Mapping Functionally Important and Stabilising Regions in Biotherapeutic Proteins, using NMR and Mutagenesis

A thesis submitted to University College London for the degree of DOCTOR OF PHILOSOPHY

By

Mark-Adam Walter Kellerman

University College London

Department of Biochemical Engineering

Declaration

I, Mark-Adam Walter Kellerman confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Candidate:

Date: 10/11/2022

UCL Research Paper Declaration Form

I acknowledge permission of ACS Molecular Pharmaceutics to include in chapter 3 of this thesis portions of the publication (subjected to academic peer review) titled "NMR Reveals Functionally Relevant Thermally Induced Structural Changes within the Native Ensemble of G-CSF" (published on 10/08/2022). The contributions from all authors are as follows; I acquired and processed the data and wrote the manuscript, Dr Almeida helped with data analysis, Dr Rudd helped with data acquisition, interpretation and manuscript drafting and Prof. Dalby and Dr Matejtschuk helped with data interpretation and manuscript drafting.

Candidate:

Date: 10/11/2022

Senior Author:

Date: 18/11/2022

Abstract

Structure-function relationships in proteins refer to a trade-off between stability and bioactivity, moulded by evolution of the molecule. Identifying which protein amino acid residues jeopardise global or local stability for the benefit of bioactivity would reveal residues pivotal to this structure-function trade-off. Demonstrated here is the use of varied-temperature ^{15}N-^{1}H heteronuclear single quantum coherence (HSQC) nuclear magnetic resonance (NMR) spectroscopy to probe the microenvironment and dynamics of residues in granulocyte-colony stimulating factor (G-CSF). This experimental approach was also used to investigate (de-) stabilising mechanisms of action for previously studied excipients with G-CSF. Combining NMR with *in silico* analysis revealed four structural clusters that are subject to localised conformational changes (some of which are key to bioactivity) or partial unfolding prior to global unfolding at higher temperatures. Mechanisms by which excipients influence these important structural changes and implement their own structural clusters reflects their impact on stability and function. This approach was leveraged for semi-rational mutant/formulation design. These mutants were tested for fitness with respect to thermostability and functionality. The Mutants P65V and E45Q were constructed to elicit mutation-excipient interactions, and presented the largest impact on the respective fitness. Hence, this study proposes an approach to profile residues, thus highlighting their roles in stability and bioactivity while exposing potential mutation-excipient interactions. This permits a semi-rational protein engineering approach to optimise desirable protein fitness characteristics.

Impact Statement

With biotherapeutics comprising such a large proportion of pharmaceutical sales, a deeper understanding of protein stability and structure function-relationships could provide substantial help to bioengineering strategies. Various biophysical methods can be used to assess the stability of proteins, including their conformational stability determined from changes in intrinsic tryptophan fluorescence during thermal or chemical denaturation. Colloidal stability can also be determined from aggregation onset temperatures, zeta potentials and B_{22} values (Roberts, Das and Sahin, 2011; Saito *et al.*, 2012; Thiagarajan *et al.*, 2016). Nevertheless, a blind spot exists when the stability of a protein is assessed simply by optical methods during thermal or chemical denaturation. They do not acquire any information regarding the changes that individual residues experience prior to and during denaturation or aggregation. They also ignore changes in the distribution of conformations within the native ensemble prior to denaturation that may be functionally relevant. Thermal fluctuations of residues within the native ensemble are also considered to be an important aspect of the mechanisms that lead to aggregation behaviours (Codina *et al.*, 2019). This study will address this lack of resolution on residues to provide a semi-rational basis for constructing mutants and formulations. Therefore, this would provide protein engineers with more economical approaches to creating improved biotherapeutics, which would stimulate the biobetters market.

Furthermore, machine learning approaches have shown that the thermal dependence of fluorescence spectra under only native conditions, are sufficient to predict their subsequent melting temperatures (Zhang *et al*., 2021), highlighting the underlying importance of native ensemble dynamics in defining the pathways to global conformational unfolding. In addition, this underlines the predictive capabilities of *in silico* methods and their potential for improving protein engineering capabilities. This study will further interrogate the predictive capabilities of *in silico* methods by cross-validating conclusion from biophysical techniques with computational analysis. These comparisons will yield information regarding structure-function mechanisms and highlight which computational tools (or combinations thereof) best reflect experimental observations. Moreover, the per-residue resolution from NMR and *in silico* modelling permits a near 1:1 comparison, providing a basis for more informed decisions regarding which residues to mutate.

Recent advancements in our understanding of protein evolution has opened up new avenues for mutagenesis. The combination of NMR and computational modelling in this study aims to provide a more mechanistic view for evolutional phenomena such as epistasis, thus progressing approaches to achieve mutational additivity. Maximising the effects of each mutation is pivotal to the biobetter market because this class of therapeutics need to be an improvement over the original biologic. Moreover, improving the predictive capabilities of *in silico* modelling would provide protein engineers with much more powerful screening techniques, thus enhancing the potential for higher quality biotherapeutics at a lower cost. Therefore, presented in this study is a workflow elucidating high-resolution protein characteristics to stimulate the biobetters market and yield better predictive capabilities for protein stability and function: an important addition to the current field of protein structure prediction.

Acknowledgments

I would first like to thank my supervisors Professor Paul Dalby (UCL) and Dr Paul Matejtschuk (MHRA) for drawing up the project and giving me the opportunity to be apart of such a great field of research.

I would like to thank the department of Biochemical Engineering at UCL for their assistance in many aspects of the degree journey. Many thanks also to the EPSRC CDT for their financial support. I am very appreciative for the time and knowledge that Dr Cheng Zhang and Dr Akash Pandya dedicated to helping me understand various experimental and *in silico* modelling aspects. My MRes and undergraduate students Sara Ouadhi and Jinhui Kim provided this project with

very valuable data. I fully appreciated the time and effort they put into the work and enjoyed having them in the lab.

I would also like to thank the team at MHRA for their tremendous support in the many aspects of my research. Dr Adrian Bristow was very kind to donation the G-CSF expression system. Chinwe Duru, Kiran Malik and Ernest Ezeajughi provided me with support in training and were a pleasure to have around. Dr Chris Ball provided training with initial protein expression and bioassays, the helpful impact of which cannot be overstated. Both were very accommodating in their lab. Many thanks to Dr Tim and Theresa for the time they gave me with their NMR facilities and their help with data processing.

Finally, I would like to dedicate this thesis to thank the many family members that supported me throughout the project. Hard times were always made brighter with them.

Contents

Figures

Tables

Definitions:

G-CSF - Granulocyte colony-stimulating factor

∑Δδ - Cumulative change in chemical shift

PI – Peak intensity

90th percentile of PI - The upper 10% threshold for the total distribution of PI data.

Percentage increase in PI - The percentage increase in PI (typically between 295K and 305K)

Linear/non-linear "peak trajectory" - Linearity of peak maxima movement across the HSQC spectra

Linear/non-linear "∑Δδ-temperature relationship"- Linearity of ∑Δδ vs temperature lines

Sub-cluster – Cluster of residues from the top 15 residues for percentage increase in PI

Structural cluster – Cluster of residues considered significant from combined ∑Δδ, PI and percentage increase in PI

SASA - Solvent accessible surface area

MD – Molecular Dynamics

NMR - Nuclear magnetic resonance

T_m - Thermal unfolding temperature

CCM - Cross-correlation matrices

$\Delta\delta_H/\Delta T$ - Temperature coefficients

1 Introduction

1.1 Challenges Posed to the Biotherapeutic Market

Biotherapeutics account for roughly one third of all new drugs currently awaiting FDA approval or in clinical trials. These therapeutics encompass vaccines, monoclonal antibodies, cytokines, growth factors and other similar products (Select USA, 2019). A total of 210 biotechnology products are reported to be on the market with 180 more in the pipeline (Mitragotri, Burke and Langer, 2014). The birth of the biotherapeutics industry in the 1970's saw three major research milestones. These were the development of monoclonal antibody technologies, human genome sequencing and recombinant DNA technologies (Pavlou, 2003). Before the advent of this industry, small molecule drugs (SMDs) dominated the pharmaceutical market. These drugs are mostly organic compounds such as corticosteroids. SMDs mainly differ from biologics in their size, where they have a molecular weight of <1kDa unlike the larger size of >1kDa for biotherapeutics. This small size of SMDs makes them highly diffusible across membranes (Olivera, Danese and Peyrin-Biroulet, 2017). The attractive virtue of biotherapeutics is their high specificity, potency and low toxicity in comparison to SMDs (Mitragotri, Burke and Langer, 2014). Such benefits have advanced antibodies to the most widely used and clinically successful class of protein therapeutics in cancer (Jemal *et al.*, 2011).

With all of this enthusiasm for biotherapeutics, it is important to remember that low stability is an inherent disadvantage that comes with proteins. This can lead to undesirable effects such as immunogenicity upon patient delivery and reduced shelf-life. Immunogenicity alongside pharmacodynamic and pharmacokinetic failure can also lead to loss of patient response to treatment. Other disadvantages of biotherapeutic proteins include poor tissue penetration (as a result of size) and the high cost of manufacturing (Aldeghaither, Smaglo and Weiner, 2015; Ding, Hart and De Cruz, 2016). Consequently, there is a demand for addressing these disadvantages.

1.2 Biobetters can Display Better Characteristics than their Originator Product

With biologics expected to comprise 50% of the pharmaceutical industry by 2022, it is important to address means of relieving the economic burden of this class of therapeutics on healthcare systems (Dixit, 2021). In turn, this could help close the gap between countries wealthy enough to access these therapeutics and those that are not. Generic chemical drugs (i.e. copycat drugs) are deemed identical to their originator drugs since it can be proved that the effect of both drug types will be the same on patients. The copying and marketing of biological substances by other manufactures gave rise to 'biosimilars' which, as with generics, function to lower the burden of drug prices. Nevertheless, the problem that biosimilars pose to the generics market is that biologics are synthesised in living cells and since no two independently developed cell lines are considered identical, biologics cannot be completely copied. Hence, the term biosimilar refers to the fact that the generic version of a biologic is not exactly the same but similar (Declerck, 2007). However, biosimilars must be amino acid sequence identical to the originator.

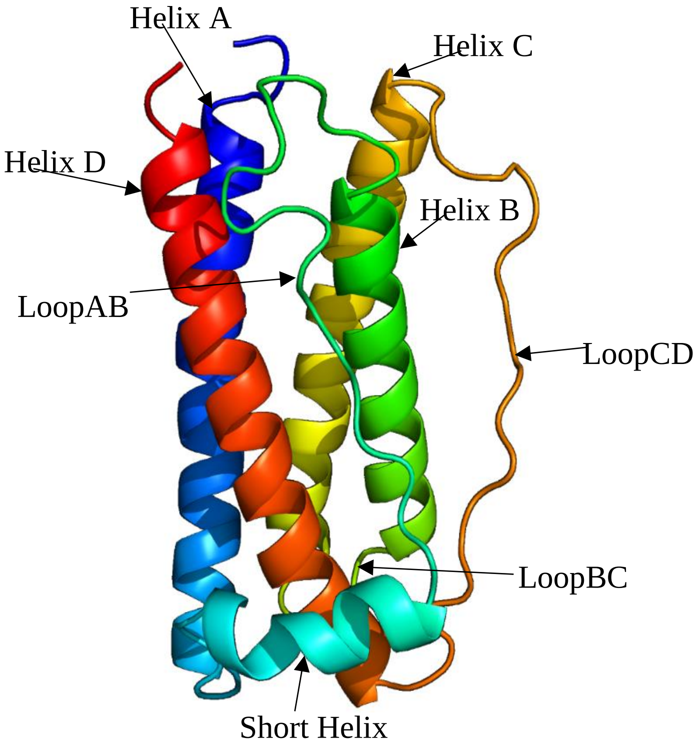
The EMA and FDA have approved 58 and 29 biosimilars respectively as of January 2021 (Honavar, 2021), with some reaching half the price of the innovator biologic (Gota *et al*., 2016; Danese, Bonovas, and Peyrin-Biroulet, 2017). However, many factors can hinder the market penetration of biosimilars. These include gaps in knowledge regarding biosimilars and their safety by physicians and patients, unclear evaluations of "biosimilarity" meaning, unclear pathways of biosimilar development and patent extension (evergreening) of innovator products (Calvo, 2021; Dixit, 2021; I-MAK, 2021).

Biobetters aim to alleviate some of these hurdles that biosimilars face because they are highly differentiated and display superior characteristics to the originator biologic. As a result, biobetters do not need to wait for originator patents to expire. These therapeutics have their own hurdles to face, such as a longer (and more expensive) process to market and establishment/demonstration of biosuperiority (Dixit, 2021). Nevertheless, with benefits to the patient such as increased potency, longer half-life and reduced immunogenicity, the potential of this market cannot be overlooked (Honavar, 2021). There are many ways to engineer biotherapeutic proteins to create biobetters. These include mutagenesis, PEGylation, antibody-drug conjugation and humanisation (Chen and Arnold, 1993; Piedmonte, and Treuheit, 2008; Scheinfeld, 2003). Knowledge of which approach (or combination thereof) is effective for a particular biologic requires rationality. Mutagenesis can prove to be an enormous challenge because the number of mutational possibilities grows exponentially with protein size.

Granulocyte colony-stimulating factor (G-CSF) is available in recombinant human (rh-) form as a drug against severe neutropenia. RhG-CSF is marketed in 2 forms, namely filgrastim (under the trade name Neupogen®, Amgen) which is purified from *E. coli* and Granocyte® which is purified from Chinese hamster ovary cells. Neulasta®, a PEGylated version of filgrastim, is an example of a biobetter (Molineux, 2004; Strohl, 2015). G-CSF will be used as a model protein in this study to investigate areas significant to stability and bioactivity.

1.3 G-CSF Mechanism of Action

Neutrophils are phagocytes that function to initiate and maintain inflammation at sites of infection for anti-bacterial immunity. The principal immune-regulatory protein (cytokine) in neutrophil development and function is G-CSF, which is induced by inflammatory stimuli such as interleukin-1β and tumour necrosis factor-alpha (Panopoulos and Watowich, 2008). Human G-CSF (HG-CSF) has a molecular weight of 19.6 kDa, larger than the 18.8 kDa of murine G-CSF, which is due to the O-glycosylation of HG-CSF at T133. Additionally, HG-CSF has 5 cysteine residues, from which two intermolecular disulphide bridges form between 4 residues, important to its activity, with the remaining cysteine free and partially solvent accessible (Souza *et al.*, 1986; Arvedson and Giffin, 2012). Substituting the free Cys-17 residue for an Alanine has exhibited improved bioactivity (Jiang *et al.*, 2011; Liu and Jiang, 2010). G-CSF has four alpha-helices with a linker region formed by the peptide sequence between helices A and B. This region passes in front of helix D and forms a 3_{10}-helix and an alpha-helix (Figure 1).



MTPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCATYKLCHPEELVLLGHSLGIPW
APLSSCPSQALQLAGCLSQLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIW
QQMEELGMAPALQPTQGAMPAFASAFQRRAGGVLVASHLQSFLEVSYRVLRHLAQP.

Figure 1. The Structure of G-CSF. PDB:2D9Q of G-CSF structure and its amino acid sequence.

The principal immune-regulatory cytokine in neutrophil development and function is Granulocyte colony-stimulating factor (G-CSF; Panopoulos and Watowich, 2008). The multiple cells that express G-CSF range from endothelial cells to bone marrow stromal cells, while G-CSF receptors (G-CSFRs) are expressed on both haematopoietic and non-haematopoietic cells (Roberts *et al*., 1997). G-CSF (Figure 2 in green) binds to G-CSFR (orange) in a 2:2

stoichiometry through cross-over interactions between the G-CSFR Ig-like domain and the neighbouring G-CSF. G-CSFR activation is suggested to cause neutrophil mobilization by inducing haematopoietic cells to generate secondary signals that act in *trans* to stimulate neutrophil release from bone marrow (Liu, Poursine-Laurent and Link, 2000; Semerad *et al.*, 2002). Residues from two sites on G-CSF are involved in receptor binding. These sites are the major site/site II (residues K16, G19, Q20, R22, K23, L108, D109 and D112) and the minor site/site III (residues Y39, L41, E46, V48, L49, S53, F144 and R147; Tamada *et al*., 2006).

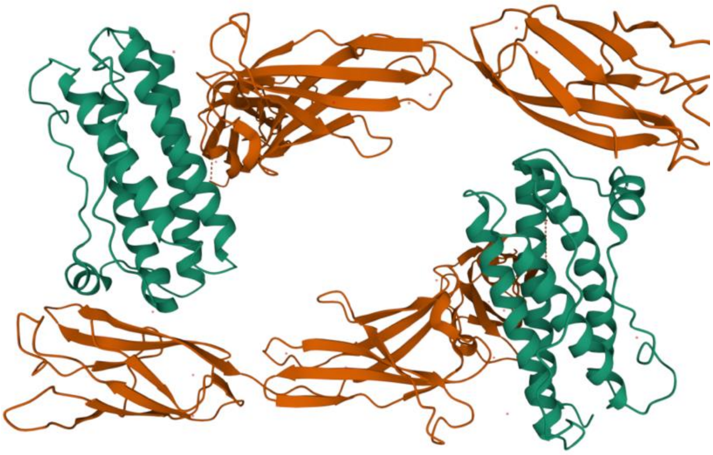


Figure 2. G-CSF in Complex with its Receptor. Crystal structure of two G-CSF molecules (green) in complex two G-CSF receptors (orange)
Taken from Tamada et al., 2006.

1.4 Interactions that Guide Protein Folding

1.4.1 Proteins Fold toward a Lower Energy Native State

Proteins play significant roles in a plethora of functions; facilitating transmembrane migration and transmission of information from DNA to RNA just to name a few. Considering this, proteins can be described as the workhorse of cellular machinery (Keskin *et al.*, 2008). Amino acids polymerize via peptide bonds to form the primary (polypeptide) structure of proteins. A feature of the peptide bond which confines the polypeptide backbone to certain conformations is its planar structure. This feature restricts bond angles around Cα - N and C - Cα to phi and psi angles respectively. Local folding of the polypeptide chain form a higher-order structure (HOS) known as the protein's secondary structure which typically consists of α-helix, β-sheet, β-turn and random coil. These secondary structure elements assemble to comprise a second HOS called the tertiary structure. The tertiary and primary structure of a protein endow it with signature conformation and physicochemical properties which determine its function. A third HOS, quaternary structure, exists and refers to the interaction within multi-subunit proteins.

The folding of a protein to its native state is thought to follow the "funnel-shaped folding energy landscape" where unfolded conformations are at the top of the funnel and narrowing of the funnel signifies fewer possible folding conformations (Figure 3). Protein folding can come at an entropic cost due to increased structural order. However, a compensating enthalpic and entropic force can arise from the summation of weak non-covalent interactions that form via the polypeptide backbone and amino acid side chains (Berkowitz and Houde, 2014). These interactions include hydrogen bonding, the hydrophobic effect and van der Waals (VDW). Also stabilising the HOS are disulphide bonds which can occur within a polypeptide chain, two different chains within the same protein or even between different proteins. Proteins are believed to contain small modules capable of folding quasi-independently due to local interactions guiding the protein to a folded structure by folding in a single cooperative step. These units are called "foldons" (Panchenko, Luthey-Schulten and Wolynes, 1996). Nevertheless, proteins can encounter states where they are not as conformationally stable as their native state. In this case they are trapped (indicated as smaller funnels in Figure 3) due to the activation energy needed to return them to a more stable form. Folded proteins may not have reached their optimal fold (final native state) and so this near optimal state is commonly referred to as 'molten globular state'.

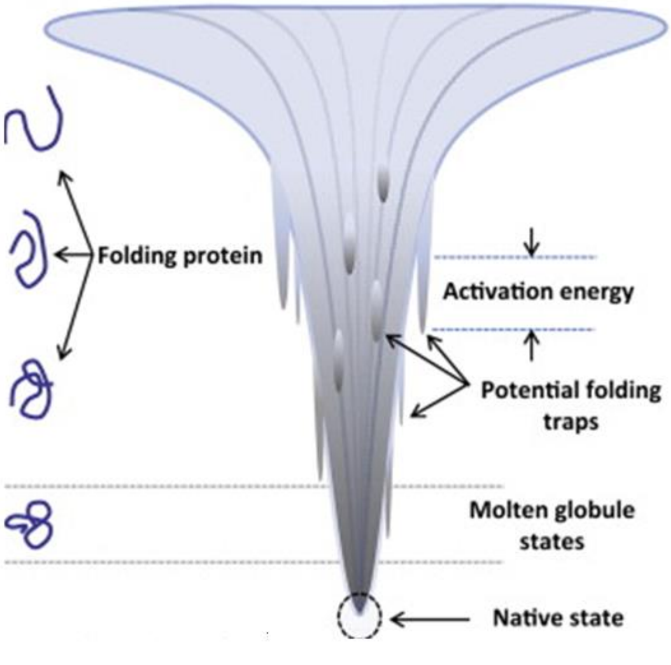


Figure 3. Funnel-Shaped Folding Energy Landscape. Completely unfolded proteins begin at the top of this funnel. They become more folded lower down the energy landscape funnel toward the (fully folded) native state, illustrated here as the bottom of the funnel.
(Taken from Berkowitz and Houde, 2014)

1.4.2 Protein Solvation Induces Protein Folding

Protein solvation by water plays an important role in packing and stabilisation via hydrogen bond networks and electrostatic interaction screening (Levy and Onuchic, 2004). The latter is particularly important to hydrophilic effects driving protein folding, where high densities of surface charges impose a large energetic penalty for folding, which can be lowered by increasing ionic strength (Arbely *et al*., 2010). Dehydration of water from hydrophobic cores can also play a significant role in protein folding (Levy and Onuchic, 2004). Solvents surrounding protein can be characterised as forming either the hydration shell or bulk solvent (Chen, Weber and Harrison, 2008), while individually bound water can hydrogen bond with buried polar or charged residues (Roberts and Mancera, 2008). The hydration shell is the closest layer to the protein surface and is divided into two non-overlapping shells (Figure 4). The first shell (red) has a radius of 2.75 Å, hydrogen bonding with the protein and the second shell (blue) has a radius of 3.65 Å, forming clathrate-like water via VDW interactions with nonpolar residues (Chen, Weber and Harrison, 2008; Parui and Jana, 2019). MD simulations suggest the hydration shell can regulate protein dynamics through its surface interactions, permitting only low-frequency motions by adding springs between the shell and protein surface (Majumdar, Kim and Na, 2020). Considering these characteristics of water can reveal the behaviours that excipients exhibit in solution with protein, as will be shown later in this study.

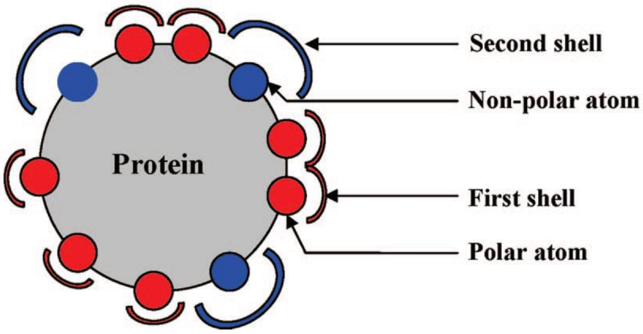


Figure 4. Protein Hydration Shell Structure. Two layers of the hydration shell surrounding the protein model. Shells are coloured as a red arc (first shell) and blue arc (second shell) and residues are coloured as a red circle (polar) and blue circle (non-polar).
(Taken from Chen, Weber and Harrison, 2008).

