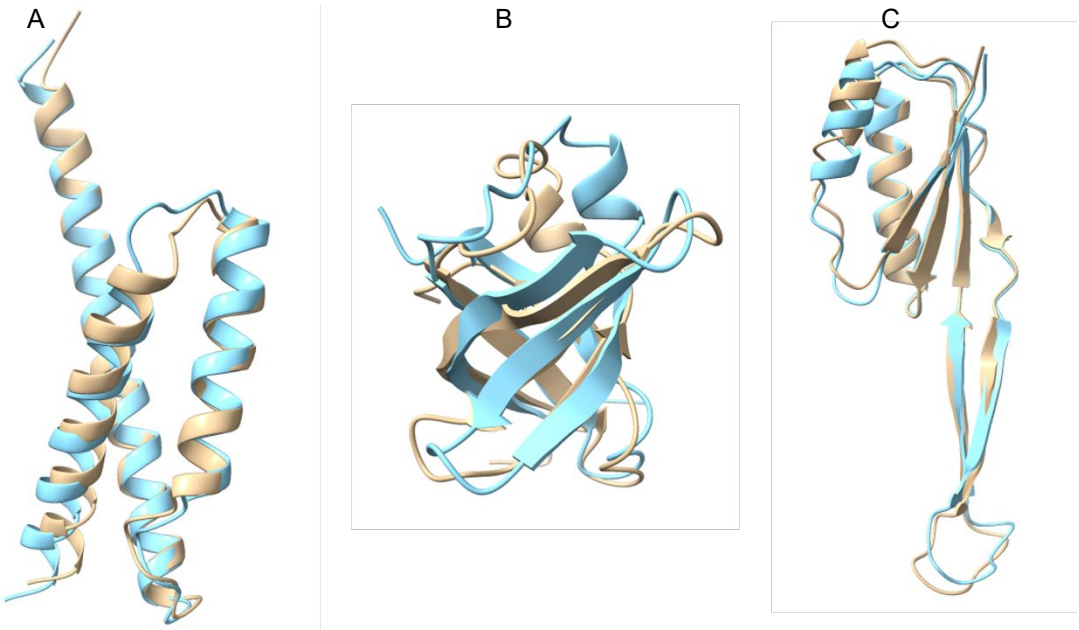


Appendix Figure 4.6 – Ramachandran plot analysis for the wt-IF1 and RA-IF1 variant. A. Ramachandran plot for the wild-type IF1. The x- and y-axis show the phi (Φ) and psi (Ψ) torsional angles respectively, and the individual amino acids are shown as circles. B. Ramachandran plot for the SWISS-MODEL generated structure for the 17_chem variant. The percentages of allowed Φ-Ψ dihedral angle combinations are 79% and 76% for the wt-IF1 and the 17_chem variant,

respectively. The Ramachandran analysis results for other RA-IF1 variants are provided in Appendix Table 4.1 B.



Appendix Figure 4.7 – Ramachandran plot analysis for the wt-rpS10 and RA-rpS10 variant. A. Ramachandran plot for the wild-type rpS10. The x- and y-axis show the phi (Φ) and psi (Ψ) torsional angles respectively, and the individual amino acids are shown as circles. B. Ramachandran plot for the SWISS-MODEL generated structure for the 16_chem variant. The percentages of allowed Φ-Ψ dihedral angle combinations are 98% and 95% for the wt-rpS10 and the 17_chem variant, respectively. The Ramachandran analysis results for other RA-rpS10 variants are provided in Appendix Table 4.1 C.



Appendix Figure 4.8 – AlphaFold structure predictions for the 8A alphabet variants of the three model proteins. A. Chorismate mutase, B. IF1, C. rpS10. The structure prediction is performed using the AlphaFold plugin in ChimeraX. The structures predicted by AlphaFold are shown in cyan, and the homology modeled structure of the 8A variants are shown in golden. For each protein, both the structures have been superimposed to calculate the RMSDs, which is 1 Å for all the three proteins.

Appendix Table 4.1 – Ramachandran analysis for the RA-variants of the three test proteins. The percentages of favourable Φ-Ψ dihedral angle combinations for the amino acids in the RAP variants of the test proteins chorismate mutase, IF1, and rpS10 are shown in tables A, B, and C, respectively.