1.5 Protein Instability

The stabilisation of a folded protein is only marginal considering that the weak non-covalent interactions can be broken as a function of time (Berkowitz and Houde, 2014). The disruption of these interactions permits the display of a conformational dynamic where, in solution, a protein exists as an ensemble of different conformations as opposed to a single one (Bellissent-Funel *et* *al.*, 2016). Protein instability can generally be classed as chemical or physical, where chemical instability involves formation of new chemical entities. On the other hand, physical instability is where the chemical status is not changed and the physical state is. There are also cases where chemical instability can lead to physical instability. Two main types of physical instability are denaturation and aggregation.

1.5.1 Physically-Induced Protein Instability

1.5.1.2 Denaturation

Denaturation refers to a loss of 3D structure, which both temperature and pressure are able to induce (Manning *et al.*, 2010). The Gibbs-Helmholtz equation describes the Gibbs free energy change (ΔG) upon protein unfolding, which can occur at high temperatures and low temperatures (cold-denaturation) (Baldwin, 1986). Thermal denaturation of proteins occurs when temperature is raised to a point where the solvent and protein configurational entropy outweighs enthalpic stabilisation (Bellissent-Funel *et al.*, 2016). The primary reason for cold denaturation is solvent penetration into the hydrophobic core of proteins (Ramírez-Sarmiento *et al.*, 2013). Increasing the pressure of a system decreases the system's volume. Therefore, the main reason given to pressure-induced protein denaturation is water penetration, where the protein becomes more solvated by water molecules (Panick *et al.*, 1998; Bellissent-Funel *et al.*, 2016). Of note, chaotropes such as urea can chemically denature proteins by binding them and reducing their chemical potential (Manning *et al.*, 2010).

1.5.1.3 Characteristics and Mechanisms of Aggregation

The capability of aggregation to cause adverse side effects, such as immune responses, deems it an important problem for drug manufacturers to tackle. Another challenge for manufacturers is being able to better predict aggregation propensities of proteins at an early stage of development to maximize their likelihood of making it to market (Jain *et al.*, 2017). Protein aggregates can be classified by several characteristics of protein-protein interactions. This includes bond type (weak noncovalent versus strong covalent interactions), reversibility, size (small soluble oligomers versus larger insoluble oligomers) and protein conformation (predominantly native versus non-native structure) (Mahler *et al.*, 2009). Philo *et al* suggested five general aggregation mechanisms; association of native monomers, aggregation of conformationally altered monomers, nucleation-controlled aggregation, aggregation of chemically-modified monomers and surface-induced aggregation (Philo and Arakawa, 2009). Identifying the cause of aggregation can therefore prove to be a tough endeavour due to the many pathways through which it can occur. There are many potential causes of aggregation from protein manufacture to patient delivery which can be brought about by exposing the protein to damaging conditions such as freezing, solid-liquid interfaces and extreme pH.

The rate-limiting step in protein aggregation is thought to be the level of transient reactive species that are partially unfolded but similar to the native state (shown as M* in the scheme below). Therefore, a protein's intrinsic conformational stability is considered to be a highly significant factor in aggregation (Manning *et al.*, 2010). The osmotic second virial coefficient (B_{22}) is a measure of unfavourable solution behaviours that arise from two-body (protein-protein) interactions:

$$B_{22} = \frac{2\pi}{M^2}\int_0^\infty r^2\left(1 - e^{-u(r)/kT}\right)dr \; \boldsymbol{Eqn.1}$$

Protein molecular weight is represented here by M, intermolecular separation distance by r, interaction potential by u(r), Boltzman constant by k and absolute temperature by T. A negative B_{22} value indicates attractive forces between proteins where protein-protein interactions are favoured over protein-solvent interactions. A positive B_{22} value indicates the opposite. B_{22} can, therefore, reflect protein colloidal stability where a more positive B_{22} value would indicate higher colloidal stability. VDW forces and electrostatic interactions are a major contribution to colloidal interactions. Chi *et al* described aggregation as being a collective effect of colloidal and conformational stability (measured by the free energy of unfolding: ΔG_{unf}). RhG-CSF is known to aggregate rapidly under physiological solution conditions (pH 7 phosphate-buffered saline and 37^{o}C) unlike in conditions of pH 3.5 and low ionic strength solutions where aggregation was minimal. At this low pH, rhG-CSF is highly positively charged which causes high colloidal stability due to strong electrostatic repulsion between monomers (Chi *et al.*, 2003; Israelachvili, 2011). It was also observed that the aggregation behaviour of rhG-CSF drastically varied across the pH range of 2 to 7 (Chi *et al.*, 2003). A similar phenomenon was seen by Chakroun *et al* who observed a pH and ionic-strength dependent likelihood of the antibody fragment A33Fab to aggregate. They suggested that conformational stability was a better predictor of aggregation kinetics of this fragment at higher temperatures because protein aggregation would be dominated by the unfolded state (Chakroun *et al.*, 2016). Robinson *et al* also found this to be true for G-CSF where measures of Tagg (temperature at which aggregates are first detected) was only a useful measure when the proportion of unfolded protein was high (Robinson *et al.*, 2018). An aggregation mechanism put forward by Krishnan *et al* is depicted in the scheme below:

$M \leftrightarrow M^*$

$2M^* \rightarrow M_2$

$M^* + M_x \rightarrow M_{x+1}$

In this scheme, M represents the structurally undisturbed G-CSF monomer, whereas M* represents the structurally perturbed monomer (expanded transition state) which has a 15% increased surface in comparison to M. M* can react irreversibly with itself, to form the dimer represented as M_2, or with existing aggregates, M_x, to from the larger aggregates shown as M_{x+1}. Therefore, it is understandable that in solution conditions where B_{22} values are low (such as physiological conditions) and the colloidal stability is low, aggregation is rate-limited by conformational stability (Raso *et al.*, 2005). Elucidation of this partially unfolded state, the exposed residues and important conformational changes, could be very valuable to protein engineering strategies.

1.6 Chemically-Induced Protein Instability

1.6.1 Deamidation

Hydrolysis describes a chemical reaction where bonds are broken by the addition of water. Hydrolytic reactions of concern to protein stability include deamidation, proteolysis, β elimination and racemization. The most common of these reactions is deamidation which occurs primarily at asparagine residues, modifying it to form aspartate, but can also affect glutamine residues at a much lower rate. The aspartate side chain is acidic, unlike the neutral side chain of asparagine, and will therefore change the net charge of the protein and protein sequence. In addition, at a neutral to basic pH, a structurally isomeric 'isoAsparagine' residue can dominate the population, which in turn can alter the peptide backbone by introducing an extra methyl group. Moreover, deamidation has been shown to reduce conformational stability thus making it a promoter of aggregation. Influencing the rate of deamidation are protein sequence, temperature and solution pH. Faster rates have been observed in peptides where the 'N+1' residue (the adjacent amino acid on the C-terminal side of asparagine or glutamine) is small. The protein HOS also influences rates, as more rapid deamidation is seen in flexible loop regions in contrast to structurally rigid regions (Nilsson, Driscoll and Raleigh, 2009; Topp *et al.*, 2010).

1.6.2 β elimination and racemization

β elimination and racemization are similar reactions in their initial step which involves abstraction of a proton from the α carbon of an amino acid. The product of racemization can be either a D or L form of the original amino acid. On the other hand, β elimination can result in the expulsion of a persulphide anion which has been implicated in intermolecular crosslink formation, thus making this reaction more closely linked to aggregation. High temperature and pH have been reported to increase the rate of both reactions (Volkin and Klibanov, 1987; Topp *et al.*, 2010).

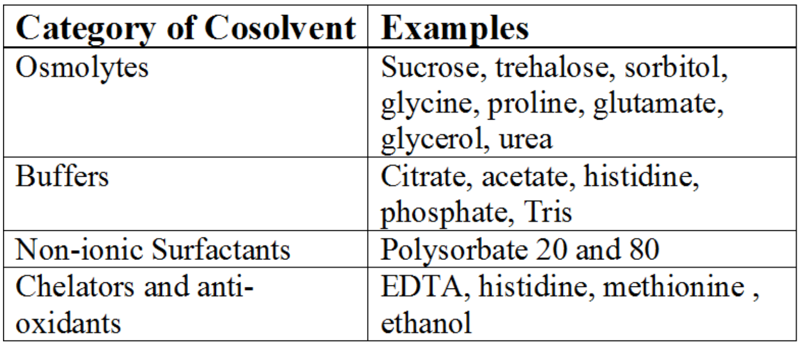
1.6.3 Oxidation

Although oxidation can occur at several amino acid side chains, those most susceptible possess sulphur or aromatic containing side chains. The three general oxidation pathways are metal-catalysed oxidation (MCO), photooxidation and free-radical cascade oxidation. Binding of some transition metals like Fe, Mn and Cu to proteins are redox-active and can confer oxidative damage by generating reactive oxidation species (ROS). This emphasizes vigilance even more so for manufacturers since these metals be leached from storage containers, manufacturing equipment and also from excipients used in formulation (particularly sugars and polymers) (Topp *et al.*, 2010). A proposed mechanism for promotion of aggregation by oxidation is that the hydrophobicity of buried protein cores can be significantly reduced by methionine side chain oxidation (Uversky *et al.*, 2002). Strategies to minimise oxidation of biopharmaceuticals includes reducing the vial headspace or addition of chelating agents to reduce the MCO (Manning *et al.*, 2010).

1.7 Protein-Engineering Methods to Increase Stability

1.7.1 Formulation with Cosolvent

Formulation with cosolvents is employed to stabilise the pharmaceutical protein against many of the aforementioned physical and chemical instability pressures. Cosolvents can range in molecular weight, examples of which are organised in Table 1 based on the effects they have and detailed below (Kamerzell *et al.*, 2011).



Category of Cosolvent	Examples
Osmolytes	Sucrose, trehalose, sorbitol, glycine, proline, glutamate, glycerol, urea
Buffers	Citrate, acetate, histidine, phosphate, Tris
Non-ionic Surfactants	Polysorbate 20 and 80
Chelators and anti-oxidants	EDTA, histidine, methionine , ethanol

Table 1. Types of Cosolvent. Shown here are the different categories of excipients with corresponding examples represented in the right-hand column.
(Adapted from Kamerzell *et al.*, 2011)

Osmolytes are naturally occurring small organic molecules capable of increasing thermodynamic stability of intracellular proteins and in some cases inducing cooperative protein folding into functional native-species (Kumar, 2009; Manning *et al.*, 2010). The osmophobic nature of the protein backbone enables osmolytes to force their way into the core of folded proteins. Osmolytes accomplish this by increasing the surface free energy of water and by being excluded from the water-protein interface, which shifts the system towards minimization of this interface area (i.e. towards a more folded state) (Arakawa and Timasheff, 1985).

Non-ionic surfactants primarily stabilise proteins by outcompeting and blocking them from hydrophobic surfaces such as air-water interfaces. The exact stabilising mechanism of action by which many non-ionic surfactants bind and stabilise proteins is not fully understood. However, studies showing reduced adsorption of G-CSF to polyvinyl chloride by addition of HSA and poloxamer 407 suggests that this was due to a decreased fraction of G-CSF available for surface binding (Wang, Udeani and Johnston, 1995). Another study suggests that the ratio of polysorbate to protein results in different protein-excipient interactions, hinting at the possibility that the conformation of the excipient determines its interaction with protein (Deechongkit *et al.*, 2009).

The use of buffering agents in formulation stems from the dependence of the chemical integrity of amino acid residues on pH. They can therefore be used to optimize solution pH for conformational and colloidal stability (Kamerzell *et al.*, 2011). Nevertheless, this is not their sole use because at high protein concentrations the protein can provide most of the buffering capacity instead (Gokarn *et al.*, 2008). Alternate stabilisation mechanisms for buffers have been reported. These include free radical scavenging abilities (antioxidant effects), altering protein-surfactant binding characteristics, direct binding at low buffer concentrations and optimizing electrostatic interactions within a native protein by rejecting conformers that lead to repulsive charge-charge conformations (Spassov, Karshikoff and Ladenstein, 1994; Min Won *et al.*, 1998; Porasuphatana *et al.*, 2001).

1.7.1.1 Strength of Protein-Excipient Interaction does not Dictate Stabilising Effect of the Excipient

Protein in solution can sometimes have higher affinity for cosolvents than for water. In this scenario, termed “preferential interaction”, the cosolvent can overcome the hydration shell and interact with the protein surface. The opposite scenario is “preferential exclusion”, where the protein surface has a higher affinity for water in the hydration shell. The addition of any cosolvent lowers the activity of water through translational and rotational perturbations (Timasheff, 2002). Immobilisation of water or cosolvent to the protein surface also incurs translational and rotational entropy loss (Irudayam, and Henchman, 2009). Therefore, there must be a favourable enthalpy-entropy compensation for protein-cosolvent interaction (Du *et al*., 2016).

Excipients are typically added at high concentrations to protein formulations, due to their weak binding affinity (Kheddo *et al*., 2016), to alleviate protein interactions and partially unfolded species, thus reducing aggregation (Tosstorff *et al.*, 2019). Amino acid excipients in particular have been shown to improve conformational stability and lower aggregation kinetics. An example being arginine-glutamate, where an equimolar mixture of L-Arg and L-Glu reduces self-association and improves conformational stability (Kheddo *et al.*, 2014; Zhang *et al.*, 2016). The general consensus is that excipients binding to folded protein, as opposed to partially unfolded, should be maximised to increase stability. However, discriminating between binding to folded or unfolded conformations via experimentation can be very challenging. This is due to the weak excipient binding affinity being outside the sensitivity of methods like surface plasmon resonance (SPR), isothermal titration calorimetry (ITC) and fluorescence polarisation (FP) (Du *et al.*, 2016) as well as difficulty in determining partially unfolded structures. In addition, studies have shown excipient binding strength to have weak negative to no correlation with thermal stability (Zalar, Svilenov and Golovanov, 2020). This emphasises the point that no single formulation is beneficial for all proteins.

1.7.2 Lyophilisation

Lyophilisation, also known as freeze-drying (FD), is a process used for its ability to stabilise protein subject to temperature changes, improve long-term storage stability and ease handling during transportation (Manning *et al.*, 2010). Alternate drying methods can also be used such as spray drying and foam drying (Abdul-Fattah *et al.*, 2007). FD typically involves the stages of freezing, primary drying and secondary drying. Freezing is where the majority of water is removed from the protein drug and excipients (Tang and Pikal, 2004). However, this stage can induce many destabilising stresses for protein. This includes increasing protein concentration (and thus protein-protein interactions), changing pH due to crystallization of buffer salts, reducing hydrophilic interactions due to dehydration caused by ice formation and formation of ice-aqueous interfaces (Strambini and Gabellieri, 1996; Tsuruta, Ishimoto and Masuoka, 1998; Pikal-Cleland *et al.*, 2002). As illustrated in Figure 5, there are three states in which frozen solids can exist; amorphous, crystalline and polycrystalline (Zhang, 2017). Amorphous solids are the common state for materials like proteins, some sugars and polymers. They possess a short-range crystal-like order, residual crystallinity and varying areas of density. Amorphous solids also have higher entropy and free energy in comparison to corresponding crystals and due to this higher instability, they are able to undergo structural relaxation or crystallization (Yu, 2001). Upon freezing, the transition of a liquid's viscosity, entropy, enthalpy and volume to that observed in a 'glassy state' is termed the 'glass transition temperature (T_g)' (Ringe and Petsko, 2003). The equal for crystallized solutes is the 'eutectic temperature (T_e)' (Tang and Pikal, 2004).

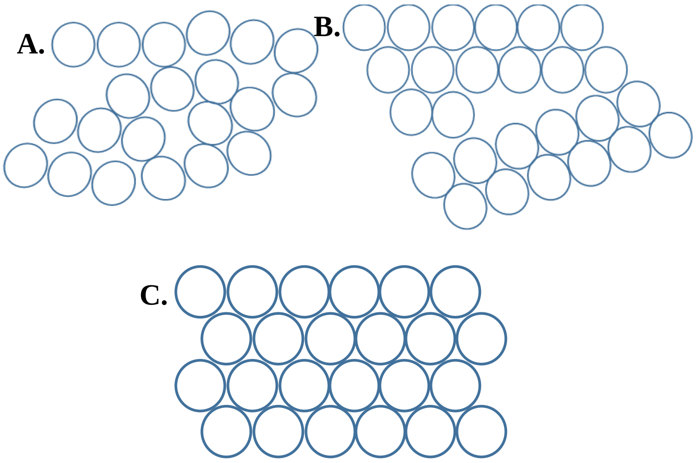


Figure 5. The Different States of Frozen Solids. The different order of molecules observed in the frozen state are illustrated here with the following representations; **A.** amorphous state, **B.** polycrystalline and **C.** crystalline. Each hollow circle represents a single molecule.
(Adapted from Zhang, 2017 and Yu, 2001)

The next step of FD, primary drying, typically consumes most of the FD cycle time and can be used to optimize the product temperature. At this stage, the ice formed during freezing is sublimated off the product by pulling a vacuum at low temperatures. Sublimation incorporates heat- and mass-transfer whereby water avoids the liquid state and passes directly from solid state to vapour state (Nireesha *et al*., 2013). Secondary drying follows on from primary drying and serves to reduce the residual moisture from the amorphous product to an optimal level for stability (typically less than 1%). Consideration of formulation is also important for FD. Bulking agents like mannitol or glycine are used to provide mechanical stability to the cake structure produced after FD a mixture. The idea behind this is that the amorphous phase collapses onto the surface of the crystalline phase at temperatures between T_g and T_e (Tang and Pikal, 2004; Manning *et al.*, 2010). A more cost effective approach for optimizing FD formulation has been done on an ultra-scale down (USD) level using ‘design of experiments’ method (Grant *et al.*, 2012). Therefore, a deeper understanding of protein-excipient interactions in the lyophilised state and comparison to the liquid state could broaden our understanding into stabilising mechanisms of excipients. More understanding is also needed with regards to the effect of reconstitution on protein conformation and stability (Zhang *et al.*, 1995).

1.7.3 Attempting Rational Mutagenesis to Improve Stability

Changing the sequence of amino acids in proteins, otherwise known as mutagenesis, can also improve stability. Many rational/semi-rational approaches have been taken to improve protein stability via mutagenesis. This includes directed evolution, rigidifying flexible sites (RFS), protein design automation (PDA), adding novel disulphide bonds and optimizing surface charge-

charge interactions (Luo *et al*, 2002; Dombkowski, 2003; Gribenko *et al*, 2009; Dalby, 2011; Yu and Huang, 2014).

Directed evolution refers to a range of molecular biology techniques that permit the evolutionary process to be mimicked in the laboratory. This includes random mutagenesis (using error-prone PCR), recombination of functional sequences, targeting desired residues for random mutagenesis and rational design (i.e. implementing mutations towards consensus sequences) (Chen and Arnold, 1993; Stemmer, 1994; Miyazaki and Arnold, 1999; Lehmann and Wyss, 2001; Dalby, 2011). Other mutagenesis approaches seek to identify and then modify regions that impact protein stability; namely flexible regions.

1.7.3.1 Identifying Flexible Sites

Common methods for identifying flexible regions are the B-FITTER program, computational molecular dynamics (MD) simulations and Floppy Inclusion and Rigid Substructure Topography (FIRST). B-FITTER is a program that calculates the flexibility for an amino acid residue by averaging the B-factor value from the protein crystallographic data. B-factor values are used to represent flexibility because they indicate atomic displacement parameters obtained from the crystallographic data (Parthasarathy and Murthy, 2002). Hence, residues with higher B-factor values are more flexible. However, this approach has some limitations, such as; fluctuations tend to be larger in solution than in the crystal state and B-factor values may greatly differ between proteins due to crystal quality just to name a few.

MD differs from the B-FITTER approach in that it simulates the motional properties of atoms in a protein structure (obtained from the Protein Data Bank - PDB). Flexibility of residues is determined from MD simulations using average root mean square fluctuation (RMSF) values. These average RMSF values reflect the mean amplitude of each residue relative to a mean reference position during the simulation. Therefore, higher RMSF values indicates greater flexibility (Yu and Huang, 2014b).

1.7.3.2 Rigidifying the Identified flexible Regions

After screening for the flexible regions, approaches that can be taken to rigidify them include iterative saturation mutagenesis (ISM), introduction of proline, addition of salt bridges, adding novel disulphide bonds, optimising surface charge-charge interactions and structure-guided consensus mutagenesis. ISM is an effective approach for directed evolution because it targets flexible sites that can be identified from B-factor values and then subjects these sites to random mutagenesis (Reetz and Carballeira, 2007). Computational programs have also been developed for rational design of proteins with optimal surface charge-charge interactions and disulphide bonding (Dombkowski, 2003; Gribenko *et al.*, 2009).

Rosetta is a computational protein design program. It takes a PDB structure and uses Monte Carlo optimization along with simulated annealing to identify and suggest amino acid sequences that pack well, bury their hydrophobic atoms and satisfy the hydrogen bonding potential of polar atoms (Liu and Kuhlman, 2006). Therefore, in effect, Rosetta aims to predict stabilising sequences. The RosettaDesign application has been employed previously by Yong Hwan Kim *et al*. Here B factor values were used to identify eleven target residues for thermostabilisation in *Coprinus cinereus* peroxidase (CiP). RosettaDesign was then used to create 8 mutants, two of which showed increased thermostability as well as conserved bioactivity in comparison to the wild-type (WT) (Kim *et al.*, 2010). Rosetta aims to predict the change in stability (ΔΔG) of a monomeric protein induced by a point mutation. An increase in stability is represented by a negative ΔΔG, whereas a positive ΔΔG indicates a decrease in stability (Kellogg, Leaver-Fay and Baker, 2011). The Rosetta ddg_monomer application is able to make these calculations for each residue within a protein, endowing it with a high throughput capability. Nevertheless, a crucial limitation of Rosetta ddg_monomer is that it is not suitable for predicting several mutations simultaneously.

1.7.3.3 Combining Mutations may be Required to Significantly Enhance Thermostability

Since individual mutations contribute relatively little to stabilising large proteins, multiple simultaneous mutations are often required to achieve higher stability (Zhao and Arnold, 2002; Goldenzweig *et al.*, 2016). However, combining two or more mutations in the same protein brings into question the interactions between these mutations. Interaction between mutations is known as intragenic epistasis. Epistasis is where the fitness effect of one mutation depends on the genetic background at another loci and is believed to be a main factor in determining short- and long-term protein molecular evolution (Parera and Martinez, 2014). The epistatic interactions between mutations can either be mathematically additive or non-additive. In additive epistasis, $\Delta XY = \Delta X + \Delta Y$ where ΔXY (a double mutant) is the collective contribution from experimental changes observed in single mutants X and Y.

However, mutations that independently make a positive contribution may interact in a non-additive way when combined. This non-additivity can take the form of positive epistasis, negative epistasis (partially additive), negative sign epistasis or reciprocal sign epistasis. Positive epistasis occurs when $\Delta XY = \Delta X + \Delta Y + I$, where I represents interaction between mutation X and Y in cases where the side chains of the two residues are close in contact with one another or when one or both mutations switch the reaction mechanism. Negative epistasis acts in a way that makes the fitness of the double mutant not as bad as either single mutant alone, where $\Delta XY < \Delta X + \Delta Y$. Hence, only partial additivity in the double mutant is observed (the double mutant phenotype is smaller than expected under additivity). In negative sign epistasis, ΔXY < ΔX or ΔY and in reciprocal sign epistasis ΔXY < 0 (Reetz, 2013; Starr and Thornton, 2016; Yu and Dalby, 2018a).

Therefore, achieving additivity in mutagenesis could be an effective approach in biobetter manufacture owing to its capability of making mutants that elicit optimal combined effects of the individual mutations within the protein. This additivity most likely occurs when the structural regions of mutated residues do not substantially overlap (Yu and Dalby, 2018a). Epistasis between spatially distant residues is believed to be mediated through a communicating amino acid network of interactions. Although explanations for what mediates this long-range communication network of interactions remain parsimonious, protein dynamical and thermodynamic properties has been suggested to be an influence (Whitley and Lee, 2009; Posfai *et al.*, 2018).

1.8 Experimental Methods for Analysing Protein Stability and Bioactivity

There are many biophysical approaches to probe a protein's HOS, dynamics and aggregation. Table 2 classifies various commonly used techniques methods and part Figure 6 accompanies this by ordering them in terms of resolution. Means of testing G-CSF bioactivity are not so plentiful but still serve as a good way to determine structure-function relationships. NFS-60 cell bioassays are a common way to probe bioactivity by monitoring the proliferation of cells (Weinstein *et al.*, 1986). An enzyme immunoassay approach has also been shown to be a good way to measure G-CSF activity (Motojima, *et al*., 1989).

Interrogation of a protein's quaternary structure (aggregation), size, shape and mass can be done by studying its hydrodynamic properties (i.e. its global shape). Hydrodynamics is assessed by monitoring protein movement through a liquid medium in response to driving forces such as thermal kinetic energy of the protein (exploited in SEC) or a high centrifugal field as exploited in sedimentation velocity AUC (SV-AUC). These techniques yield important information about unique sample characteristics like the heterogeneity of protein conformers (e.g. aggregates).

Electrophoretic methods include SDS-PAGE and native gels. SDS-PAGE can be used to elucidate the covalent nature of aggregates whereas native gels are able to reveal molecular shape, weight and intrinsic charge due to the nondenaturing conditions of the gel. Therefore, SDS-PAGE is useful for aggregation studies but not on studies of HOS.

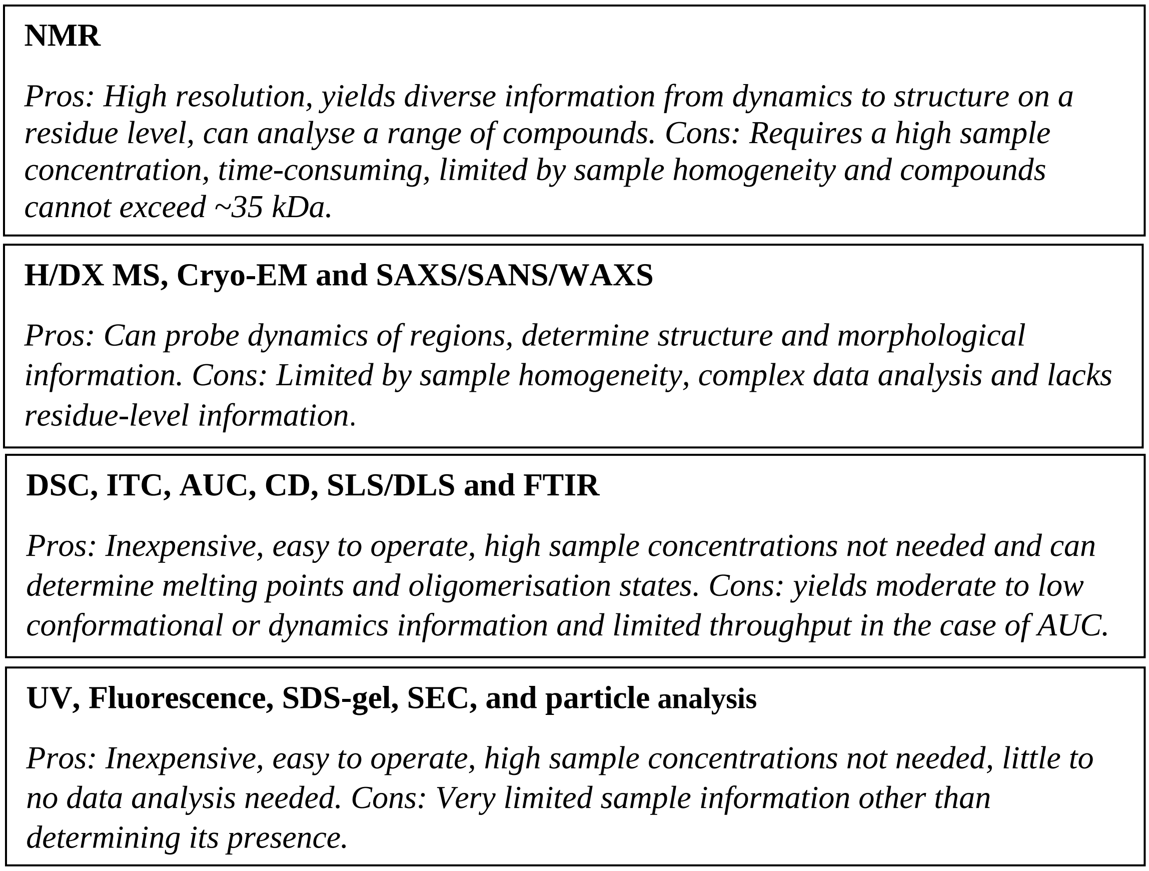
Thermodynamic techniques study the HOS by monitoring the flow of heat as the protein is exposed to increasing temperatures. The theory behind this is that changes in the energy associated with the conformational changes translate into changes in the heat flow. Two common techniques that accomplish this are ITC and DSC. A fairly complementary technique that monitors the change in HOS is Spectroscopy.

Spectroscopy exploits the presence of chromophore and fluorophore entities within protein that can absorb electromagnetic radiation. The output spectra, therefore, will change as the microenvironments of the fluorophores and chromophores change as a result of HOS perturbations.

1.8.1 Nuclear Magnetic Resonance Spectroscopy is Information-rich

Nuclear magnetic resonance (NMR) is also a spectroscopic method and is shown to be in the highest tier in Figure 6 because it gives the highest resolution of HOS. It is capable of mapping the position of the amino acid residues and investigating local structural environment and dynamics. This made possible by the exploitation of the magnetic properties of certain atomic nuclei such as ^{1}H, ^{15}N and ^{13}C (Levitt, 2013). When these nuclei are incorporated into proteins and introduced to an external magnetic field (B_0), the rotation of these nuclei will align with this field. Another magnetic field applied perpendicular to the B_0 field causes rotation to transverse into another plane before precession back to B_0. The rate of this precession varies based on local magnetic environment, which is influenced by local protein structure and solvent (Kleckner and Foster, 2011).

NMR has also proven to be a robust method for comparing the HOS of filgrastim biosimilars. MS (hydrogen/deuterium exchange in particular) can be a complementary tool to NMR as it can elucidate the physicochemical environment of amino acids, thus providing information on conformation, dynamics and protein-protein interactions. (Houde and Berkowitz, 2014; Ghasriani *et al.*, 2016). Solid state (ss-) NMR and ssHDX-MS are also useful for assessing dynamics and conformation in the lyophilised (solid) state. However, in this study, NMR will be combined with a thermal melt. Combining NMR with a changing variable is a valuable method because it can probe the structural and dynamic sensitivity of protein regions to these environmental changes (Trainor *et al*., 2020; Aubin *et al*., 2015).



NMR

Pros: High resolution, yields diverse information from dynamics to structure on a residue level, can analyse a range of compounds. Cons: Requires a high sample concentration, time-consuming, limited by sample homogeneity and compounds cannot exceed ~35 kDa.

H/DX MS, Cryo-EM and SAXS/SANS/WAXS

Pros: Can probe dynamics of regions, determine structure and morphological information. Cons: Limited by sample homogeneity, complex data analysis and lacks residue-level information.

DSC, ITC, AUC, CD, SLS/DLS and FTIR

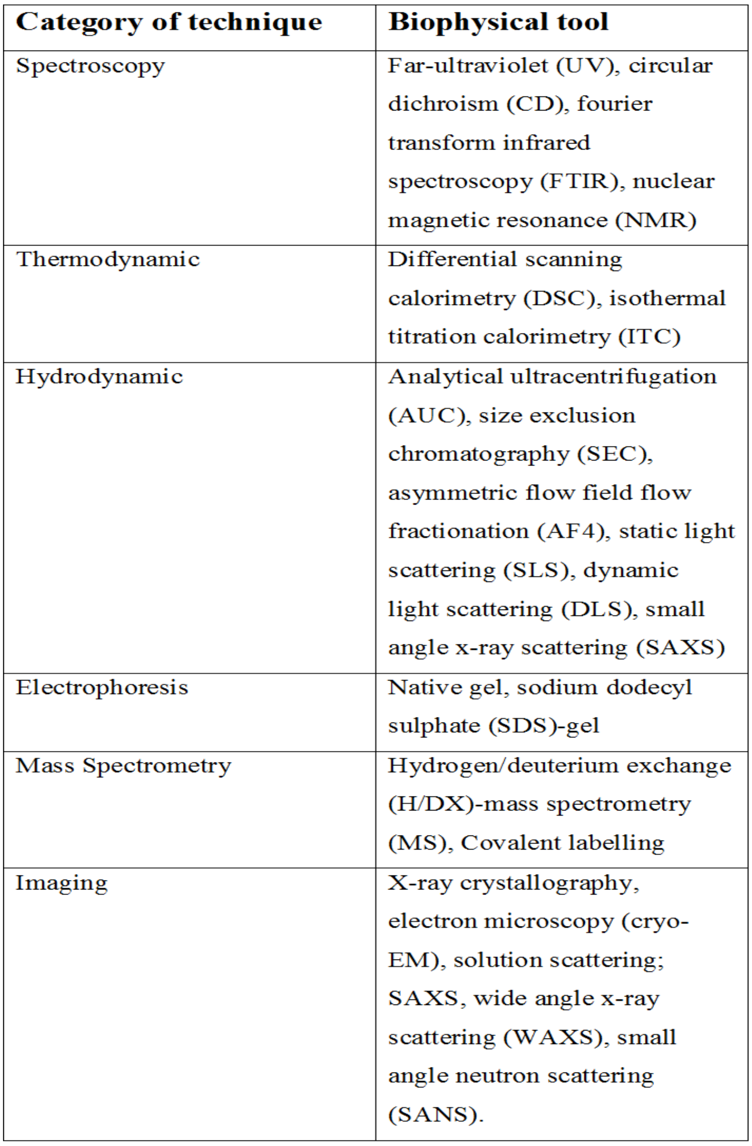
Pros: Inexpensive, easy to operate, high sample concentrations not needed and can determine melting points and oligomerisation states. Cons: yields moderate to low conformational or dynamics information and limited throughput in the case of AUC.

UV, Fluorescence, SDS-gel, SEC, and particle analysis

Pros: Inexpensive, easy to operate, high sample concentrations not needed, little to no data analysis needed. Cons: Very limited sample information other than determining its presence.

Figure 6. Hierarchy of Biophysical Tools. Hierarchy depicting the biophysical tools in their order of resolution, where the higher resolution tools are towards the top. The collective pros and cons of the techniques in each tier are also detailed in italics.

(Adapted from Houde and Berkowitz, 2014)



Category of technique	Biophysical tool
Spectroscopy	Far-ultraviolet (UV), circular dichroism (CD), fourier transform infrared spectroscopy (FTIR), nuclear magnetic resonance (NMR)
Thermodynamic	Differential scanning calorimetry (DSC), isothermal titration calorimetry (ITC)
Hydrodynamic	Analytical ultracentrifugation (AUC), size exclusion chromatography (SEC), asymmetric flow field flow fractionation (AF4), static light scattering (SLS), dynamic light scattering (DLS), small angle x-ray scattering (SAXS)
Electrophoresis	Native gel, sodium dodecyl sulphate (SDS)-gel
Mass Spectrometry	Hydrogen/deuterium exchange (H/DX)-mass spectrometry (MS), Covalent labelling
Imaging	X-ray crystallography, electron microscopy (cryo-EM), solution scattering; SAXS, wide angle x-ray scattering (WAXS), small angle neutron scattering (SANS).

Table 2. Characterisation Tools. Biophysical tools corresponding to the type of technique to which they belong.

1.9 Project Aims and Objectives

This project aims to identify regions in G-CSF important to structural resilience and function, and determine how excipients exhibit their mechanism of action on these significant regions. To this end, varied-temperature NMR will be used to probe regions that experience significant environmental and dynamic changes. These observations will be cross-referenced with *in silico* modelling to validate and illustrate structural and functional mechanisms. Mutations will also be constructed in these regions of interest to evaluate their effect on molecule stability and functionality. Varied-temperature NMR will also probe the mechanisms of action exhibited by excipients and mutational analysis used to confirm these mechanisms as well as the significance of the regions they act upon.

2. Materials and Method

All sterile filtration was performed with Millex-GP 0.2 μM, 33 mm, polyethersulfone (PES) sterile syringe filters (Millipore, Hertfordshire, UK)

2.1 Molecular biology

2.1.1 DNA Purification and Site-directed Mutagenesis

The pET21a plasmid with the WT human granulocyte colony stimulating factor (G-CSF) gene was provided by Dr Adrian Bristow (NIBSC; Bristow et al., 2012) in BL21 (DE3) Escherichia coli cells. Plasmid extraction was performed with a QIAprep Spin Miniprep Kit and procedure (QIAGEN Ltd, West Sussex, UK). Overnight 10 mL cultures of BL21 (DE3) E. coli cells were grown at 37^{o}C with 250 rpm agitation in Luria Bertani (LB) media containing 1 mM ampicillin (Amp). The final elution step was altered so that the elution buffer stayed on the column for five minutes and the final elution volume was 20 μL. This ensured that the final DNA concentration, measured using NanoDrop A260 values (Thermo Fisher Scientific Inc., Wilmington, USA), was high enough for sequencing (section 2.1.3).

Primers (Figure S.17) for site-directed mutagenesis were codon optimized using the OpenWetWare site (https://openwetware.org/wiki/Escherichia_coli/Codon_usage) and designed with the .py script in Figure S.16 so that their melting temperature was between 65^{o}C and 75^{o}C and there was no GC clamping. Invitrogen (Massachusetts, USA) conducted primer synthesis. Site-directed mutagenesis was performed using a QuikChange Lightning site-directed mutagenesis kit (Agilent Technologies, Inc., Santa Clara, CA, USA), as per instructions, and polymerase chain reaction (PCR) for mutant strand synthesis. PCR was performed with a C1000 Touch™ Thermal Cycler (Bio-Rad, Hercules, CA, USA). Each PCR reaction was set up with 125 ng of oligonucleotide primers, 50 ng of WT G-CSF plasmid template and a three-minute extension time. Parental (non-mutated) DNA was digested post-PCR reaction with Dpn I restriction enzyme.

2.1.2 DNA gel Electrophoresis

Successful mutant plasmid amplification was confirmed with a restriction digest using EcoRI-HF restriction enzyme and 1X NEBuffer (New England BioLabs Inc, Ipswich, USA). Digest samples were then incubated at 37°C for 1 hr before being mixed with 1X loading dye (New England BioLabs Inc, Ipswich, USA). DNA gels were prepared with 1% (w/v) agarose and 1X Invitrogen SYBR Safe staining reagent (Thermo Fisher Scientific Inc., Massachusetts, USA). Digest samples were loaded into the gel wells along with a 1 kb DNA ladder (New England BioLabs Inc, Ipswich, USA) and electrophoresis was performed at a constant voltage of 80 V for 1 hr in 1X TAE

(Thermo Fisher Scientific Inc., Massachusetts, USA) running buffer. A single band at ~5.9 Kbp, observed with a Geldoc 2000 (Bio-Rad, Hercules, CA, USA), confirmed successful plasmid amplification.

2.1.3 Mutant Plasmid Transformation and Sequencing

Following confirmation of successful PCR, 2 μL of Dpn I treated mutant plasmid sample was transformed into 45 μL of XL10-Gold ultracompetent cells that came with the QuikChange Lightning kit (section 2.1.1). To aid cell recovery, each transformation reaction was mixed with NZY+ broth. Subsequently, 50 μL of transformed cells were plated onto LB/Amp agar plates containing 80 μg/ml X-gal and 20 mM IPTG (Generon Ltd, Maidenhead, UK), and then incubated at 37 °C overnight. A single colony for each mutant could then be picked and the DNA purified, as described in section 2.1.1, and aliquoted for sequencing and later transformation (section 2.1.4). Plasmid sequencing was conducted by Source BioScience (Nottingham, UK) for WT and mutants, using their stock T7F primers, confirming that the correct mutation had been made. NZY+ broth and LB/AMP agar plates were prepared according to the QuikChange Lightning kit manual.

2.1.4 Glycerol Stocks

After confirming from sequencing that mutagenesis was successful, 250 ng of the purified mutant DNA was transformed into E. coli BL21 (DE3) competent cells (New England BioLabs Inc, Massachusetts, USA). Cell recovery was aided by SOC outgrowth medium, also provided by New England BioLabs, followed by plating on LB/Amp agar plates and storage at 37 °C overnight. A single colony for each mutant was then picked, grown in 10 mL of LB/Amp media overnight (37 °C and 250 rpm) and mixed with 50% (v/v) of sterile filtered glycerol solution at a 1:1 ratio before storage at -80°C.

2.2 Cell Culture

Media for WT and mutant cultures were autoclaved for 20 minutes at 120°C and all subsequent additions filter sterilized. Moreover, any transfer of media or components was performed in a safety cabinet. Seed cultures of 10 mL Terrific Broth (TB/Amp) in 50 mL falcon tubes were prepared from glycerol stocks and incubated at 37°C, 250 rpm overnight. Sterile 500 mL TB/Amp media in 2 L baffled flasks were inoculated with seed cultures and further incubated at 37°C, 250 rpm. Expression was induced at an OD600 of 0.6 by spiking in Isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM.

Minimal media was prepared for NMR using 15N labelled (NH4)2SO4, PO4/NaCl, Na2SO4, EDTA trace elements, MgSO4, CaCl2, d-Biotin, Thiamine and d-Glucose. PO4/NaCl, Na2SO4, EDTA trace elements were autoclaved at 120°C for 20 minutes. The remaining components, along with Amp, were filter sterilised. A 100 mL seed culture of transformed E. coli BL21 (DE3) competent cells (New England BioLabs Inc, Ipswich, USA) in minimal media/Amp was incubated overnight at 37^{o}C with shaking at 250 rpm. This seed culture was then transferred to 2 L of minimal media/Amp in baffled flasks (i.e. two 5 L flasks with 1 L of media) and incubated (37^{o}C with shaking at 180 rpm) overnight. Expression was induced at an OD600 of 0.6, by spiking with IPTG to a final concentration of 1 mM. The culture was incubated overnight at 37^{o}C with shaking at 180 rpm.

2.3 Primary Separations

Cells were harvested with centrifugation at 7080 x g for 20 minutes (4^{o}C) using an Avanti J-20 XPI (Beckman Coulter, Inc). Cell pellets were then washed in 40 mL of 10 mM phosphate buffer saline (PBS) and centrifuged at 7728 x g for 30 minutes (4°C) into ~4 g pellets in 50 mL falcon tubes. To help with cell lysis, these pellets were stored at -20 °C. Each cell pellet was defrosted by leaving to stand for 30 minutes at room temperature (RT) and then resuspended in 40 mL 10 mM PBS. Cell lysis was carried out by giving the resuspended pellets a single pass through a APV LAB40 high pressure homogeniser at 1000 Bar and storing them on ice. Cell lysis was also aided by adding sodium deoxycholate at 1 mg/mL and rolling at room temperature for 15 minutes. The high viscosity of the lysate from DNA release was reduced by addition of 20 µL Benzonase nuclease (25 U/mL; Merck Millipore) and rolling continued for 15 minutes. The lysate was centrifuged at 17,700 x g, 30 min, 4°C (Avanti J20 XPI; Beckman Coulter, Inc., Fullerton, CA, USA) to pellet the GCSF inclusion bodies (IB). After removal of the supernatant, the IB pellet was washed twice to remove host cell impurities. In all steps, the pellets were resuspended at 1:40 (w/v) in wash buffer using a homogeniser 850 (Thermo Fisher Scientific Inc., Wilmington, USA) and repelleted via centrifugation at 17,700 x g for 30 minutes (4°C). Wash A contained 50 mM Tris pH 8, 5 mM EDTA and 2% Triton X-100 (w/v), wash B contained 50 mM Tris pH 8, 5 mM EDTA and 1 M NaCl.

Pellet solubilisation was achieved using a pH shift procedure, which included re-suspension in 10 mL of 4 M urea and pH adjustment to pH 12 using strong NaOH, followed by rolling for 30 min at RT. Refold was achieved by diluting this solution dropwise by 20X into 1 M Arginine.HCl buffer pH 8.25, followed by rolling for > 12 h at RT. Refolding was quenched by pH adjustment to 4.25 using strong glacial acetic acid followed by rolling at RT for 2.5 hours. The refold was clarified by centrifugation at 17,700 x g, 20 min, 4°C, (Avanti J-20 XPI; Beckman Coulter, Inc., Fullerton, CA, USA), the supernatant retained and concentrated to a final volume of 10 mL using an Amicon stirred cell (with 10 kDa, 29.7 mm diameter ultra-centrifugal filter units; Merck Millipore). Concentration was continued with Amicon Ultra-15 10 kDa cut off membrane centrifugal filters (Merck Millipore, Billerica, Massachusetts, USA) at 1,389 x g and 4 °C.

2.4 Purification and Concentration

The 10 mL concentrated sample was purified by size exclusion chromatography (SEC) on an ÄKTA™ Explorer (GE Healthcare Life Sciences, Germany) using a HiLoad® 26/60 Superdex® 200 prep grade column (GE Healthcare Life Sciences, Germany; 2.6 cm internal diameter; i.d., 60 cm bed height, 320 mL column volume; CV). A 10 mL injection loop was used to load the sample onto the column whilst eluted was performed isocratically in 50 mM Sodium Acetate pH 4.25 at 2.5 mL/min. Fractions with > 0.1 mg/mL concentration were pooled and concentrated to a final stock concentrations using Amicon Ultra-15 10 kDa cut off membrane centrifugal filters at 1890 x g and 4°C. These concentrations were 1.7 mg/mL (0.09 mM) for NMR samples and 0.3-0.5 mg/mL for non-NMR samples.

2.5 Formulation with Excipients

When formulated with excipients, mutant and WT samples were mixed at a 1:1 ratio (protein to excipient) to a final protein concentration of 0.15 mg/mL (pH 4.25, 50 mM sodium acetate) and 1X excipient concentration (from a 2X stock). Excipient solutions were prepared by mixing the solid excipient with 50 mM sodium acetate at pH 4.25 so that protein buffer was not diluted. These formulations were used in lyophilisation (section 2.10), bioactivity assays (section 2.7) and thermal degradation studies (section 2.8). Isotopocally labelled WT-GCSF samples for NMR were mixed with excipients at a 9:1 ratio (protein to excipient) to achieve a protein concentration of 1.53 mg/mL (0.08 mM) and 1X excipient concentration from 10X stock.