A

RA-CM variants	Ramachandran favoured Φ-Ψ angles
16AAA_1_chem	97%
16AAA_1_cons	98%
16AAA_2_chem	98%
16AAA_2_cons	98%
14AAA_1_chem	97%
14AAA_1_cons	98%
14AAA_2_chem	97%
14AAA_2_cons	98%
12AAA_1_chem	96%
12AAA_1_cons	95%
12AAA_2_chem	94%
12AAA_2_cons	95%
10AAA_1_chem	94%
10AAA_1_cons	97%
10AAA_2_chem	94%
10AAA_2_cons	96%
9AAA_1_chem	98%
9AAA_1_cons	94%
8AAA_1_chem	94%
8AAA_1_cons	95%

B

RA-IF1 variants	Ramachandran favoured Φ-Ψ angles
17AAA_chem	79%
17AAA_cons	78%
16AAA_1_chem	78%
16AAA_1_cons	79%
16AAA_2_chem	78%
16AAA_2_cons	77%
14AAA_1_chem	77%
14AAA_1_cons	75%
14AAA_2_chem	76%
14AAA_2_cons	75%
12AAA_1_chem	75%
12AAA_1_cons	74%
12AAA_2_chem	74%
12AAA_2_cons	73%
10AAA_1_chem	77%
10AAA_1_cons	74%
10AAA_2_chem	75%
10AAA_2_cons	75%
8AAA_1_chem	76%
8AAA_1_cons	74%

C

RA-rps10 variants	Ramachandran favoured Φ-Ψ angles
16AAA_chem	95%
16AAA_cons	94%
14AAA_1_chem	96%
14AAA_1_cons	95%
14AAA_2_chem	94%
14AAA_2_cons	95%
12AAA_1_chem	93%
12AAA_1_cons	94%
12AAA_2_chem	94%
12AAA_2_cons	95%
10AAA_1_chem	94%
10AAA_1_cons	93%
10AAA_2_chem	94%
10AAA_2_cons	96%
9AAA_1_chem	96%
9AAA_1_cons	96%
9AAA_2_chem	94%
9AAA_2_cons	95%
8AAA_1_chem	94%
8AAA_1_cons	93%

Appendix Table 4.2 – Scoring and ranking of RA-chorismate mutase variants. The left column shows the alphabet sizes of the variants, and the top row shows the computational analysis techniques employed to study the dynamic and structural properties of the variants. Performance of all variants is compared to the wild-type protein for each analysis and scores are assigned based on the scoring algorithm shown in Figure 4.2. The final scores of the variant are used to select the best variants for each alphabet size.

AAA size	RMSD	RMSF	Network shape compared to wt	How many of top10 BC residues from wt are preserved	Residue heatmaps compared to wt	Major motions compared to wt	Folding pathway compared to wt	Residue contacts compared to wt	AlphaFold prediction	Final score
wt_CM_18	4.0	1.9				4		155		
16_1_chem	7.2	2.7	similar	4	similar	2	similar	142	similar	1
16_1_cons	4.8	2.3	similar	6	similar	4	similar	151	similar	9
16_2_chem	7.6	3.0	different	3	different	3	similar	140	similar	-3
16_2_cons	5.2	2.4	similar	6	similar	4	similar	148	similar	9
14_1_chem	6.2	2.9	similar	5	similar	3	similar	138	similar	3
14_1_cons	5.3	2.6	similar	6	different	4	similar	143	similar	7
14_2_chem	7.1	3.1	similar	4	similar	2	different	140	similar	-1
14_2_cons	4.8	2.2	similar	7	similar	3	similar	146	similar	9
12_1_chem	4.9	2.4	similar	5	similar	4	different	149	similar	7
12_1_cons	5.5	2.8	different	3	similar	2	different	141	similar	-3
12_2_chem	5.8	2.7	similar	3	different	2	similar	141	similar	-1
12_2_cons	5.1	2.1	similar	6	similar	3	similar	143	similar	9
10_1_chem	6.6	3.0	different	2	different	1	different	143	similar	-5
10_1_cons	6.1	2.6	similar	4	similar	3	similar	147	similar	9
10_2_chem	6.9	2.9	similar	3	different	2	different	142	similar	-3
10_2_cons	5.7	2.4	similar	7	similar	4	similar	150	similar	9
9_1_chem	4.6	1.9	similar	5	different	3	similar	148	similar	5
9_1_cons	5.3	2.3	similar	2	similar	2	similar	140	similar	1
9_2_chem	5.9	2.8	different	3	different	1	similar	142	similar	-3
9_2_cons	4.8	2.5	similar	5	similar	3	similar	148	similar	9
8_1_chem	5.5	2.7	similar	5	similar	2	similar	151	similar	
8_1_cons	5.3	2.6	similar	6	similar	3	similar	146	similar	

Appendix Table 4.3 – Scoring and ranking of RA-IF1 variants. The left column shows the alphabet sizes of the variants, and the top row shows the computational analysis techniques employed to study the dynamic and structural properties of the variants. Performance of all variants is compared to the wild-type protein for each analysis and scores are assigned based on the scoring algorithm shown in Figure 4.2. The final scores of the variant are used to select the best variants for each alphabet size.