2.6 Protein Quantification

Quantification of protein content in sample was performed by measuring absorbance at A_{280} with a NanoDrop (Thermo Fisher Scientific Inc., Wilmington, USA). The concentration was calculated using the Beer-Lambert Law (Eqn.2), where c is the concentration, A is absorbance, L is the path length and ε is the extinction coefficient of 0.86 (Herman, Boone and Lu, 2002). Extinction coefficients for mutants were calculated from the number (n) of tyrosine (Y), tryptophan (W) and cysteine (C) residues using Eqn.3 (Edelhoch, 1967; Gill and Von Hippel, 1989). Here εW is 5690 $M^{-1}cm^{-1}$, εY is 1280 $M^{-1}cm^{-1}$ and εC is 120 $M^{-1}cm^{-1}$.

$$c = \frac{A}{\varepsilon L} \textbf{ Eqn. 2}$$

$$\varepsilon_{mutant} = nW.\varepsilon W + nY.\varepsilon Y + nC.\varepsilon C \textbf{ Eqn. 3}$$

2.7 Bioactivity

To quantify units of functionality for the purified G-CSF variants (at various formulations), CellTiTer 96Aqueous One Solution Cell Proliferation Assays (Promega) were performed with murine GNFS-60 cells; A bioassay developed by Wadhwa *et al*., 2011. This assay also consisted of 96-well sterile, clear, TC-treated polystyrene microplates (Falcon) and RPMI-1640 medium (Sigma-Aldrich, Missouri, USA). Components of RPMI-1640 (assay) medium were 0.5% (v/v) Penicillin-Streptomycin solution (stock: 10,000 units/mL penicillin and 10 mg/mL streptomycin), 1% (v/v) 200mM L-glutamine (Sigma-Aldrich, Missouri, USA) and 5% (v/v) Fetal Bovine Serum (FBS). All solutions were warmed to 37°C, dilutions and cell culture manipulations were performed in a safety cabinet and incubation was done at 37°C, 5% CO_2 in a humidified incubator unless otherwise stated.

Before performing the assay, the GNFS-60 cells, stored in a liquid nitrogen freezer (in 1mL cryogenic vials), were thawed and fed for at least a week. Thawing consited of immediate warming in a bench top 37°C waterbath. For cell feeding, the cells were suspended and grown in 50 mL of growth medium (assay medium containing 2 ng/mL of r-HuG-CSF; Amgen, Uxbridge, UK) in T-75 flasks (sterile plastic and vented cap). Flasks were split to ~10^5 cells/mL every 2-3 days, depending on cell concentration, and were cultered upright in the incubator.

Once ready for the assay, these GNFS-60 cells were washed to remove residual GCSF. This washing procedure consisted of centrifugation (for 10 minutes at 1,300 × g) in 50 mL sterile plastic falcon tubes followed by resuspension of the pellet in 40 mL of assay medium. Washing was repeated for a total of four washes and the final pellet was resuspended in 10 mL of assay media. Cells were then counted with a Countess Automated Cell Counter (Invitrogen, Life Technologies Corp, Paisley, UK) after adding 11 μL of Trypan blue viability stain (Sigma-Aldrich, Missouri, USA) to 11 μL of cells and pipetting 10 μL of this to each chamber of the cell counter slide. The cells were diluted to a final concentration of 2×10^5 cells/mL. All G-CSF formulation samples to be tested and the NIBSC 2nd international reference standard were diluted to 2 ng/mL in assay media and 100 μL added to the first wells of the microplates. The reference standard and a negative control of just assay media were added so that they flanked the G-CSF variant samples. To the wells following the first row was added 50 μL of assay media. Serial dilutions down the plate were conducted using 50 μL from the first row, followed by the addition of 50 µL of cells (at 2 × 10^5 cells/mL) to each well. Therefore, the protein concentration ranged from 1 ng/mL to 0.008 ng/mL across the plate. These plates were then covered and incubated for 48 hours.

Analysis of the cell response was measured by adding 20 µL of Cell Counting Kit-8 (CCK-8; Stratech, Ely, UK) to each well, incubating the plates for a further 3-4 hours and then measuring the absorbance of the wells at 450 nm. Absorbance was measured using a Spectramax 340PC (Molecular Devices LLC, Wokingham, UK) with five seconds of plate shaking before the reading.

2.8 Accelerated Thermal Degradation

Thermostability of G-CSF mutants compared to WT was assessed using thermal unfolding temperature (T_m) values, i.e. the point at which 50% of the protein population was unfolded, obtained from the Unit/UNcle (Unchained Laboratories, UK). T_m values were determined from the barycentric mean (BCM) of the protein intrinsic fluorescence spectra at 280-460 nm (266 nm excitation) at each temperature along the thermal ramp by fitting BCM to the van't Hoff equation (eqn.4). In this equation, I_T represents the observed signal, I_N and I_D are the native and denatured baseline intercepts, a and b are the native and denatured baseline slopes, T is the temperature, ΔH_{vh} is the van't Hoff enthalpy and R is the gas constant (Consalvi *et al.*, 2000). The thermal ramp was conducted using linear heating from 20°C to 90°C at 1°C/minute and a 30 second starting incubation. Mutants and WT samples formulated to a final concentration of 0.5 mg/mL (or 0.3 mg/mL for the P56V mutant and comparative WT sample), as mentioned in section 5.1.4, were run on the Unit/Uncle in quadruplicates. Variants formulated with excipients were run in triplicate. Each well of the Uni cuvette was loaded with 9 µL samples and loaded onto the Unit/Uncle.

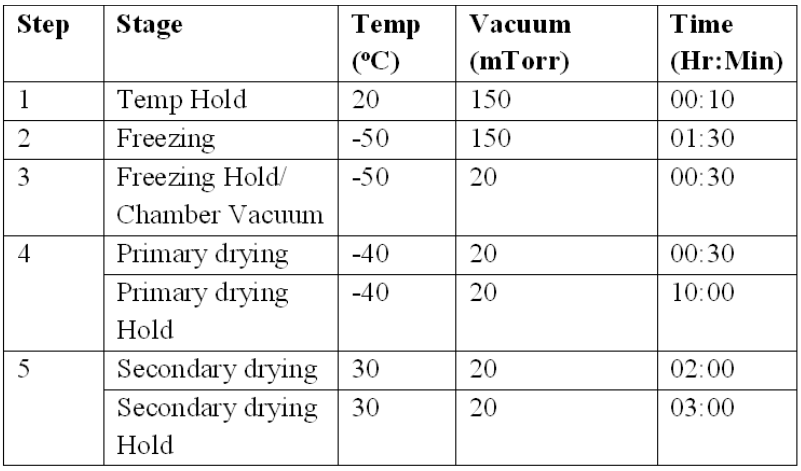
$$I_T = \frac{(I_N+aT)+(I_D+bT)\exp[\frac{\Delta H_{vh}}{R}\left(\frac{1}{T_m}-\frac{1}{T}\right)]}{1+\exp[\frac{\Delta H_{vh}}{R}\left(\frac{1}{T_m}-\frac{1}{T}\right)]} \quad \textbf{Eqn.4}$$

2.9 Mass Spec

Molecular weight of purified G-CSF variants were determined with liquid chromatography mass spectrometry (LC-MS) using a Agilent 6510 QTOF system. Samples were prepared at 0.3 mg/mL, 50 mM sodium acetate pH 4.25 and 10 µL loaded onto a PLRP-S, 1000A, 8 µM, 150 mm x 2.1 mm column maintained at 60°C. Separation was achieved using a gradient elution at 0.3 mL/min with mobile phase A (water with 0.1% formic acid) and B (acetonitrile, with 0.1% formic acid).

2.10 Lyophilisation

Lyophilisation of G-CSF variant formulations was conducted by loading 100 µL in TC-treated polystyrene 96-well flat-bottom plates (Greiner Bio-one Ltd, Gloucestershire, UK). The plate skirts were removed so that the bottom of the wells made contact with the Virtis Genesis 25EL freeze-drier shelf. The lyophilisation cycle used (Table 3) contained an anneal step because mannitol was used in formulation.



Step	Stage	Temp (ºC)	Vacuum (mTorr)	Time (Hr:Min)
1	Temp Hold	20	150	00:10
2	Freezing	-50	150	01:30
3	Freezing Hold/ Chamber Vacuum	-50	20	00:30
4	Primary drying	-40	20	00:30
	Primary drying Hold	-40	20	10:00
5	Secondary drying	30	20	02:00
	Secondary drying Hold	30	20	03:00

Table 3. Lyophilisation Cycle. Freezing hold and application of the chamber vacuum was conducted simultaneously.

2.11 Non-reduced SDS-PAGE

Protein purity was examined during expression and purification using non-reduced sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE). Samples from stages preceding protein refolding were prepared so that 500 µL was centrifuged for 5 minutes at 13,523 x g and the supernatant decanted before resuspending the pellet in 200 µL of 1X Novex NuPage LDS (Thermo Fisher Scientific Inc., Wilmington, USA). For post-refold analysis, samples were mixed with Milli-Q water and LDS to achieve 0.1 mg/mL protein concentration and 1X LDS.

Samples were heated at 90°C for 5 minutes before being loaded at 10 µL into Novex NuPage 15-well 4-12% Bis-Tris precast gels with 1X NuPage MES running buffer (Thermo Fisher Scientific Inc., Wilmington, USA). Pre-refold samples were centrifuged at 13,523 x g for 5 minutes and 10 µL of supernatant loaded following this heat treatment step. A PageRuler Prestained Protein (Thermo Fisher Scientific Inc., Wilmington, USA) with molecular weight markers ranging from 10 to 180 kDa was loaded at 6 µL into the first lane on every gel. Electrophoresis was conducted at a constant voltage of 200 V for 35 minutes. Gels were imaged with an Amersham Imager 600

(GE Healthcare Bio-Sciences, PA, USA) by staining with InstantBlue (Expedieon Ltd, Cambridgeshire, UK) for >1 hour and destaining with distilled water overnight.

2.12 Nuclear magnetic resonance (NMR) Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy was performed using a 700 MHz Bruker Avance NEO spectrometer fitted with a Bruker AEON refrigerated magnet and QCI-F cryoprobe. The ^{15}N-^{1}H HSQC spectroscopy experiments were performed using the *hsqcetfpf3gpsi* pulse sequence, while ^{1}H NMR spectra were collected using the *zgesgp* pulse sequence. Spectra were recorded in the temperature range 295 K to 323 K, at incremental steps of 2 K. To control for thermal drift of signals, 5 µL of trimethylsilylpropionic acid (TSP) was added to the protein sample. While TSP is not expected to affect chemical shifts or display any behavior similar to the excipients in this study, it was added to all samples to control for any changes it may induce.

2.12.1 Processing of NMR Spectra and Further Analysis

Both ^{1}H and ^{15}N-^{1}H HSQC experiments were processed in Topspin 4.0.8 (Bruker, Coventry UK). Signals from experiments at each temperature point were zeroed to the TSP signal. CcpNmr Analysis 2.4.2 (Vranken *et al*., 2005) was then used for further analysis in order to calculate Δδ and peak intensity.

2.13 Equations for NMR Observable Interpretations

The NMR observables δ and PI were scrutinized to give ∑Δδ, percentage change in PI, 90th percentiles of both observables and also residue correlation for both observables.

2.13.1 ∑Δδ

Calculation of ∑Δδ is illustrated in Figure S.1. ∑Δδ at each temperature is the cumulative change in microenvironment at that temperature, for example ∑Δδ at 297 K = Δδ from 295 K to 297 K and ∑Δδ at 301 K = (Δδ from 295 K to 297 K) + (Δδ from 297 K to 299 K) + (Δδ from 299 K to 301 K).

2.13.2 90th/95th Percentile for ∑Δδ

The 90th and 95th percentile for ∑Δδ was calculated at each temperature point along the thermal melt. The normal distribution of the ∑Δδ data set at each temperature was calculated and residues with a ∑Δδ above the 90th/95th percentile threshold of this distribution were considered to be in the 90th and 95th percentile respectively. The normal distribution equation is given as:

$$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ \textbf{Eqn. 5}$$

Where μ is the distribution mean, σ^2 is the variance and x is the independent variable.

2.13.3 90th Percentile for PI

The same normal distribution equation was used to calculate the 90th percentiles for PI. Here, the normal distribution was calculated for all data points across the thermal melt and residues with a PI value above the 90th percentile threshold of this distribution were determined to be 90th percentile.

2.13.4 Percentage Change

The percentage change in PI was calculated between the PI value at the start of the melt and maximum point of the melt for respective residues:

$$Percentage\ Change = \frac{Maximum\ PI - Start\ PI}{Start\ PI}\ \textbf{Eqn. 6}$$

2.13.5 Cross Correlation for $\Delta\delta$ and PI

Spearman's correlation (ρ) was used to calculate the correlation between residues. Coefficients were derived using δ and PI values at consecutive temperature points along the thermal melt. The Spearman's equation used was as follows:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}\ \textbf{Eqn. 7}$$

Where d_i is the difference between a pair of ranks and n is the number of observations.

2.14 Molecular Dynamics

2.14.1 Rosetta

The Cartesian_ddg application within the Rosetta software suite was used to relax PDB:2D9Q. Gromacs was used to clean PDB:2D9Q with the "grep -v HOH" command and then renumbered so that residue 7 was residue 1. This renumbered PDB was then relaxed and the lowest energy PDB was taken for another relaxation step. The lowest energy PDB from the second relaxation step was then used as the relaxed structure in this study. This structure was also carried forward to the following step where the ddg score (ddg = $\Delta G_{\text{Mutant}} - \Delta G_{\text{Wild-type}}$) for all possible single mutants (3192) was calculated using the .py file detailed in Figure S.15.

2.14.2 Calculating Solvent Accessible Surface Area

Solvent accessible surface area (SASA) was calculated using the lowest energy PDB from Rosetta on the online server ProtSA (Estrada *et al.*, 2009). A probe radius of 0.14 nm was used to perform this calculation.

2.14.3 APR Software

Consensus APRs were determined using AmylPred 2 (Tsolis *et al.*, 2013) based on 10 APR scanning software: namely AGGRESCAN, Amyloidogenic Pattern, Average Packing Density, Beta-strand contiguity, Hexapeptide Conf. Energy, NetCSSP, Pafig, SecStr, TANGO and WALTZ. The APR scanning software used in this study that was not part of this consensus is PASTA 2.0 (Walsh *et al.*, 2014).

2.14.4 Protein in Water

Simulation of G-CSF (Protein Data Bank ID code 2D9Q) was performed with the molecular dynamics (MD) software Gromacs version 2019. The pH of PDB structures was altered using the online server PDB2PQR *Version 2.1.1* (Dolinsky*, et al.*, 2007) with the CHARMM27 force field (ff). Molecular topology was generated with the program gmx pdb2gmx using the CHARMM27 ff and TIP3P water model. A cube (-bt cubic) defined the box type in which the protein was centered with at least 1.0 nm from the box edge (defined by -d 1.0). Sufficient Na^+ or Cl^- ions were added to neutralise in the system, which was energy minimised using the steepest descent method by submitting the jobs to the UCL high performance computing facility: Myriad. 100 ps position-restricted simulations were conducted under an NVT (with a constant number of particles, volume and temperature) and NPT ensemble (with a constant number of particles, pressure and temperature). All bond lengths were constrained using the LINCS algorithm and the time step of simulation was set to 2 fs. Each MD simulation was performed for a minimum of 100 ns.

Running coarse-grained (CG-) MD with the SIRAH 2.0 ff (Machado *et al.*, 2019) required the use of the AMBER ff when setting the protonation state with the PDB2PQR server. The SIRAH ff, tools and relevant molecular structures were downloaded from www.sirahff.com. The atomistic structure of PDB:2D9Q was mapped to its CG representation using the cgconv tool. Subsequent steps to prepare the molecule for simulation were conducted in Gromacs version 2019 as described above, however, system solvation and neutralisation were performed with WT4 and NaW/ClW molecules respectively (Machado *et al.*, 2019).

2.14.5 Protein in Buffer and Excipient

Parameterisation of excipient and buffer was conducted by uploading respective .mol2 files to CGenff (https://cgenff.umaryland.edu/initguess/) and downloading the output .str file, provided the penalty score was below 50. The CHARMM .py script “cgenff_charmm2gmx_py3.py” was downloaded (http://mackerell.umaryland.edu/charmm_ff.shtml#gromacs) and used along with the CHARMM36 ff (http://mackerell.umaryland.edu/charmm_ff.shtml) to build the topology of excipient and buffer files using their .mol2 and .str files:

```
python cgenff_charmm2gmx_py3.py SOR Sorbitol.mol2 Sorb.str
charmm36-mar2019.ff
```

To complete this step, numpy and networkx (v 1.11) were installed. Topology of the protein was constructed as described above (using the CHARMM36 ff). The position restraints for buffer and excipient, generated using the gmx genrestr command, could then be added to the protein .itp file. Following this, the protein was centered in a dodecahedron box with at least 1.0 nm from the box edge, and excipient and buffer molecules were added using the gmx insert-molecules command. The number of molecules to insert was calculate based on the volume of the system using the equation:

$$Moles = Concentration\ (M) \times Volume\ (L)$$

$$Moles = \frac{Number\ of\ molecules}{Avagadro's\ number}$$

$$Number\ of\ molecules = Concentration\ (M) \times Volume\ (L) \times Avagadro's\ Number$$ **Eqn. 8**

Therefore, the final system topology was adjusted to reflect the addition of buffer and excipient by adding the following:

```
; Include Sorbitol/ACE Topology
#include "SOR/ACE.itp"
#ifdef POSRES_SOR/ACE
#include "POSRES_SOR/ACE.itp"
#endif

; Include Sorbitol/ACE parameters
#include "sor/ACE.prm"

[ molecules ]
; Compound          #mols
Protein_chain_A       1
ACE                  2
SOR                 8
```

10mM sodium acetate was used for all simulations. System solvation and the following steps were conducted as previously described.

2.14.6 MD Analysis

2.14.6.1 Dynamics

Periodicity was accounted for after the MD run using the gmx trjconv command. Radius of gyration (Rg) and root-mean-square-deviation/–fluctuation (RMSD/RMSF) were calculated using gmx gyrate, rms and rmsf commands respectively. All analysis was done using the protein backbone.

2.14.6.2 Bio3D

Converting the .xtc simulation file to a .dcd file permitted the analysis of simulations in Bio3D. This conversion was performed by executing the catdcd file (http://www.ks.uiuc.edu/Development/MDTools/catdcd/license.cgi?files/catdcd-4.0b.tar.gz) and using the stride command to record every 100 frames (nanosecond) of the original simulation so that the file size was not too large. These frames were superimposed onto the α-carbon positions of the PDB:2D9Q using the commands:

```
ca.inds <- atom.select(pdb, elety="CA")
xyz <- fit.xyz(fixed=pdb$xyz,
mobile=dcd,
       fixed.inds=ca.inds$xyz,
       mobile.inds=ca.inds$xyz)
```

Dynamic cross-correlation maps (DCCMs) were produced from this superposition using the commands:

```
cij<-dccm(xyz[,ca.inds$xyz])
plot(cij)
```

Principal component analysis (PCA) was also conducted using this superposition with the following commands:

```
pc <- pca.xyz(xyz[,ca.inds$xyz])
plot(pc, col=bwr.colors(nrow(xyz)) )
```

Normal mode analysis (NMA) was performed by executing the nma command (setting rm.gaps to false) after superimposing the frames 10 and 90 with the pdbaln command. Contact mapping did not require superposition and instead, after the α-carbon selection step, the cmap command was executed. The torsion.pdb command calculated torsion angles.

2.14.6.3 iGEMDOCK

After setting the protonation state for PDB:2D9Q and paramterising the excipient, molecular docking was run with iGEMDOCK. Docking was set to accurate with a population size of 800 for 80 generations and 10 solutions (Yang and Shen, 2005).

2.14.6.4 Coevolution

The CoeViz (Baker and Porollo, 2016) application in POLYVIEW-2D, accessed via the SABLE server, calculated coevolution for G-CSF residues. These coevolution scores were computed from the multiple sequence alignment using Pearson correlation, which was weighted by phylogeny background (described in Baker and Porollo, 2016).

2.14.6.5 Calculations for Excipient Analysis

Radial Distribution Frequency (RDF) was calculated by creating an index file, selecting the oxygen atoms of all water molecules, which was used to reference the hydration shells in with the gmx rdf command. Distance distribution between protein and excipient was determined by calculating the minimal distances between residues and excipient molecules over the simulation using the gmx pairdist command. Residue interaction probabilities was then derived by counting (with the Excel 2016 countif function) the number of distances below 0.35 nm (for hydrogen bonding) and 0.6 nm (for VDW interactions) and dividing this by the total number of distance data points for each residue.

Chapter 3: NMR Reveals Thermally-induced Changes Pivotal to the G-CSF Structural-function Relationship

Little is reported on major conformational changes in G-CSF that are significant to bioactivity or stability/aggregation. However, a G-CSF aggregation mechanism has been proposed in which a highly reactive and structurally perturbed monomer functions as an aggregation seed (Krishnan *et al*., 2002). This perturbation was suggested to be in loop AB of G-CSF by Raso *et al*., based on a change in intrinsic fluorescence and the location of tryptophan residue W58. The aggregation of G-CSF is potentially rate-limited by conformational stability (Raso *et al*., 2005; Robinson *et al*., 2018) consistent with such an aggregation-prone intermediate state. Peptide-level hydrogen-deuterium exchange mass spectrometry (HDX-MS) recently also confirmed the sensitivity of aggregation rates and thermal stability upon mutation or formulation, to changes the exchange rates of residues within loop AB, loop CD, and the beginning of loop BC (Wood *et al*., 2020; Wood *et al*., 2022). Identifying specific residues that instigate or are directly affected by significant structural changes like this, in response to mutations or thermal perturbations, could reveal important structural features and mechanisms that affect function and stability, and thus also guide future rational/semi-rational protein engineering.

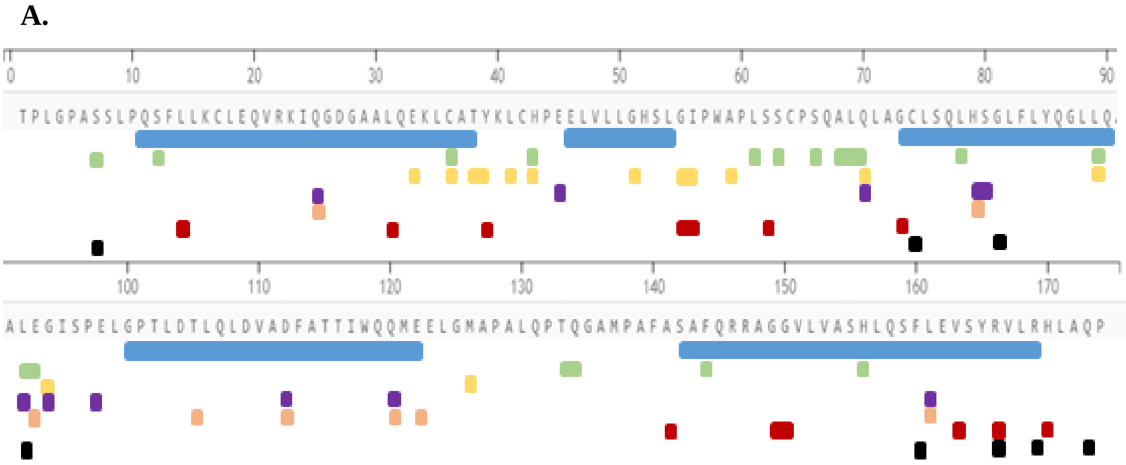
High-resolution insights on the residue-level dynamics over a range of native temperatures would provide valuable insights into key structural changes within the native ensemble that may be relevant to both function and the propensity to denature or aggregate. An NMR HSQC maps the microchemical environment of protein residues with chemical shift peaks. Therefore, changes in these peak positions signify changes in residue microchemical environment. Observing residue-level NMR chemical shift and peak intensity changes over a range of temperatures from 295 K to 323 K, this chapter explores the changes that individual residues in G-CSF experience through the early stages of thermal denaturation prior to the global transition. The peak intensity of signals in NMR typically represent the population of a species in the solution, e.g. the more G-CSF molecules are in a particular conformation, the higher the observed peak intensities (Kleckner and Foster, 2011; Dong *et al*., 2017). Additionally, dynamics can influence peak intensity and there exists a plethora of NMR experiments to probe protein dynamics (Igumenova, Frederick and Wand, 2006; Lakomek *et al*., 2008; Zeeb and Balbach, 2004). Higher residue mobility decreases R_2 ($1/T_2$) relaxation rates and increases peak intensity (Caulkins *et al*., 2018; Palmer III, 1997; Viles *et al*., 2001). This chapter attempts to identify dynamic residues in G-CSF by collectively accounting for their change in microchemical environment and peak intensities.

Using this approach, it was possible to resolve key events during the earlier thermal ramping towards the global transition temperature. This identified high-priority residues as potential targets for mutagenesis based on the significant changes they experience both locally and far in space. Structural changes were also identified within loop AB that supports previous observations that this loop can conformationally rearrange to form an aggregation-prone state (Raso *et al.*, 2005; Wood *et al.*, 2020; Wood *et al.*, 2022). Finally, subtle conformational changes were revealed in binding site III residues that may be significant in pre-organising the active site for receptor binding. Involvement of a key histidine residue suggests a pH-dependence that may adapt G-CSF activity within the lower-pH long bone marrow where it acts *in vivo* (Nikolaeva, 2018).

3.1 Results

3.1.1 Assigning G-CSF 2D ^{15}N-^{1}H HSQC Spectra at Different Temperatures

From the 2D ^{15}N-^{1}H HSQC spectra of 0.09 mM wild-type G-CSF in 50 mM sodium acetate, pH 4.25 (Figure 7b), I was able to assign a maximum of 115 peaks out of the 160 assignable peaks published by Zink *et al* using CcpNmr Analysis 2.4.2 (Vranken *et al*., 2005; Zink *et al*., 1994). I applied a thermal ramp from 295 K to 323 K, which was just below the thermal melting transition temperature, and recorded spectra at every 2 K to track the movement of peaks by measuring the changes in their chemical shift positions (Δδ). This allowed us to monitor residue environmental changes (including partial unfolding events and conformational transitions) up until the point of global unfolding. The number of assignable peaks decreased from 115 at 295 K, to 106 at the final temperature of 323 K, where some peaks became co-incident with others while others disappeared altogether. All NMR experiments in this study were conducted in singlicate, which can limit the conclusions to be made solely from NMR data. However, as this study will show, significant conformational changes are consistent across different formulations (Figure 17, 21 and 29). A sequence map of G-CSF with significant residues referenced throughout the study is highlighted in Figure 7a.


A.
0
10
20
30
40
50
60
70
80
90
TPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLSQLHSGLFLYQGLLQ
100
110
120
130
140
150
160
170
ALEGISPELGPTLDTLQLDVADFATTIWQQMEELGMAPALQPTQGAMPAFASAFQRRAGGVLVASHLQSFLEVSYRVLRHLAQP

🟦 Secondary Structure

🟩 90th percentile of ∑Δδ

🟨 Top 15 residues with the highest percentage change in PI

🟪 Excipient-Induced Δδ for 25mM Histidine (90th percentile)

🟧 Excipient-Induced Δδ for 25mM Arginine (90th percentile)

🟥 WT peak trajectory linearity (linearity below 0.5)

⬛ Conserved residues (95th percentile for self-correlation)

B.

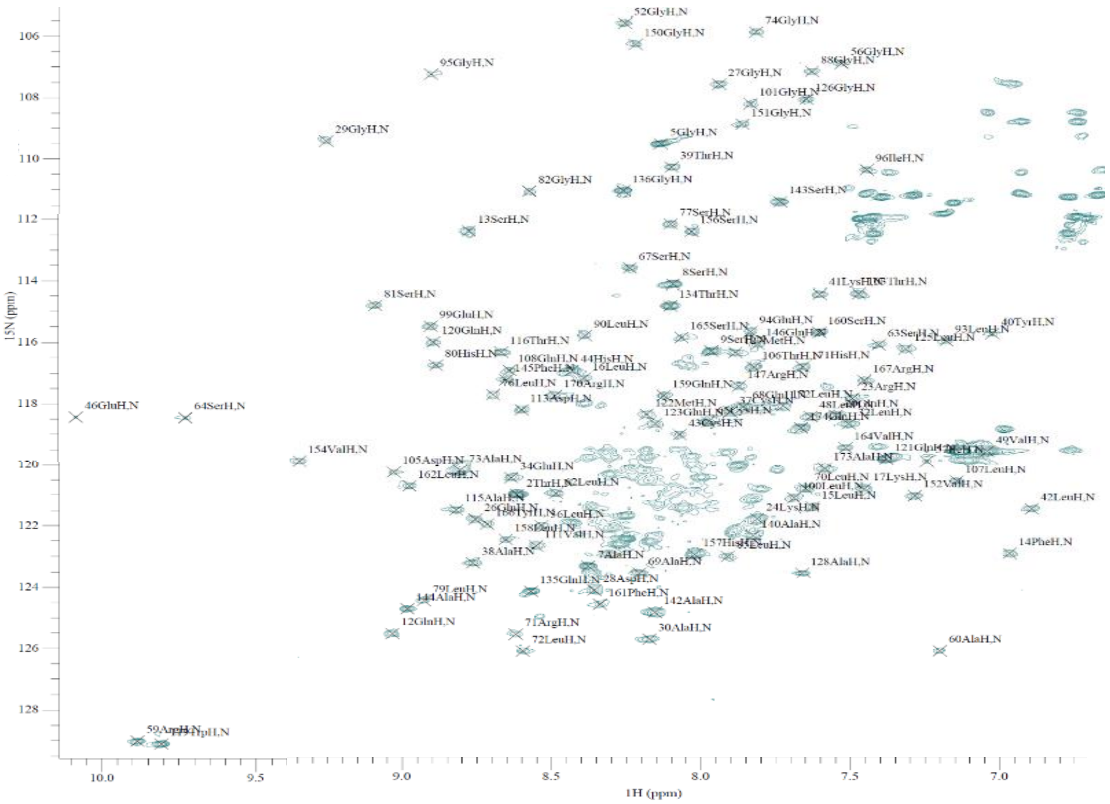


Figure 7. G-CSF Sequence and Assigned ^{15}N-^{1}H HSQC. **A.** The colour-coded bars (legend is below the sequence) indicate which residues belong to the respective category. **B.** The ^{15}N-^{1}H HSQC spectrum of 0.09 mM WT G-CSF (pH 4.25, 50 mM sodium acetate) collected at 303 K (all residue numbers are shifted by +1) and assigned in CcpNmr Analysis 2.4.2 (Vranken et al., 2005).

The chemical-shift distances travelled by peaks during the thermal melt were measured and the trajectories characterised as linear or non-linear (defined in Table S.1). A typical example of a linear peak maxima trajectory during the thermal melt (295K to 305K) is shown in Figure 8a for residue Q134. In total, 68 residues had linear trajectories. By comparison, 44 residues had non-linear peak trajectories over the thermal melt such as that in Figure 8b, which indicated a more complex pathway in their change in microenvironment, with intermediate conformations being populated. There is a concentration of some of the most non-linear trajectories around the C-terminus of helix D (V163, R166, H170) and the proximal loop AB residues S62 and G73, the significance of which is later discussed. Some signals could not be assigned throughout the entire temperature range. For example, the peak from residue E45 disappeared at 303 K and above. The signals from other residues, namely Q67, M126, E93, G87 and S155, were lost after appearing in the same position as a signal for another residue experiencing the same microchemical environment at that temperature. Residues with more than three temperature points missing were not included in further analysis.

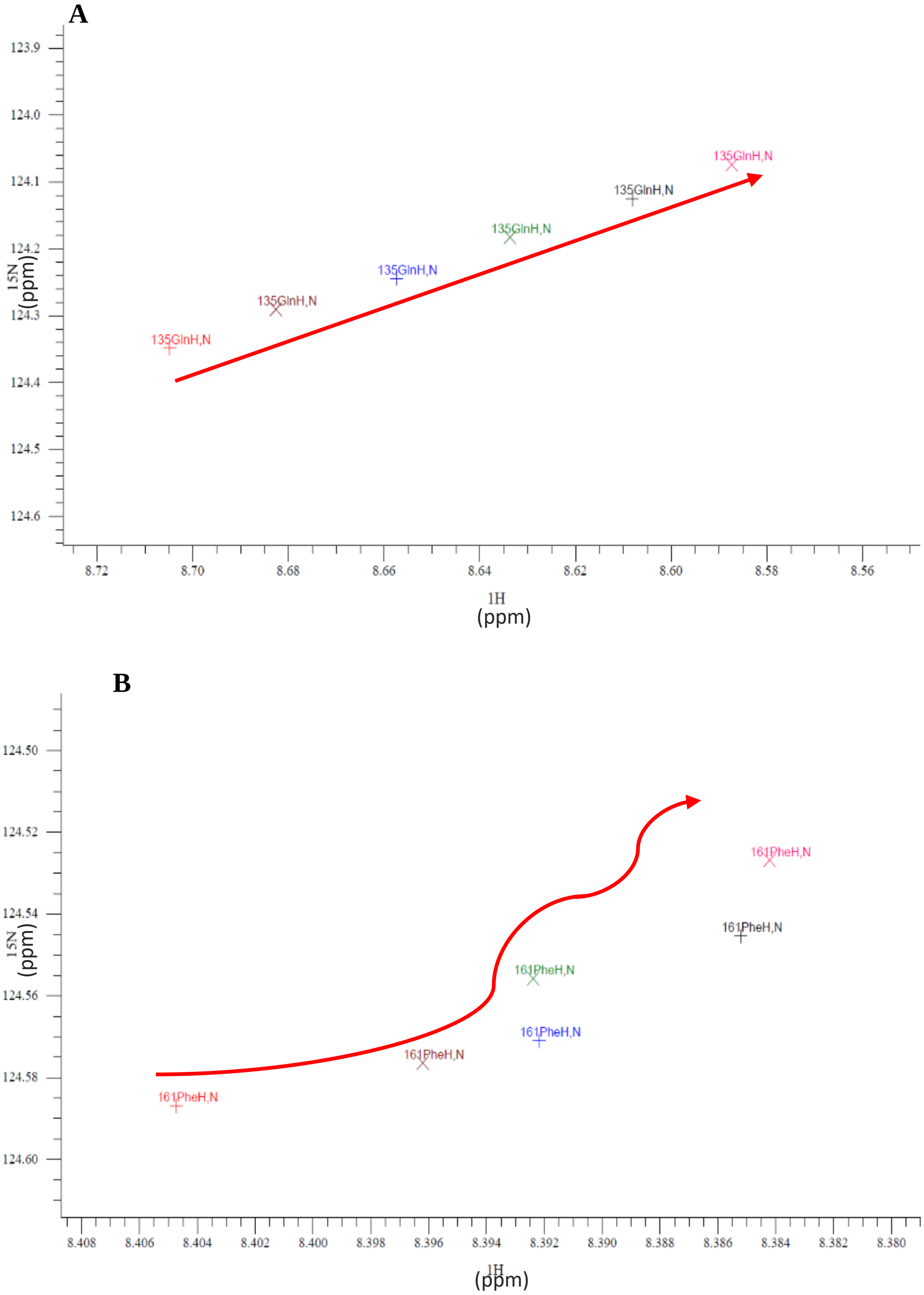


Figure 8. Residue Peak Migration Over the Thermal Melt. Each cross in **A.** and **B.** represents the maxima from the peaks for residues Q134 and F160, respectively, from 295 K (represented

with the black cross) to 305 K (represented with the pink cross). The red arrows indicate the trajectory of these maxima during the thermal melt.

3.1.2 Tracking the Cumulative Change in Residue Microchemical Environment

The peaks from different residues also moved at various rates and to varied extents, as shown with examples in Figure 8. To highlight this, a cumulative change in chemical shift (∑Δδ) was calculated for each residue as a function of temperature, starting from ∑Δδ = 0 Δδ at 295 K, as shown in Figure S.1.

The vast majority of residues had a linear "∑Δδ–temperature relationship" for the entire thermal ramp (Figure S.2), indicating a gradual rise in thermally-induced mobility throughout the structure, but with no clear conformational changes in local structure. Their ∑Δδ values were normally distributed (Figure 16), where the distribution broadened with increasing temperature, although the majority of residues remained with a total ∑Δδ of <0.2 Δppm even at 323 K. This indicates that the changes in microenvironment across the majority of the residues were mostly related to gradually increasing mobility in the native ensemble, with increasing temperature. On the other hand, a few key residues underwent significantly larger changes, surpassing the threshold of ∑Δδ (at 323K) = 0.2 Δppm by at least 50%, indicative of residues with microenvironments much more susceptible to temperature than the majority.

Table 4a highlights residues ranked in the 90th percentile according to their ∑Δδ at each temperature. Residue Q70 (the top grey line in Figure S.2) clearly ranked highest by ∑Δδ over the whole thermal melt except at 297 K. The relationship between residue-level ∑Δδ and location in the crystal structure of G-CSF (Protein Data Bank ID code 2D9Q: Tamada *et al.*, 2006) was also visualised by highlighting residues in the 90th percentile in Figure 9a. A more detailed colour-mapping of ∑Δδ values for all residues, and at each temperature is available in the supplementary information (Figure S.3). This shows the gradual increase in ∑Δδ for most residues as temperature increases, but also highlights the positions of the residues that had stronger responses to temperature, which can be easily seen as early as 307 K.

Residues in the 90th percentile were mainly in loop AB, aside from residues H156, A37, L89, L78, F144 and E45, found in structural clusters formed from parts of helix A, B, D and the short helix (Figure S.3). This suggests that there were three or four localised regions of structure susceptible to conformational change or partial unfolding at temperatures lower than for global unfolding. This will be discussed after further analysis below.

Interestingly, for a few residues (namely G55, I56, A59, S63 and F144), the ∑Δδ–temperature relationship was not entirely linear and a minor transition occurred at approximately 305/307 K, as visualised in Figures S.6 and S.7. This transition represents a minor conformational rearrangement clustered within the first half of loop AB and its interactions with helix D, which is adjacent to the GCSF-R binding site III (Tamada *et al*., 2006). Notably, residue S63 climbs Table 4a rapidly at above 305 K as it experienced larger changes in its microenvironment during the localised conformational transition discussed above.

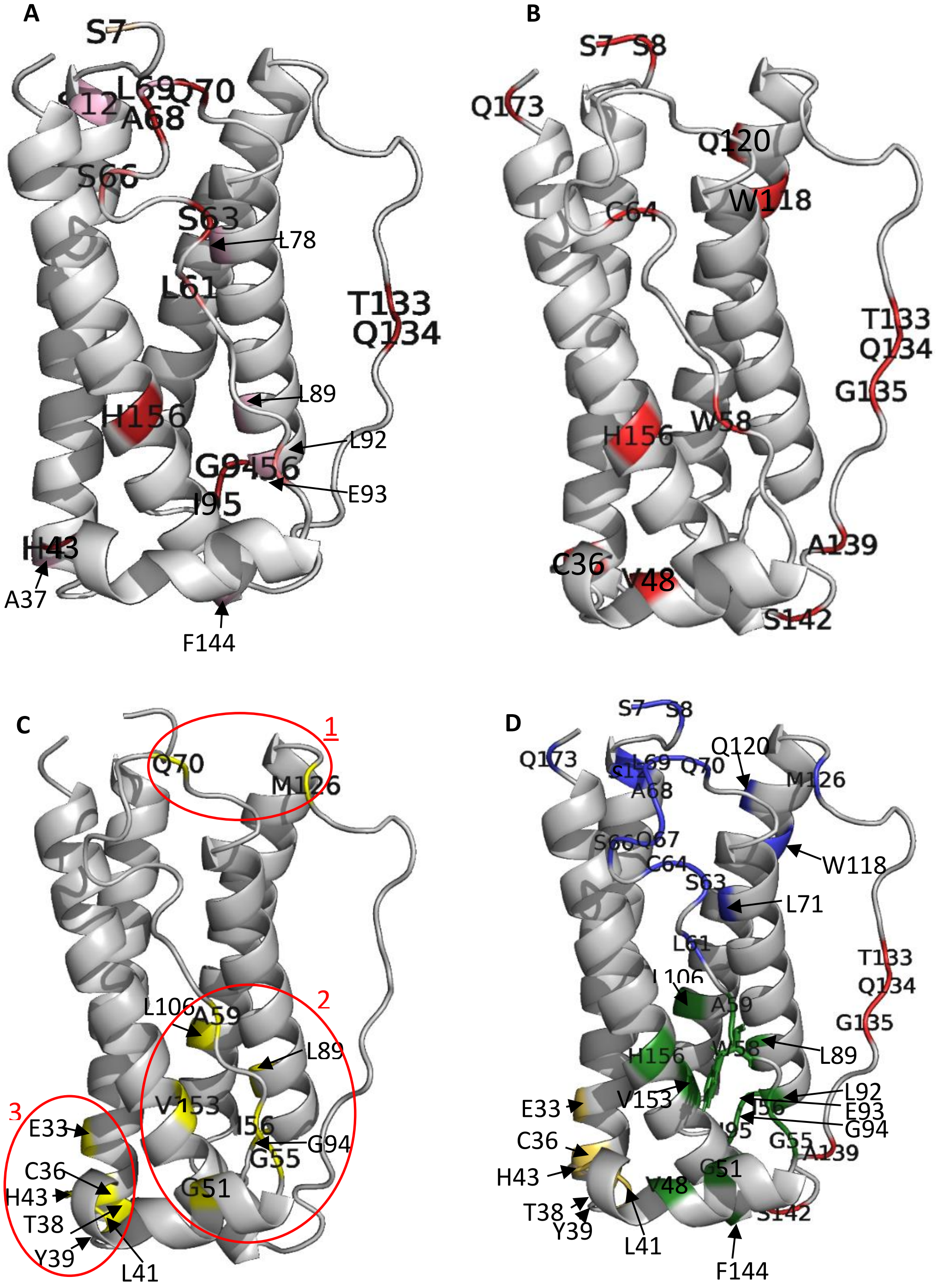
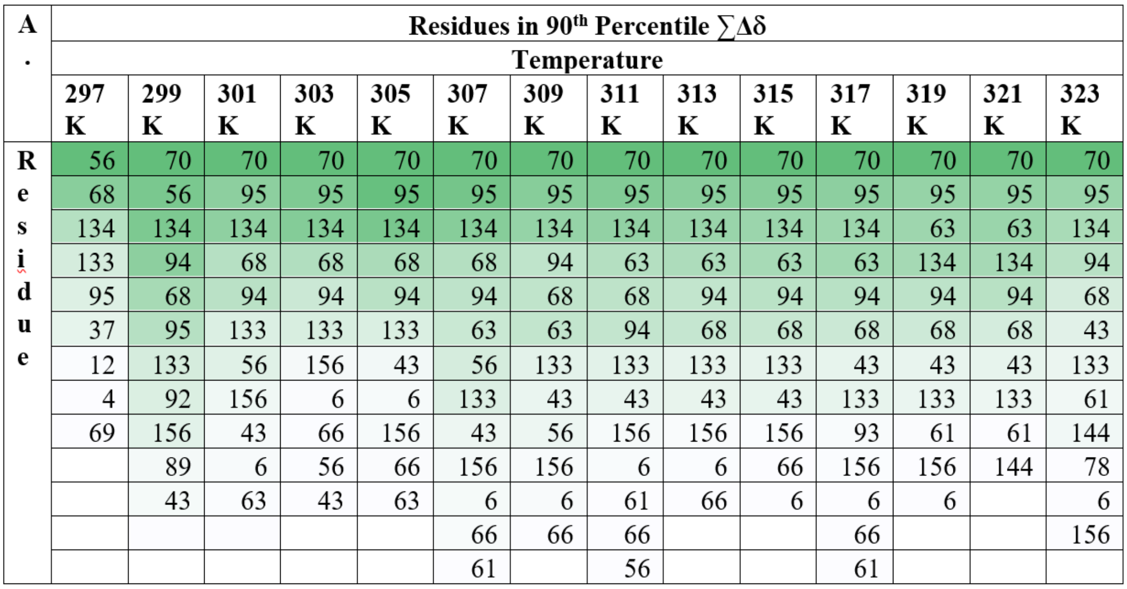
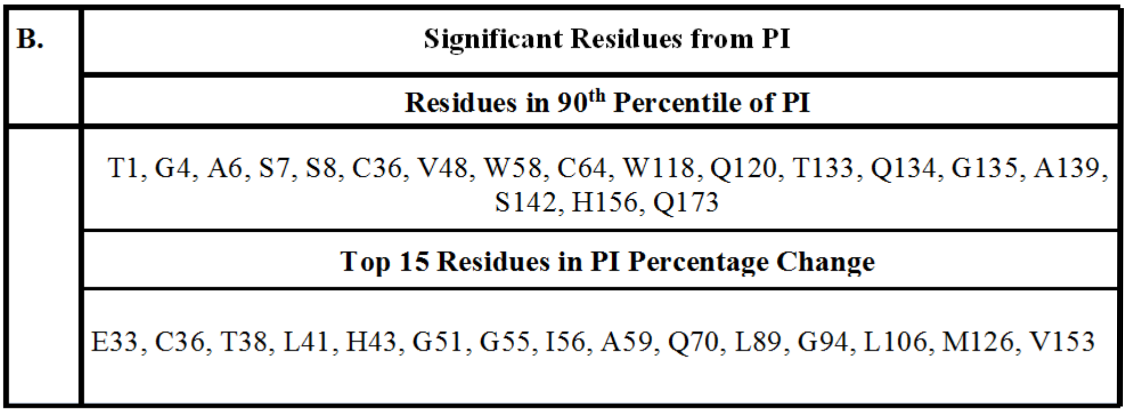


Figure 9. Mapping NMR Observables onto G-CSF. Each observable parameter is mapped onto the G-CSF crystal structure. **A** Residues in the 90th percentile of ∑Δδ Residues that appeared in the 90th percentile for up to two of the 14 temperatures are coloured pink, salmon residues appeared at 3 to 9 temperatures, and deep red residues appeared for at least ten temperatures. In **B**, all 90th percentile residues for PI (absolute change) are coloured red. In **C**, the 15 residues showing the highest percentage increase in PI over the temperature range are coloured yellow. These form 3 sub-clusters, circled red. **D** combines all residues in **A-C** to reveal four final structural clusters. Structural cluster 1 is blue, 2 is green, 3 is yellow and 4 is red. Residue W58, assigned previously to observed hyper-fluorescence, is shown as sticks. PDB:2D9Q is missing its first 6 residues, therefore S7 is highlighted in place of earlier residues.