AAA size	RMSD	RMSF	Network shape compared to wt	How many of top10 BC residues from wt are preserved	Residue heatmaps compared to wt	Major motions compared to wt	Folding pathway compared to wt	Residue contacts compared to wt	AlphaFold prediction	Final score
wt_IF1_18	2.0	1.2				4		178		
17_1_chem	3.2	2.9	similar	4	different	2	similar	170	similar	-1
17_1_cons	2.8	2.0	different	5	different	3	similar	174	similar	5
16_1_chem	2.5	1.3	similar	7	similar	4	similar	176	similar	9
16_1_cons	3.4	1.8	similar	4	similar	3	different	160	similar	1
16_2_chem	2.5	1.4	similar	6	similar	2	similar	175	similar	7
16_2_cons	3.7	1.3	similar	4	similar	3	similar	164	similar	4
14_1_chem	2.1	1.5	similar	7	different	3	similar	174	similar	7
14_1_cons	2.8	1.9	similar	3	similar	3	similar	170	similar	3
14_2_chem	2.5	1.7	similar	5	similar	2	similar	175	similar	7
14_2_cons	3.0	2.0	different	4	similar	3	different	170	similar	-1
12_1_chem	1.9	1.2	similar	5	different	3	similar	173	similar	7
12_1_cons	2.4	1.8	different	4	similar	2	different	166	similar	0
12_2_chem	2.4	1.7	similar	4	similar	4	similar	169	similar	9
12_2_cons	3.2	2.3	similar	3	similar	3	similar	164	similar	1
10_1_chem	2.8	1.8	similar	6	different	4	different	170	similar	5
10_1_cons	3.3	2.9	similar	3	similar	2	similar	169	similar	3
10_2_chem	2.6	1.5	different	5	similar	3	similar	169	similar	7
10_2_cons	2.6	1.6	similar	3	different	3	different	161	similar	0
8_1_chem	2.8	1.7	different	4	similar	3	similar	173	similar	
8_1_cons	2.5	1.3	similar	3	similar	3	similar	176	similar	

Appendix Table 4.4 – Scoring and ranking of RA-rpS10 variants. The left column shows the alphabet sizes of the variants, and the top row shows the computational analysis techniques employed to study the dynamic and structural properties of the variants. Performance of all variants is compared to the wild-type protein for each analysis and scores are assigned based on the scoring algorithm shown in Figure 4.2. The final scores of the variant are used to select the best variants for each alphabet size.

AAA size	RMSD	RMSF	Network shape compared to wt	How many of top10 BC residues from wt are preserved	Residue heatmaps compared to wt	Major motions compared to wt	Folding pathway compared to wt	Residue contacts compared to wt	AlphaFold prediction	Final score
wt_rpsJ_18	5.9	2.3				3		61		
16_1_chem	5.4	2.0	similar	5	similar	3	similar	56	similar	7
16_1_cons	5.7	2.2	similar	6	different	3	different	60	similar	6
14_1_chem	6.1	2.0	different	4	similar	3	similar	54	similar	2
14_1_cons	5.8	2.1	similar	7	similar	3	different	57	similar	6
14_2_chem	5.9	1.9	similar	6	similar	4	different	58	similar	5
14_2_cons	6.4	2.5	different	6	different	3	similar	57	similar	3
12_1_chem	5.8	2.2	similar	7	similar	3	different	60	similar	7
12_1_cons	6.6	2.2	similar	8	similar	3	different	61	similar	7
12_2_chem	6.7	2.1	similar	6	similar	3	different	59	similar	2
12_2_cons	7.4	2.6	similar	5	similar	3	different	60	similar	3
10_1_chem	6.2	2.1	similar	5	similar	4	similar	59	similar	6
10_1_cons	6.0	1.9	similar	5	similar	3	similar	60	similar	9
10_2_chem	6.3	2.0	similar	6	different	3	similar	58	similar	4
10_2_cons	6.3	2.0	different	5	similar	4	different	58	similar	0
9_1_chem	7.0	2.5	different	7	different	3	similar	54	similar	1
9_1_cons	6.1	2.0	similar	6	similar	3	similar	56	similar	9
9_2_chem	5.8	2.2	similar	5	different	3	similar	59	similar	5
9_2_cons	6.9	2.9	similar	4	different	4	similar	59	similar	1
8_1_chem	7.1	2.8	similar	5	similar	4	different	57	similar	
8_1_cons	6.4	2.5	similar	4	different	4	similar	60	similar	