A.	Residues in 90th Percentile ∑Δδ													
	Temperature													
	297 K	299 K	301 K	303 K	305 K	307 K	309 K	311 K	313 K	315 K	317 K	319 K	321 K	323 K
Residue	56	70	70	70	70	70	70	70	70	70	70	70	70	70
	68	56	95	95	95	95	95	95	95	95	95	95	95	95
	134	134	134	134	134	134	134	134	134	134	134	63	63	134
	133	94	68	68	68	68	94	63	63	63	63	134	134	94
	95	68	94	94	94	94	68	68	94	94	94	94	94	68
	37	95	133	133	133	63	63	94	68	68	68	68	68	43
	12	133	56	156	43	56	133	133	133	133	43	43	43	133
	4	92	156	6	6	133	43	43	43	43	133	133	133	61
	69	156	43	66	156	43	56	156	156	156	93	61	61	144
		89	6	56	66	156	156	6	6	66	156	156	144	78
		43	63	43	63	6	6	61	66	6	6	6		6
						66	66	66			66			156
						61		56			61			



B.	Significant Residues from PI
	Residues in 90th Percentile of PI
	T1, G4, A6, S7, S8, C36, V48, W58, C64, W118, Q120, T133, Q134, G135, A139, S142, H156, Q173
	Top 15 Residues in PI Percentage Change
	E33, C36, T38, L41, H43, G51, G55, I56, A59, Q70, L89, G94, L106, M126, V153

Table 4. A. Residues in the 90th percentile of the ∑Δδ normal distribution at each temperature point, highlighted on a green scale, with darker green representing higher ∑Δδ. **B.** A list of significant residues determined from PI. Residues with a maximum PI value (typically around 305K) that are in the 90th percentile are shown in the top half of the table. The top 15 residues with the highest percentage change in PI are shown in the bottom half.

3.1.3 Variation of Residue Signal Peak Intensities with Temperature

Over the thermal ramp, the peak intensities (PI) for all residues, in general, increased with temperature up to ~305 K and plateaued before decreasing at ~313K and above (Figure S.4) and could be fitted with a second order polynomial curve for all residues. This general trend relates to the gradual increase in dynamics, and hence PI, as the temperature increases, but then as the protein begins to unfold and aggregate, the PIs decrease, leading towards zero PI as the thermal denaturation midpoint is approached (at approximately 323 K). Residues T1 to S8, V48 and Q120 are an exception to this trend as they plateaued much earlier, perhaps signaling internal rearrangement or high dynamics while the protein was in a low energised state. Residues in the 90th percentile of PI were determined as those passing the upper 10% threshold for the total distribution data for all PI (eqn.5). Residues with a maximum PI above this threshold were classed as 90th percentile (indicated in Table 4b, Figure 9b and Figure 10a with red bars). PI is strongly influenced by dynamics such that residues in the 90th percentile can be classed as relatively dynamic.

Given that both sample concentration and temperature can influence PI, I could, at least in part, be observing a general 2nd order polynomial curve for PI due to the influence of these factors (Shaoxiong Wu, 2011; Zhang, Powers and O’Day, 2020). The initial increase in PI could result from the rising temperature of the sample, which in turn increases bulk magnetization (Shaoxiong Wu, 2011). Additionally, over a thermal melt, certain local conformations within the protein can become dominant. Consequently, this would cause the PI of these residues to increase over the melt, given that PI reflects the number of nuclei resonating at a given frequency (experiencing the same micro-chemical environment; Kleckner and Foster, 2011). Global increase in mobility would also cause this initially increase in PI. An interplay between all mentioned factors is equally likely. The decrease in peak intensity towards the end of the thermal melt could result from loss of sample through unfolding or aggregation at around 321K (323K is where signals for the majority of residues are lost with NMR) (Wood, 2020). Nevertheless, all residues may not follow these global trends and may hold key information to their role in global stability.

In addition to the maximum PI, I aimed to identify residues that underwent significant changes in dynamics during thermal denaturation. These large dynamic changes could result from local unfolding events, or conformational switching within the native ensemble with relevance to stability or function. The percentage increase in PI (eqn.6) was calculated for all assigned residues (Figure 10a). Therefore, unlike the 90th percentile for PI, percentage increase in PI refers to the largest change in dynamics experienced by a given residue. Residues that were highly dynamic already at low temperatures, such as at the N-terminus (T1 to S8) tended to give low (7%-25%) increases in PI. However, some residues gave large increases in dynamics, such as G51 which experienced a 297% increase in PI. Many residues with a high percentage increase had low maximum PI values, and so had a low absolute change in PI. However, residues C36 and H43 have both high maximum PIs and a high percentage change (Figure 10b), indicating significant changes in dynamics for these residues. The vast majority of the top 15 residues experiencing the highest percentage increase in PI (also highlighted in Table 4b and in yellow in Table S.3) were generally clustered at the receptor-binding end (site III) of the protein structure, forming the sub-clusters 2 and 3 shown in (Figure 9c). This is also particularly emphasised in the split half-way through loop AB in which the N-terminal residues had large percentage increases in PI, compared to the C-terminal half of loop AB which gave high maximum PI values.

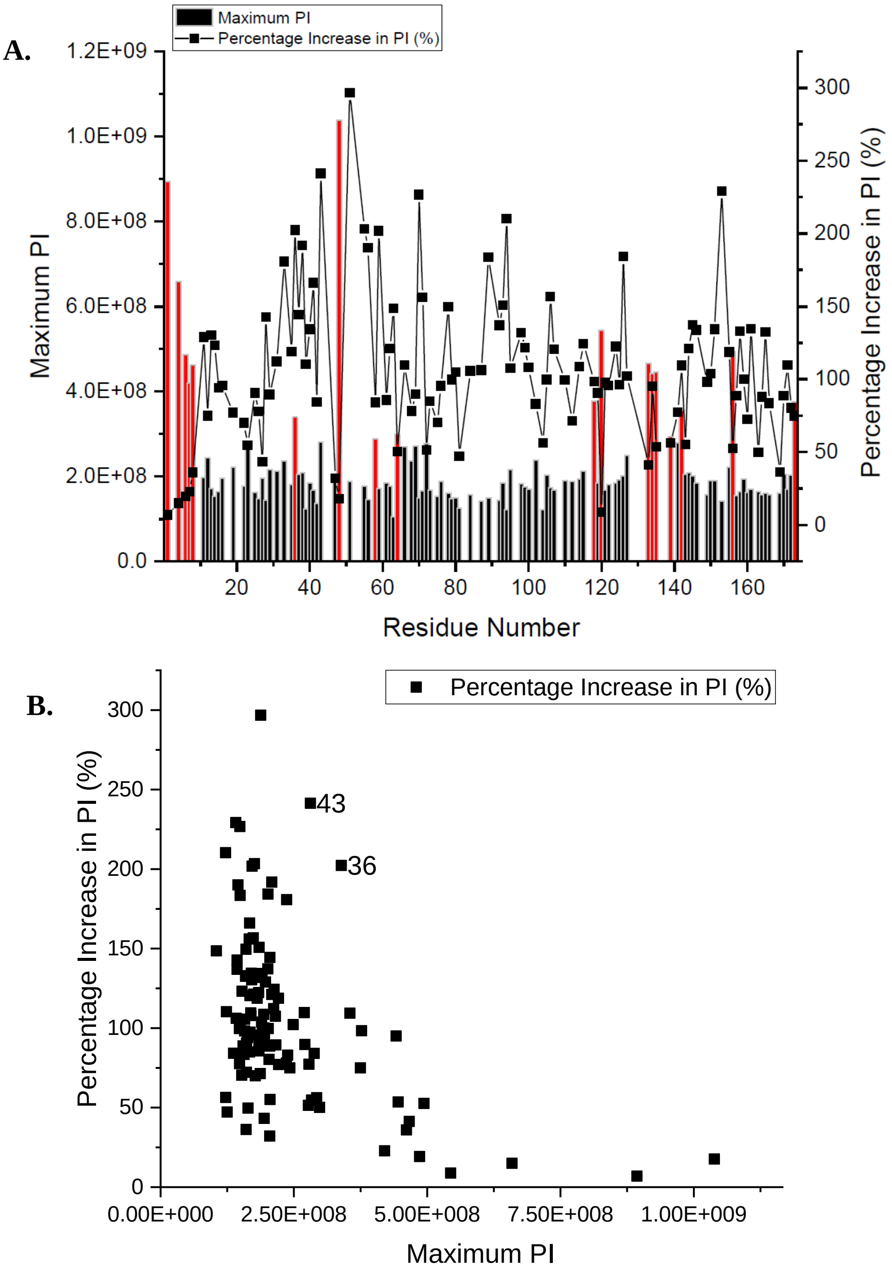


Figure 10. Maximum PI and Percentage Increase in PI. A. Maximum PI values for all residues (bars), with residues above the 90th percentile threshold for maximum PI highlighted red. The

black line indicates percentage increase in PI over the melt. **B.** Maximum PI vs percentage increase in PI (with key residues labelled).

There was considerable overlap between the residues in the 90th percentiles of $\sum\Delta\delta$, PI and %PI (Table 4), and hence also for their structural locations (Figures 9 a-c). Figure 9d combines the residues highlighted by each measure, and clearly shows that they form four structural clusters. The first is formed along the C-terminal half of loop AB (residues L61, S63, C64, S66, A68, L69, Q70 and L78), the C-terminus (Q173), the N-terminus (T1, G4, A6, S7, S8, S12) and the beginning of loop CD/end of helix C (W118, Q120 and M126). The second structural cluster spans the N-terminal half of loop AB (G55, I56, W58, A59) with some interacting residues from the short helix (V48, G51), helix D (F144, V153, H156), helix C residue L106 and loop CD residues (L89 and L92-I95). From the residues involved, this structural cluster appears to form a large hydrophobic core in which residues also have a low solvent accessibility (Figure S.8). Therefore, given that an increase in solvent accessibility can increase transfer of magnetization to solvent (thereby increasing PI), structural cluster 2 could be experiencing an expanding motion, making it more solvent accessible.

The third structural cluster resides in GCSF-R binding site III, with helix A and nearby short helix residues (E33, C36, A37, T38, L41 and H43). Finally, the fourth structural cluster is centred on loop CD residues T133, Q134, G135, A139 and S142 at the end of loop CD, near to the short helix. Clearly, these structural clusters overlap in some regions.

As discussed above, the large changes in microenvironment or dynamics for these residues in localised structural clusters, indicates localised conformational changes or partial unfolding, at low temperature (from 305 K) prior to any global unfolding or aggregation which begins (>1% unfolded) at approximately 320 K (Robinson *et al*., 2018). The focus around loop AB is consistent with previous work implicating this region in a conformational shift to form an aggregation-prone G-CSF intermediate (Raso *et al*., 2005; Ko *et al*., 2022). More recent HDX-MS studies on G-CSF formulations containing mannitol, phenylalanine or sucrose (Wood *et al*., 2020), and on single-mutant variants of G-CSF (Wood *et al*., 2022) have confirmed the role of loop AB. In addition, they revealed changes in dynamics within the short helix, loop CD and part of helix D, that correlated with aggregation propensity and the thermal melting temperatures (T_m). The NMR data, showing structural clusters 1, 2 and 3 encompassing loop AB, and structural cluster 4 within loop CD, fully supports a conformational change localised in these same regions, that is promoted through the moderate temperature increase to 307 K. The largest aggregation-prone region (APR), identified by the consensus method employed in Figure S.10 (assisted by my undergraduate student, Jinhui Kim), spans helix D. Conformational changes around loop AB have a strong potential to expose this APR. Given the non-linear trajectory for residues clustered around the N-terminus of helix D (V163, R166, H170) and proximal C-terminus of loop AB (S62 and G73), conformational changes in these regions could result from multiple states being occupied in loop AB before helix D is exposed.

A notable feature of the structural clusters identified by NMR, is that none of them are directly involved in the major binding site (site II) containing residues K16, G19, Q20, R22, K23, L108, D109 and D112. Thus, the structural rearrangements identified as the temperature is increased would not necessarily affect the integrity and function of binding site II. However, the minor binding site (site III) appears to be directly impacted, suggesting that this site can be conformationally switched on or off. Indeed, the significant distortion of loop AB may be necessary to elicit a conformational change in binding site III for receptor interaction, as eluded to in Figures S.7 and S.9. It appears, therefore, that the structural change identified as making G-CSF more aggregation-prone *in vitro*, is the same or similar to the structural change observed at 307 K *in vitro*. It is also very possible that the higher temperature structure is the functionally relevant state *in vivo* at 37 °C (310 K), although the difference in pH from this work at pH 4.25, and physiological pH of 6.7-6.9 in long bone marrow, would also likely have an influence (Nikolaeva, 2018).

3.1.4 Probing Correlations in $\Delta\delta$ and PI

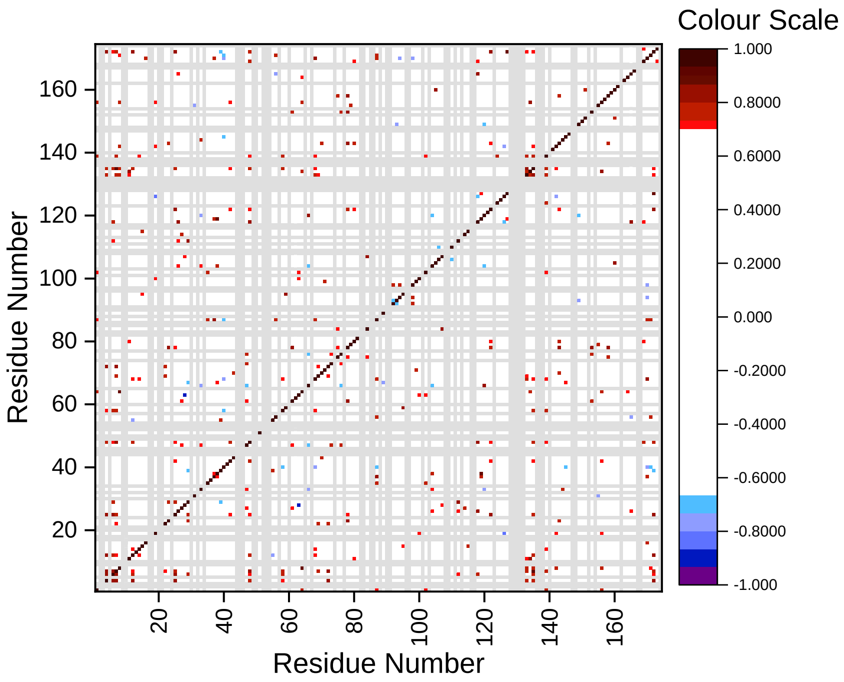
Although $\sum\Delta\delta$, PI and %PI indicated which residues were undergoing the most change under "native" conditions prior to the global thermal melt, this did not reveal how the movements in each residue related to the others, beyond simply co-locating them in structural clusters. Correlation analysis between residues could determine whether the changes in residues or the structural clusters are directly coupled during the thermal denaturation. Figure 11a shows a cross-correlation matrix (CCM) for the temperature-dependent $\Delta\delta$ of all residues in the $\sum\Delta\delta$ 90th percentile over the entire temperature range studied. Figure 11b shows a similar CCM for residues in the 90th percentile of PI, but correlating across all of their PI values at respective temperatures.

Correlations were determined in each case using the Spearmanʹs correlation coefficient (eqn.7). For $\Delta\delta$, each coefficient value was calculated between a pair of residues' $\Delta\delta$ values from consecutive temperature points over the thermal melt. Using the example of residue 90 in Figure S.1, residue Q90ʹs D_1, D_2, D_3 etc… would be correlated with residue E33ʹs D_1, D_2, D_3 etc…. In the colour scale for Figure 11a, the red shades represent positive correlations above 0.7 whilst the blue shades represent negative (anti-) correlations of less than -0.7. A white colour gate is used between 0.7 and -0.7 to remove noise, or correlations of low confidence. The diagonal black line occurs through the matrix where complete correlation occurs between the same residues. Most space in this matrix is white, indicating that most residues do not elicit strong correlations between their temperature-dependent fluctuations in Δδ. However, there are some clear regions of strong positive or negative correlation.

For PI, positive correlation values were generally higher than those for Δδ. Therefore, a higher colour gate (-0.95 to 0.95) was needed to reduce noise. Negative correlation did not occur below a value of -0.95

Interestingly, residues within loop AB itself did not tend to correlate strongly with each other in either matrix, suggesting that the overall change in conformation across the loop was not highly concerted or cooperative, but was probably more progressive with temperature. This is also reflected in the C-terminal end of loop AB having more 90th percentile PI residues (highly dynamic), but the N-terminal end having more 90th percentile %PI residues (large change in dynamics).

A. CCM for Δδ



B. CCM for PI

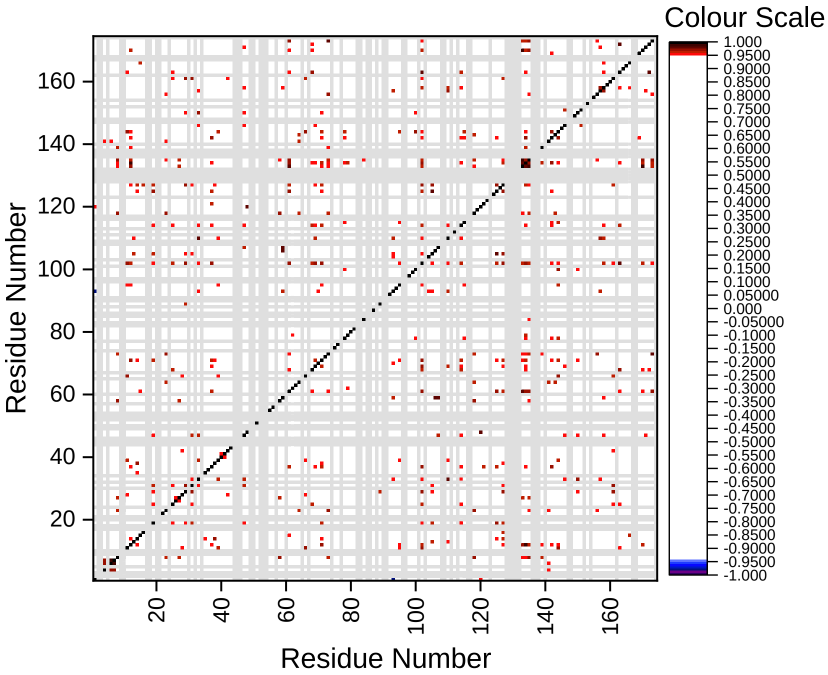


Figure 11. Cross-correlation matrices (CCM) for A. Δδ and **B**. PI. Spearman's correlation coefficient values are colour coded. In A, shades from red to black represent positive correlation (above 0.7) and shades from blue to purple represent negative correlation (below -0.7). A colouring gate of white was used between -0.7 and 0.7 to reduce noise. In B, shades from red to black represent positive correlation (above 0.95) and shades of blue represent negative correlation (below -0.95). A colouring gate of white was used between -0.95 and 0.95 to reduce noise. Unassigned residues have grey columns.

By contrast, several key clusters with strong correlations were observed in the two matrices. Some were formed within their local sequence (close to diagonals on matrices), such as at the N-terminus (G4, A6, S7 and S12), within loop CD (T133, Q134 and G135) and within loop BC (L92 and E93). However, the two regions in the N-terminus and loop CD also strongly correlated with each other despite being spatially distant (30.1 Å apart). A key linker between these regions appears to be residue W118 in helix C, which sits between them spatially, and has strong correlations with S8, T133 and G135 in PI.

The loop CD cluster (T133, Q134 and G135) was also correlated to regions of loop AB (S66, A68, L69) close to the N-terminal end of loop CD, and to H156 at the other end of loop CD, indicating increased dynamics in the centre of loop CD resulting from modified interactions with the ends of that loop. As the C-terminal end of loop AB (A68, L69) was also strongly correlated in Δδ to N-terminal residues (G4 and S12), this provides another structural link that could mediate the correlation between the N-terminus and loop CD. Thus overall, the structural changes in the

N-terminus, the C-terminal end of loop AB, the centre of loop CD and residue W118, appear to become modified in a concerted manner. The changes in the N-terminal end of loop AB is not in concert with this, but instead undergoes its own non-linear transition at ~305 K as seen for residues G55, I56, A59 and S63 (Table S.2B and Figure S.7).

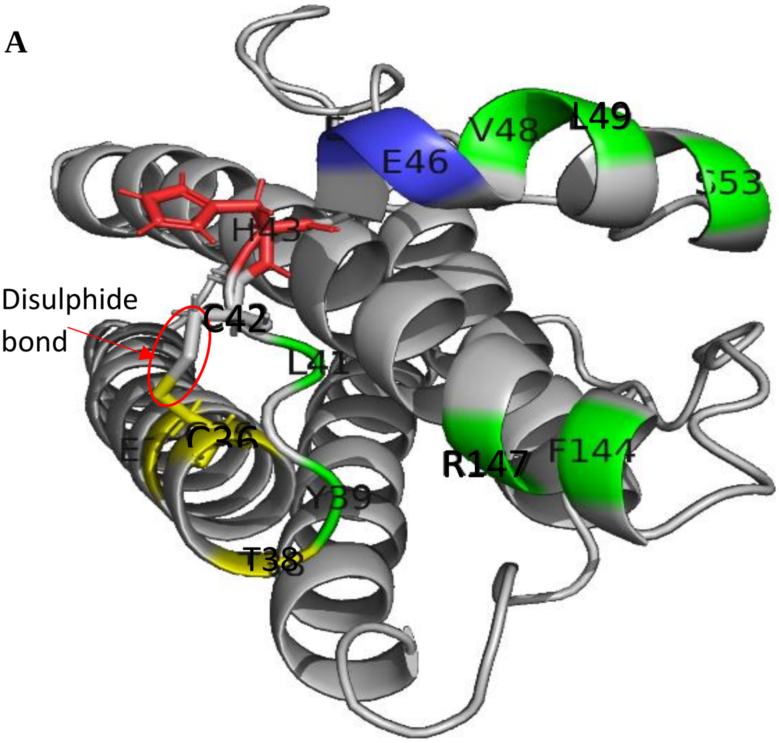
A few other individual residues show strong correlations, without forming clusters with local sequence. As such, while they may indicate coupled loss of interactions, their spatial separations and occurrence as individual residues suggests that they are more likely to be coincidentally undergoing similar changes in microenvironment.

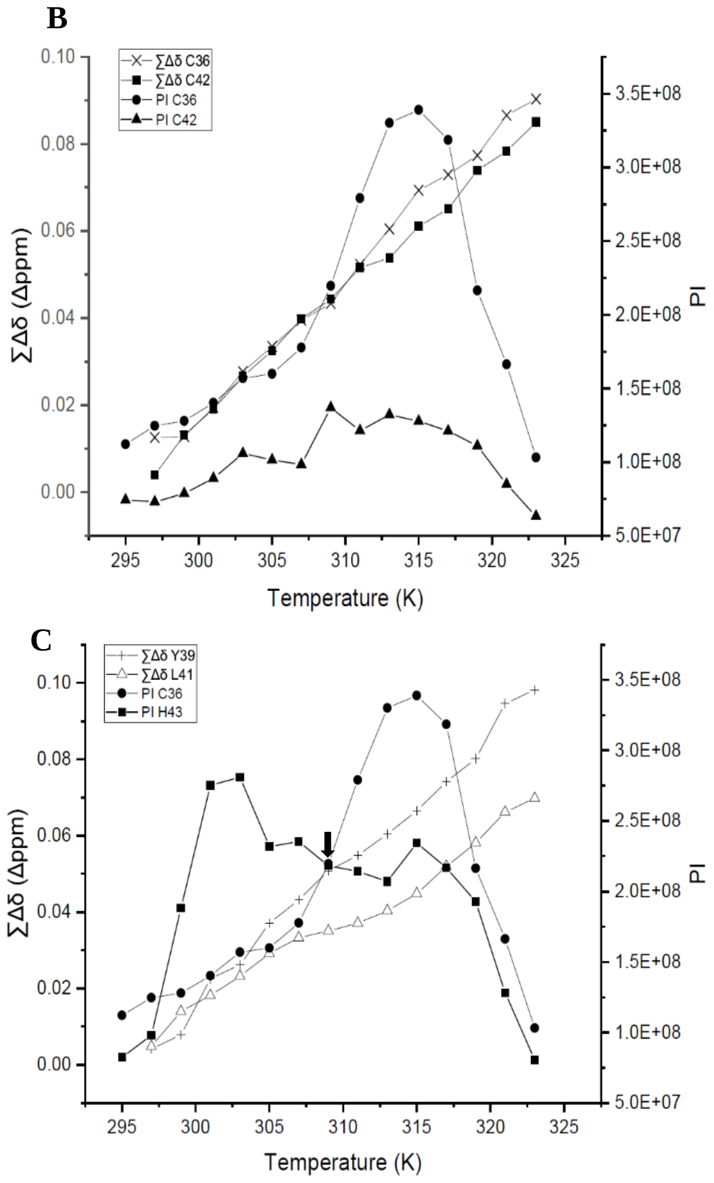
3.1.5 Characteristics of Structural Cluster 3 Reveals a Potential "Switch Mechanism"

Residue H43 displayed a notably high $\sum\Delta\delta$, maximum PI and percentage change in PI (Table 4 and Figure 10), and appears in structural cluster 3 (Figure 9d). From this cluster, residue C36 is disulphide bonded to C42, and while both residues showed a similar PI profile up to 307 K, there was a clear inflection point for C36 where its PI increased more rapidly at above 307 K and peaked at 315K (Figure 12b). PI for residue C42, on the other hand, slightly increased at 307 K as well but then decreased and stayed low after this. Similarly, $\sum\Delta\delta$ for residues C36 and C42 were similar up until 311 K, but then clearly differentiated at the same temperature as the large peak in PI for residue C36.

Residue H43 is adjacent to the disulphide bonded residue C42 (Figure 12a), and is also proximal to a very negatively charged area composed of E45 and E46 (highlighted blue). P44 is positioned such that it angles the sidechain of H43 towards these negatively charged residues. H43 also has the second highest percentage increase in PI (241%), visible as a distinct peak in PI at 303 K occurring immediately before the peak observed in PI for C36 in Figure 12c. The physiological temperature of ~309 K (indicated with a black arrow) occurs right at the transition point just after the decrease in PI for H43 and just before the peak in PI for C36. Therefore, H43 is well-placed to instigate a "switch mechanism", with attraction towards E45 and E46 placing a strain on the disulphide bond between C42 and C36, due to the pulling of the short loop containing H43. A significant PI increase was experienced by C36, and much less so for C42, because it is part of a structured α-helix with more restricted movement. The significant PI increase for C36, and decrease for H43 suggests a shift towards a new conformer with increased dynamics for C36, potentially also including breaking of the disulphide, and with decreased dynamics for H43 as it forms stronger interactions with E45 and E46. Of note, although processing of NMR observables for E45 is not shown due to more than three missing temperature points, the PI for E45 decreased simultaneously with the increase in PI for H43. This could signify increased conformational restraint on E45 as H43 interacts more with it.

Given that H43 is part of an unstructured loop between helix A and the neighbouring short helix region, the proposed "switch mechanism" would likely expose the loop region that also contains residues L41 and Y39 (highlighted green in Figure 12a). ∑Δδ of Y39 and L41 sharply increases during H43's PI maximum, with a slight slowing to that increase while C36 reached its PI maximum (Figure 12c). Both Y39 and L41 form part of the minor G-CSF receptor (GCSF-R) binding site III, highlighted green in Figure 12a (Tamada *et al*., 2006). Furthermore, they are amongst the most buried residues in both active sites for G-CSF, displaying a solvent accessible surface area (SASA) of 0.640 nm^2 and 0.099 nm^2 respectively (in Table 5 and Figure S.8) as determined using the online server ProtSA (Estrada *et al*., 2009). That makes L41 the most buried residue in both of G-CSF active sites and Y39 the fifth most buried (Table 5). Hence, exposure of these residues by a "switch mechanism" would have a significant impact on bioactivity. Although the thermal melt with WT G-CSF was not repeated in this study, Figure 29 shows the same general trend in NMR observables depicted in Figure 12b and C when this approach was done in the presence of various excipients.





Residue	SASA (nm^2)
Leu-41	0.099
Gln-20	0.160
Asp-109	0.161
Val-48	0.296
Tyr-39	0.640
Arg-147	0.662
Asp-112	0.779
Ser-53	0.937
Lys-16	1.049
Glu-19	1.064
Arg-22	1.082
Lys-23	1.128
Phe-144	1.136
Glu-46	1.140
Leu-49	1.203
Leu-108	1.223

Figure 12. The "Switch" Mechanism. A. Causal and beneficiary residues in the "switch" mechanism. The disulphide bond between C36 and C42 is indicated in the red circle. Structural cluster 3 residues are yellow, H43 is red, E45 and E46 are blue and residues in GCSF-R binding site III are green. **B** and **C** compare PI and ∑Δδ for C36, C42, Y39 and L41. A black arrow represents the point of physiological temperature (~309K) in **C**. **Table 5** shows the SASA of all of these residues in binding site III.

3.1.6 In silico structure relaxation supports NMR and a proposed "switch mechanism"

To further understand the observed changes in ∑Δδ and PI, the PDB:2D9Q structure of the receptor-bound G-CSF was relaxed using the online server Rosetta in the absence of the receptor. This would allow residues that are thermodynamically constrained in the bioactive receptor-bound state, to relax into conformations more favoured in the unbound state, and that could then be compared to the large changes observed for ∑Δδ and PI. The relaxed structure (coloured cyan) is compared with the unrelaxed PDB:2D9Q structure (coloured silver) in Figure 13.

Structural cluster 2 is the second largest of the four (Figure 9c). All of these residues, aside from G94 and A59, form a hydrophobic pocket with their side chains in close proximity (within 4.7 Å) to each other (Figure S.9b). The H atom of G94 and side chain of A59 face towards this hydrophobic region and both residues are highly buried with a SASA of 0.106 nm^2 and 0.246 nm^2 respectively. Structural cluster 2 is also very close to V48, which has the highest maximum PI by a large margin (Figure 9/10). When comparing relaxed G-CSF (cyan) with unrelaxed (silver) in Figure 13, the short helix next to the structural cluster 2 hydrophobic region clearly moves outward in the unrelaxed structure (Figure S.9b). V48 appears to lead this outward movement of the short helix. In the relaxed structure the short helix is fairly straight, whereas in the unrelaxed structure it is curved with V48 at the apex. This structural change is not picked up by NMR as a significant change in the microenvironment of V48 because it is already solvent exposed (Table 5 and Figure S.8) and so already highly dynamic. The large cluster of hydrophobic residues in structural cluster 2 that experience a significant percentage increase in PI over the thermal ramp suggests that they become more mobile as the hydrophobic core unpacks, alongside a movement in position of the short helix. The intensity of the signal can in part depend on the ability of side chains to transfer magnetization to the solution. Therefore, expanding of the hydrophobic core region could also cause these residues to become more solvent exposed and significantly increase their signal. Moreover, residues G55, I56 and G149 experience extremely non-linear peak trajectories (Figure S.5) and are within (and close to) structural cluster 2, suggesting that this N-terminal loop AB region adopts multiple conformations as it expands.

H43 was in the 95th percentile of ∑Δδ, had the second highest percentage increase in PI (Table 4 and Table S.3) and was close to being in the 90th percentile for PI (Figure S.8). This suggested that it underwent a significant environmental change affecting its dynamics as the temperature increased. Given that this residue would be positively charged under this study's experimental conditions, it would also be attracted towards the nearby negatively charged E45 and E46 residues (Figure 12a and 13). Figure 13 highlights residues H43, E45, and L41 in green (relaxed) and red (unrelaxed). H43 can be seen to move further from E45 upon relaxation, shifting from 4.8 Å apart in the receptor-bound state, to 6.2 Å apart in the relaxed unbound structure. Moreover, the backbone of the loop containing L41 moved slightly, while the sidechain became more tightly packed onto the helix D backbone in the relaxed structure, compared to a more solvent exposed position in the unrelaxed structure, undertaking a 0.8 Å shift in position. Overall, the changes observed upon relaxation into the unbound structure appear to correspond with the NMR-observed transitions in reverse, and so from higher to lower temperature structures. This places G-CSF into a more active conformation at above 309 K (36 °C) *in vivo*.

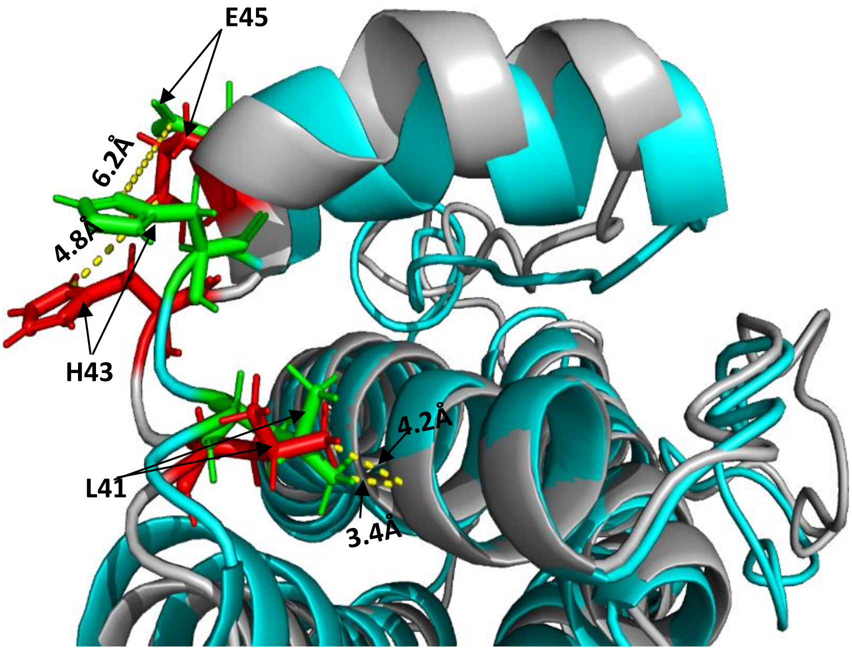


Figure 13. PDB:2D9Q vs. Relaxed Structure. PDB:2D9Q Relaxed structure in Rosetta Cartesian_ddg (cyan) overlaid on an unrelaxed PDB:2D9Q (grey). H43 and E45 highlighted green in the relaxed structure, H43 and E45 highlighted red in the unrelaxed structure. Distance between both residues is indicated in the relaxed and unrelaxed as 6.2 Å and 4.8 Å respectively. L41 is highlighted in the same manner as this with its distance from the neighbouring α-helix backbone being 3.4 Å in the relaxed structure and 4.2 Å in the unrelaxed structure.

This "switch" involving H43 could be significant to bioactivity because it is part of the same short unstructured loop as L41 and Y39. Both of these residues are part of the GCSF-R binding site and are buried (Table 5; Tamada *et al*., 2006). Hence, the aforementioned movement by H43 could pull L41 and Y39 out into solution so that they are more exposed to allow receptor binding.

Supporting this is the sharp change in the microenvironment of L41 and Y39 at the same time as the increase in H43's PI and the relaxed structure comparisons (Figure12c and 13). This "switch" appears to come at a cost to C36. The large increase in C36's PI beginning at ~307K places it in the 90th percentile of PI. However, at physiological temperature (309K), G-CSF would not experience this possible strain on C36 but would experience the benefit of the H43 "switch" (Figure 12c).

G-CSF is most stable at < pH7, ideally pH 4. Suggested contributions to this characteristic range from high colloidal stability to stronger cation-π interactions between residues W58 and H156 (both of which are very dynamic according to this study) at low pH (Ko *et al*., 2022; Chi *et al*., 2003). Furthermore, while the pH of bone marrow (where GCSF-R is present) is not well studied, some studies suggest it to be slightly acidic (Nikolaeva, 2018; Massa *et al*., 2017). The lower pH would therefore increase the attraction of H43 towards E45/E46, thus supporting a potential "switch mechanism" controlled by H43 as seen when the unrelaxed and relaxed G-CSF structures are overlaid (Figure 13). Bone marrow is also proposed to be more reducing than the intravascular environment, which could lead to a larger population of the reduced disulphide bond near H43, giving it more freedom to make the "switch" (Spencer *et al*., 2014; Woycechowsky and Raines, 2000).

Although V48 is part of the active site, its importance to bioactivity could be more than just binding to the receptor. Its dynamic nature could facilitate the expansion of binding site III (Figures 12a and S.9b), which could aid with the complementarity of this binding site to the receptor (Tamada *et al*., 2006). Additionally, it could act in combination with the H43 "switch" to help expose L41 and Y39. Faster binding has been reported when proteins are partially unfolded (increasing the capture radius), inducing long-range, water-mediated interactions with the target and finally folding upon binding. This emphasizes the importance of structure remodelling in receptor recognition and binding (Shoemaker, Portman and Wolynes, 2000). The positioning of histidines (Figure 14) in G-CSF could play an important role in this long-range interaction mechanism. All histidine side chains, aside from H43, elicit potential for intramolecular interactions with residues in loop regions (also suggested by Aubin *et al*., 2015), namely cation-π/hydrogen bonding between H79 and W118/Q131, W58 and H156/H52 and H170 and S66. These interactions are likely to be stabilising because the histidines are in structured regions, thus explaining greater stability for G-CSF at a lower pH. Therefore, the relatively higher pH of blood would enable long-range targeting for G-CSF followed by higher stability and binding affinity (from H43's "switch") at the lower pH of bone marrow.

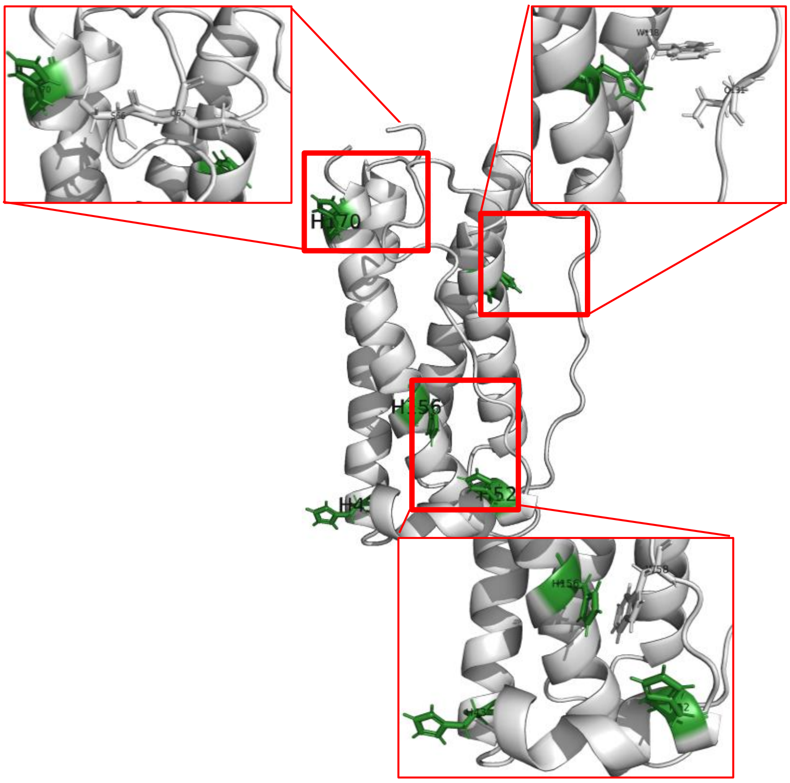


Figure 14. The Importance of Histidine Positioning for Stability. PDB:2D9Q with all histidines highlighted. Red boxes magnify potential interactions with these histidines.

3.2 Discussion

NMR was able to assign and track the mobility of G-CSF residues across a range of temperatures prior to any global unfolding. These findings were highly consistent with previous observations of the influence of loop AB and surrounding structure on G-CSF stability and aggregation propensity. Physiological temperature induced structural changes in a local structural cluster around loop AB that corresponded to regions previously linked to the formation of an aggregation-prone state. Furthermore, the same structural changes were important for "switching on" of bioactivity through remodelling of the receptor binding site III. The implication is that while the use of formulation approaches remains highly suitable for stabilising against the aggregation-inducing conformational change in a product vial or syringe, the use of protein engineering strategies to stabilise against the same structural changes may have knock-on functional effects *in vivo*. These findings also provide further insight into why the T_m values are often a poor predictor of aggregation kinetics when stored at lower temperatures (Robinson *et al*., 2018). The thermally-induced conformational switch at 307-310 K (34-37 °C) would mean that the global unfolding measurement of T_m is made from a different native state than the one present at the lower temperatures used for drug product storage. Finally, the remodelling of loop AB and binding site III involves a critical change in the position of residue H43, which points to a likely pH-sensitivity, including the reason why G-CSF is more stable at pH 4.25 *in vitro* than at physiological pH. It is also possible that the pH sensitivity is an important feature in G-CSF activation in long bone marrow which has a slightly acidic pH.

Chapter 4: Elucidating Excipient Mechanism of Action

The previous chapter explored residues that are key to G-CSF structure remodelling, and thereby stability and function. Formulating biotherapeutics with excipients is a very common protein engineering method for stabilisation in both liquid and solid state. Observing mechanisms by which excipients (de-)stabilise proteins can be challenging, with current approaches including NMR (Aubin *et al*., 2015; Ghasriani *et al*., 2020), SAXS, DCS, DLS (Xu *et al*., 2019) and molecular docking software (Barata *et al*., 2016). The high resolution of 2D NMR endows it with the ability to interrogate perturbed/interacting residues in the presence of excipients (Cui *et al*., 2020). Furthermore, varied-temperature 2D NMR (VT-NMR) permits interrogation of non-solvent hydrogen bonding as co-solvent is altered (Heisel and Krishnan, 2014). This chapter will explore how the VT-NMR approach used in chapter 3 can probe the influence of excipients on regions key to thermal resilience and structure remodelling. The purpose of this is to explain previous observations with stability when G-CSF is formulated with excipients and to predict the impact the these formulations on bioactivity.

In the case of G-CSF, sugars as cosolvents have a mild effect on thermal stability, whereas amino acids have a more pronounced impact at pH 4.25 (Wood *et al*., 2020). Phenylalanine and histidine, at 12.5 mM and 25 mM respectively, improved T_m to the same point as ~150 mM mannitol, sucrose, sorbitol and trehalose. However, all amino acid excipients became destabilising at above ~50 mM, with arginine being destabilising at all concentrations. This is in concert with studies showing little to no protein-excipient interaction when G-CSF is formulated with surfactants and sugars (Aubin *et al*., 2015; Ghasriani *et al*., 2020), while other studies show potential interaction with arginine (Wood *et al*., 2020).

Applying the VT- NMR approach used in chapter 3 to excipient studies could shed light on residues that encourage conformational diffusion to partially unfolded aggregation intermediates. Conformation diffusion describes the shift in population between different structural conformers. Therefore, the impact of excipients on the formation of intermediates, and thus possible mechanisms of conformational and colloidal de/stabilisation can be probed. The excipients examined in this chapter are phenylalanine and histidine at 12.5 mM and 25 mM respectively, which improve conformational stability (Wood *et al*., 2020), as well as 25 mM and 50 mM arginine, given its deleterious effect on conformational stability. Peak trajectory linearity (described in Table S.1 and Figure S.5) is used here to assess how cosolvent affects conformational diffusion and hydrogen temperature coefficients (described later) and to probe whether this is due to stronger hydrogen bonding (Tomlinson and Williamson, 2012).

The relationship between residue peak trajectories and temperature is empirically linear and is influenced by δ in ^{15}N and ^{1}H planes (Andersen *et al*., 1997; Tomlinson and Williamson, 2012).

While, linearity is dominated by hydrogen bonding for the amide hydrogen (^{1}H), the amide nitrogen (^{15}N) is affected by many additional factors such as torsion angles of the same and neighbouring residues (φ_i, ψ_{i-1} and χ^1), side chain rotamer, hydrogen bonding and electrostatic interactions (De Dios, Pearson and Oldfield, 1993; Wang and Markley, 2009). An increase in hydrogen bonding either side of the amide, which could be induced by addition of acidic solvent, decreases its electron density (shielding) and causes δ to appear further downfield. According to early studies by Llinás and Klein, rapid upfield movement of δ for solvent exposed amides with increasing temperature can be caused by the decrease in intermolecular hydrogen bonding (with solvent) outweighing intramolecular bonding (Llinas and Klein, 1974).

However, the prediction of amide hydrogen bond status can be very elusive given their transient nature. Some studies derive amide hydrogen/nitrogen temperature coefficients (i.e. the slope of $\Delta\delta_H/\Delta\delta_N$ vs Δtemperature) given that experimentally resolved structures have shown that ^{1}H temperature coefficients ($\Delta\delta_H/\Delta T$) more positive than -4.6 ppb/K indicate non-solvent hydrogen bonding for that amide (Cierpicki and Otlewski, 2001). Conversely, more recent studies have indicated that hydrogen bonding poorly determines $\Delta\delta_H/\Delta T$, particularly in structured regions (Tomlinson and Williamson, 2012), instead attributing general loss in structure as a better determinant. This has also led some to use relaxation dispersion NMR for more accurate predictions (Bouvignies *et al*., 2011). Nonetheless, $\Delta\delta_H/\Delta T$ will be calculated in this study to examine whether intramolecular hydrogen bonding can account for excipient-induced structural changes.

Results confirmed previous observations that strength of protein-excipient interaction is a poor measure of an excipient's ability to improve stability (Zalar, Svilenov and Golovanov, 2020). Destabilising excipient conditions increased the influence of conformational diffusion for the highly conserved residue R166, which elicits significant thermal resistance. Furthermore, although excipients influence hydrogen bonding of select residues, the general structural remodelling (including that for the “switch” mechanism) observed in WT with no excipients, remained across all conditions.

4.1 Results

4.1.1 Thermal Resistance Relatively Stays the Same Across Excipient Conditions

Comparing microenviromental changes induced by temperature permits investigation into how excipients influence regions showing little thermal resistance. The term WT will refer to the control condition (i.e. just protein and no excipient) from this point on, given that all sample conditions contained WT G-CSF. The maximum data distribution (Figure 15 and 16) at 323 K for 25 mM histidine (~0.36 Δppm) is smaller than for other conditions because signal is lost for the two highest ΣΔδ residues at 321 K (S63 and Q70). Both of these residues accumulated a ΣΔδ at 321 K similar to that for 12.5 mM phenylalanine. Data distribution for 25 mM histidine is the only condition comparable to WT at 297 K, with other excipients displaying similar distribution to WT for the following temperatures (except for 50 mM arginine). 12.5 mM phenylalanine and 25 mM arginine have probability distribution maxima ~10% lower than for WT with 25 mM histidine at 295 K (Figure 16). This could indicate that histidine has a global stabilising effect on WT at early temperatures. For 50 mM arginine, the data distribution shifted to considerably higher ΣΔδ values for ≥299 K. However, there was a much lesser distribution of data from the mean at each temperature above and including 299 K, perhaps suggesting that 50 mM arginine induced microenvironment changes to a globally similar extent during denaturation. Thus, it appears that these excipient conditions, with the exception of 50 mM arginine, generally induced similar ΣΔδ over the melt.