Appendix Table 4.5 – *Bx(n)* values of the top ten residues in wt-cm and the 16A variants. Blue columns show the top ten amino acids and their *Bx(n)* values for wt-cm. Columns 3-10 show the *Bx(n)* values of top ten residues in the 16A variants. Highlighted in green are the residues that have high betweenness centrality in wt-cm and the respective variant.

Res.ID	wt-CM	Res.ID	16-1-chem	Res.ID	16-1-cons	Res.ID	16-2-chem	Res.ID	16-2-cons
22	72.2	16	84.1	9	83.6	15	101.2	16	90.0
23	103.8	23	93.4	20	85.9	16	73.3	20	114.1
24	108.3	24	84.2	23	89.6	20	124.4	22	74.9
35	96.2	28	77.3	24	122.8	23	77.9	23	76.0
37	98.4	39	80.6	26	78.3	25	73.9	24	99.5
47	68.9	47	113.0	35	93.5	35	74.0	35	82.4
72	80.0	50	81.4	47	116.1	72	78.3	70	85.4
75	131.2	71	100.2	58	77.0	76	75.1	72	78.5
79	112.7	72	87.9	72	111.5	77	78.0	75	76.4
86	78.5	73	71.7	86	86.5	78	79.7	76	97.2

Appendix Table 4.6 – *Bx(n)* values of the top ten residues in wt-IF1 and the 16A variants. Blue columns show the top ten amino acids and their *Bx(n)* values for wt-IF1. Columns 3-10 show the *Bx(n)* values of top ten residues in the 16A variants. The residues highlighted in green are the ones that have high betweenness centrality in wt-IF1 and the respective variant.

Res.ID	wt-IF1	Res.ID	16-1-chem	Res.ID	16-1-cons	Res.ID	16-2-chem	Res.ID	16-2-cons
8	163.5	8	189.9	12	202.5	8	307.5	12	263.9
12	226.1	12	220.2	19	208.8	12	353.5	20	196.3
20	165.3	21	250.5	21	209.6	31	220.2	23	173.3
21	235.8	23	177.5	35	271.8	34	207.8	25	169.8
31	183.3	31	195.4	52	333.3	43	291.1	52	323.3
34	168.4	36	177.0	53	229.0	52	366.7	53	246.2
36	167.7	52	304.6	54	297.8	53	303.9	54	271.8
52	266.5	54	200.2	55	222.1	54	288.1	55	169.8
54	162.3	56	235.4	66	246.8	66	231.7	66	226.3
71	236.2	66	225.7	69	235.8	69	185.8	67	221.0

Appendix Table 4.7 – *Bx(n)* values of the top ten residues in wt-rpS10 and the 12A variants. Blue columns show the top ten amino acids and their *Bx(n)* values for wt-rpS10. Columns 3-10 show the *Bx(n)* values of top ten residues in the 12A variants. The residues highlighted in green are the ones that have high betweenness centrality in wt-rpS10 and the respective variant.

Res.ID	wt-rpS10	Res.ID	12-1-chem	Res.ID	12-1-cons	Res.ID	12-2-chem	Res.ID	12-2-cons
5	1047.6	5	1006.2	5	513.8	5	1032.9	5	1100.3
7	1182.1	7	789.1	7	998.1	7	865.6	6	773.7
22	583.2	44	692.0	22	713.8	46	805.5	7	599.9
44	969.2	46	860.4	44	610.3	61	648.6	10	721.6
46	867.3	47	555.4	46	957.6	64	1244.5	15	512.4
48	749.2	61	572.9	49	537.5	66	754.6	22	498.0
62	403.1	64	793.0	64	890.4	67	488.6	42	825.0
66	736.6	66	1311.3	66	720.7	68	758.9	46	976.7
67	715.9	67	498.6	67	563.3	69	540.2	48	679.5
68	630.2	68	745.0	68	961.8	84	495.5	73	490.5