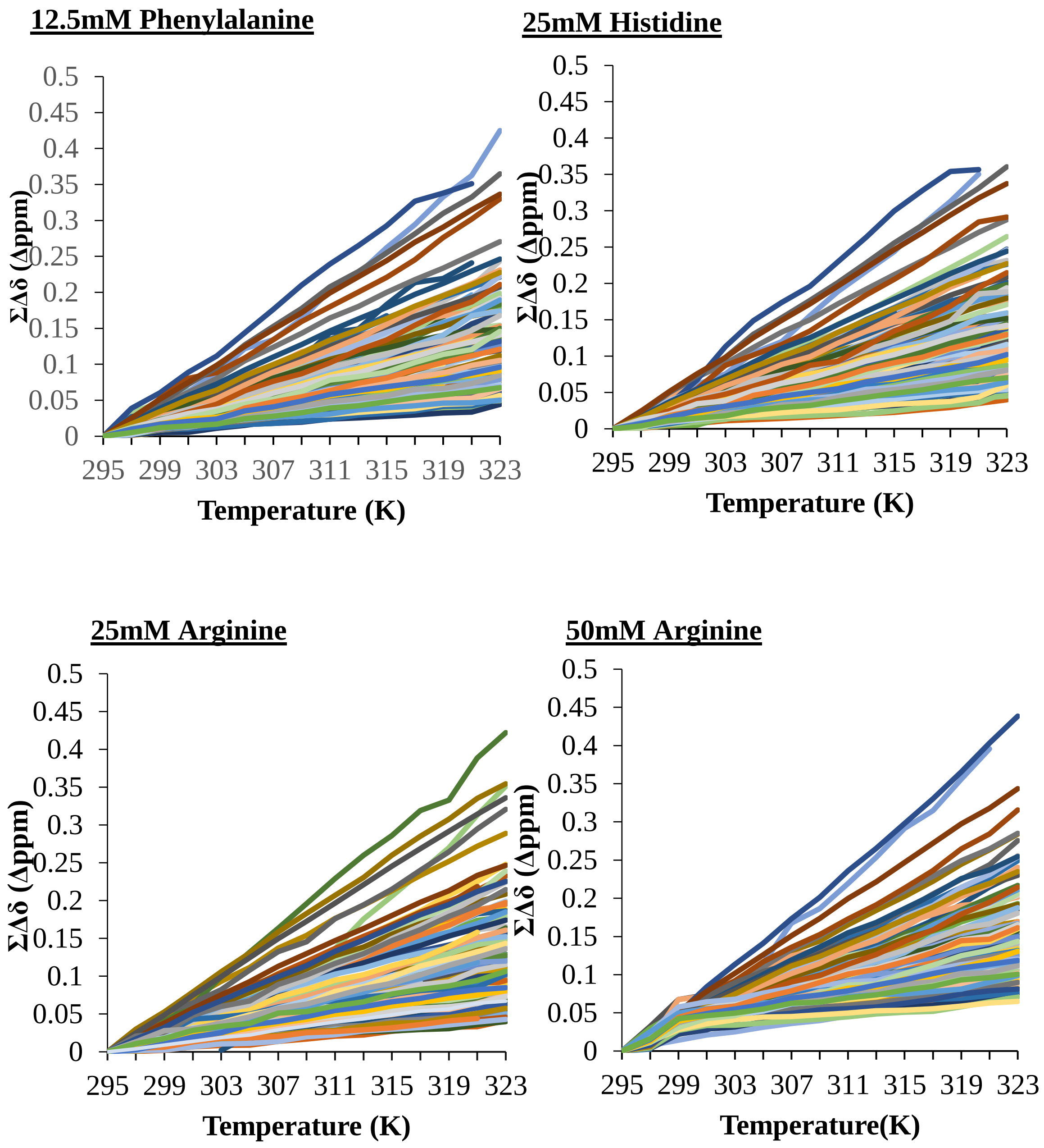
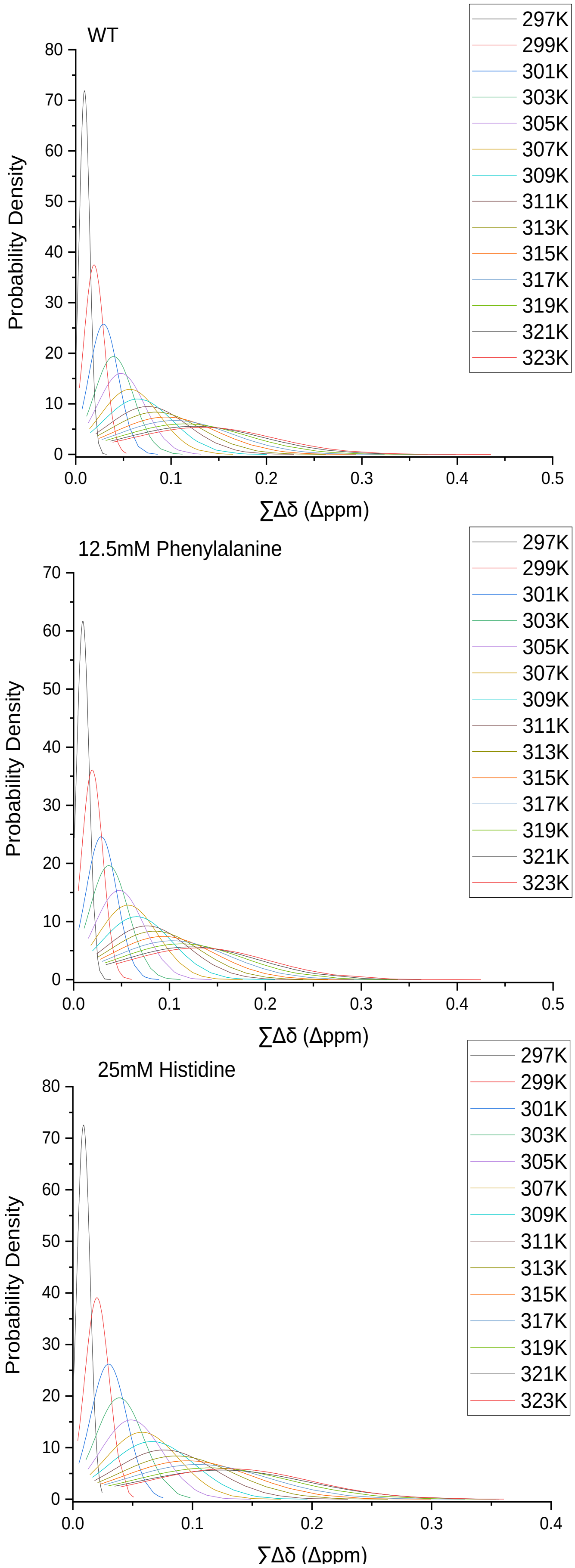


Figure 15. Temperature dependence of ∑Δδ for all assigned residues at Different Excipient Conditions. Differently coloured trends represent different residues. The starting temperature is 295 K and therefore has no cumulative Δδ distance.



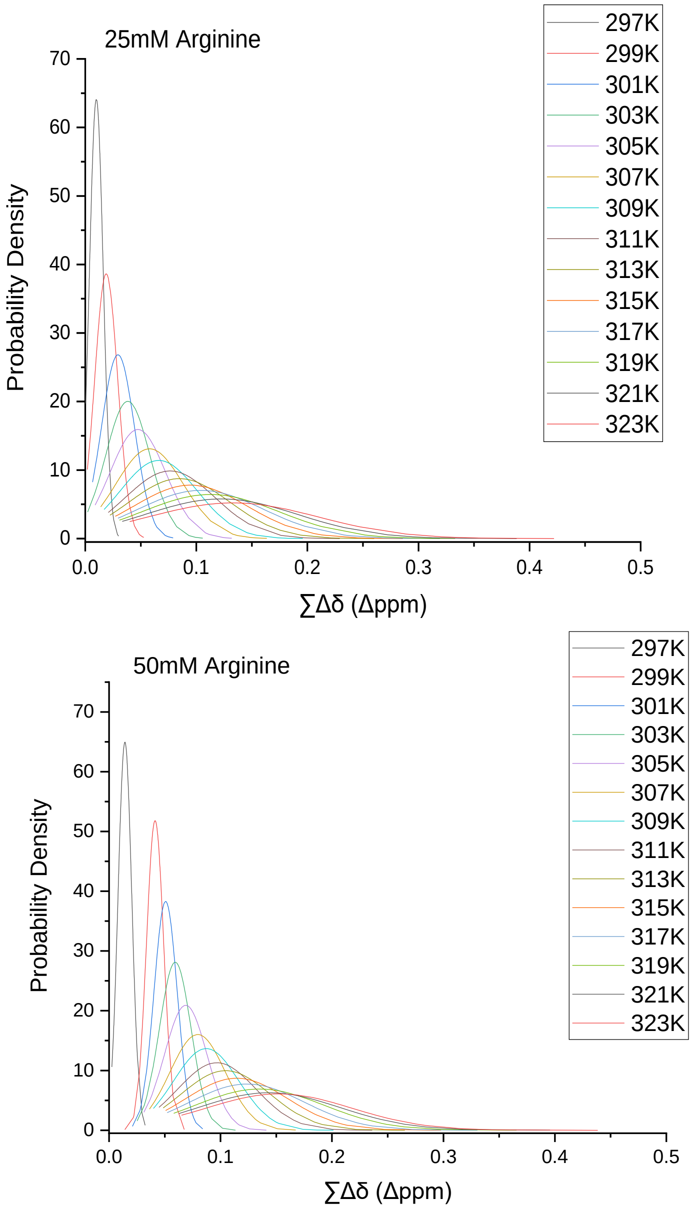


Figure 16. Normal Distribution of ∑Δδ at Each Temperature Point for Different Excipient Conditions. The temperature for each distribution is indicated in the legend.

Residues showing significant changes in ∑Δδ appear in the same unstructured loop regions (N-terminus of helix A and loops AB, BC and CD) for the different excipient conditions as when observed with no excipients. H156, located in helix D, is the only residue in a structured region that exhibits significant ∑Δδ change over the melt for all tested conditions (Figure 9a and 17). Other common observations among all conditions include residue Q70 which displayed the highest ∑Δδ for the vast majority of the melt, and ∑Δδ for residue S63 which increased rapidly at above ~305 K. This increase in S63 ∑Δδ for 25 mM Arginine, while not visible at ~305 K in Table 6, did occur at this temperature.

However, key differences in ∑Δδ appear to have occurred between conditions. The clearest of these is the chaotropic nature of Arginine; perturbing more and more residues in helical regions with increasing concentration. More specifically, the vast majority of these perturbed residues appear to be localised to helices B and D. Residues in the 95th percentile of ∑Δδ were situated in the same regions (loops AB, BC and CD) for the most of the melt for all conditions except 50 mM Arginine. For this condition, residues L15 and V48 made more than just brief appearances in the 95th percentile. Of note, it was observed in no-excipient conditions that V48, although the most dynamic residue, did not exhibit significant micro-environmental changes (Figure 9a and 17). This same pattern was observed for all excipient conditions tested here aside from 50 mM Arginine, which shows both the highest dynamics for V48 (Table 7 and Figure 17) and a significant change in ∑Δδ. Additionally, residues H43 and E45 only showed significant change in ∑Δδ, as expected for a functioning "switch" mechanism, for 12.5 mM Phenylalanine and 25 mM Histidine.

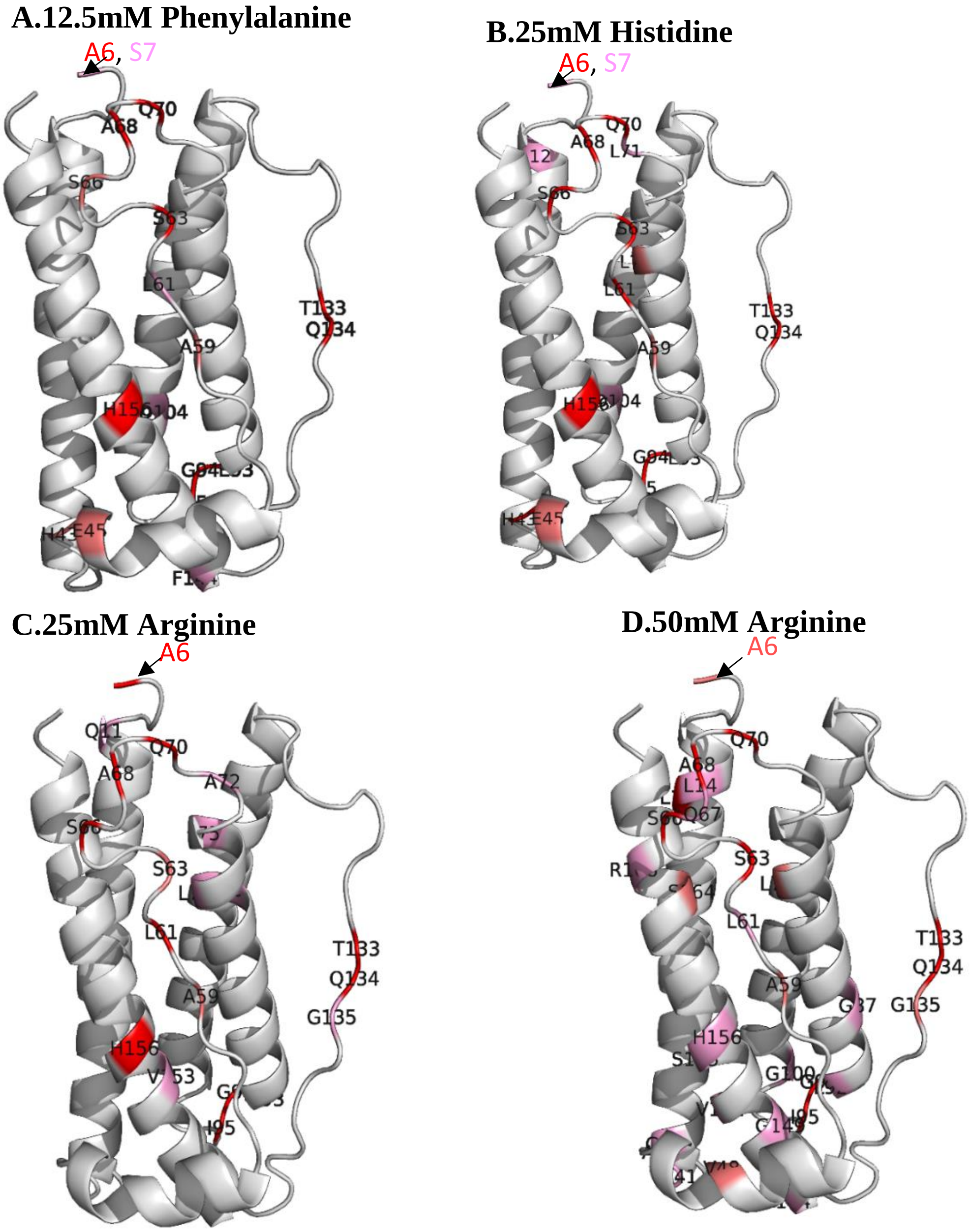


Figure 17. Mapping Significant ∑Δδ onto G-CSF in Different Excipient Conditions. Residues in the 90th percentile of ∑Δδ are coloured in the same manner as **Figure 9** for all excipient conditions (**A-D**). PDB:2D9Q is missing its first 6 residues, therefore S7 is highlighted in place of earlier residues.

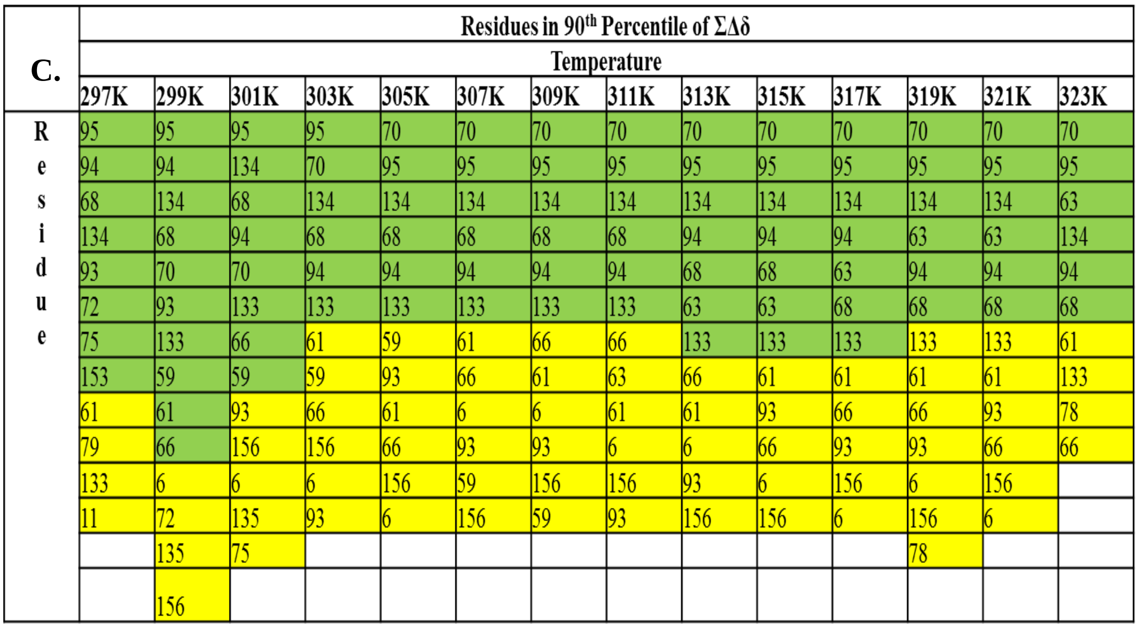
12.5mM Phenylalanine

A	Residues in 90th Percentile of ∑Δδ													
	Temperature													
	297K	299K	301K	303K	305K	307K	309K	311K	313K	315K	317K	319K	321K	323K
Residue	70	70	70	70	70	70	70	70	70	70	70	70	63	63
	43	94	94	134	95	95	95	95	95	63	63	63	70	95
	93	134	134	95	134	134	134	63	63	95	95	95	95	134
	134	43	95	94	63	94	63	134	134	134	134	134	134	94
	45	95	43	63	94	63	94	94	94	94	94	94	94	68
	63	68	63	68	68	68	68	68	68	68	68	68	68	133
	104	63	68	133	133	133	133	133	133	45	45	45	45	61
	95	133	133	43	59	6	59	45	45	133	133	133	133	144
	94	59	93	93	6	156	6	93	93	93	6	6		156
	133	93	59	45	156	66	156	6	6	66	156	156		6
	68	66	156	6	66	93	93	156	156	156	66			
		45	6	59		59	66		66	6				
		156		66										
		6		156										
		7												

25mM Histidine

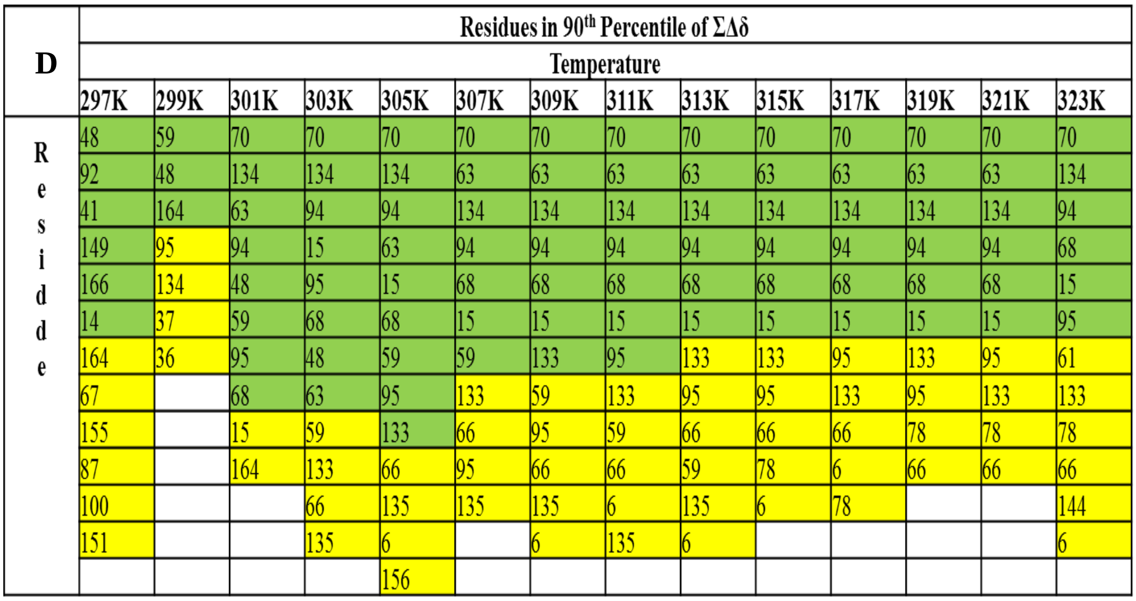
B	Residues in 90th Percentile of ΣΔδ													
	Temperature													
	297K	299K	301K	303K	305K	307K	309K	311K	313K	315K	317K	319K	321K	323K
Residue	59	134	134	70	70	70	70	70	70	70	70	70	70	95
	134	68	95	95	95	95	95	95	95	95	63	63	63	134
	78	95	68	134	134	134	134	134	134	134	95	95	95	94
	95	59	70	68	68	68	63	63	63	63	134	134	134	68
	70	78	63	94	63	63	68	68	68	68	68	94	94	43
	68	61	59	63	94	94	94	94	94	94	94	68	68	45
	61	133	94	59	133	133	133	43	133	43	43	43	43	78
	104	71	133	133	59	43	43	133	43	133	133	133	133	133
	12	6	78	156	156	6	6	6	66	61	61	61	45	61
	6	94	61	6	6	156	66	156	6	66	45	78	61	6
	93	156	156	66	61	66	61	61	61	156	66	45	78	156
	133	66	6	78	66	59	156	66	156	6	156	6	66	66
		70	43	61	78									
		7	66											

25mM Arginine



C.	Residues in 90^{th} Percentile of $\Sigma\Delta\delta$													
	Temperature													
	297K	299K	301K	303K	305K	307K	309K	311K	313K	315K	317K	319K	321K	323K
R	95	95	95	95	70	70	70	70	70	70	70	70	70	70
e	94	94	134	70	95	95	95	95	95	95	95	95	95	95
s	68	134	68	134	134	134	134	134	134	134	134	134	134	63
i	134	68	94	68	68	68	68	68	94	94	94	63	63	134
d	93	70	70	94	94	94	94	94	68	68	63	94	94	94
u	72	93	133	133	133	133	133	133	63	63	68	68	68	68
e	75	133	66	61	59	61	66	66	133	133	133	133	133	61
	153	59	59	59	93	66	61	63	66	61	61	61	61	133
	61	61	93	66	61	6	6	61	61	93	66	66	93	78
	79	66	156	156	66	93	93	6	6	66	93	93	66	66
	133	6	6	6	156	59	156	156	93	6	156	6	156	
	11	72	135	93	6	156	59	93	156	156	6	156	6	
		135	75									78		
		156												

50mM Arginine



D	Residues in 90^{th} Percentile of $\Sigma\Delta\delta$													
	Temperature													
	297K	299K	301K	303K	305K	307K	309K	311K	313K	315K	317K	319K	321K	323K
R	48	59	70	70	70	70	70	70	70	70	70	70	70	70
e	92	48	134	134	134	63	63	63	63	63	63	63	63	134
s	41	164	63	94	94	134	134	134	134	134	134	134	134	94
i	149	95	94	15	63	94	94	94	94	94	94	94	94	68
d	166	134	48	95	15	68	68	68	68	68	68	68	68	15
d	14	37	59	68	68	15	15	15	15	15	15	15	15	95
e	164	36	95	48	59	59	133	95	133	133	95	133	95	61
	67		68	63	95	133	59	133	95	95	133	95	133	133
	155		15	59	133	66	95	59	66	66	66	78	78	78
	87		164	133	66	95	66	66	59	78	6	66	66	66
	100			66	135	135	135	6	135	6	78			144
	151			135	6		6	135	6					6
					156									

Table 6. Residues are highlighted in the same manner as **Table 4A** for all excipient conditions (**A-D**).

4.1.2 Excipients Induce Structural Changes to Varying Extents

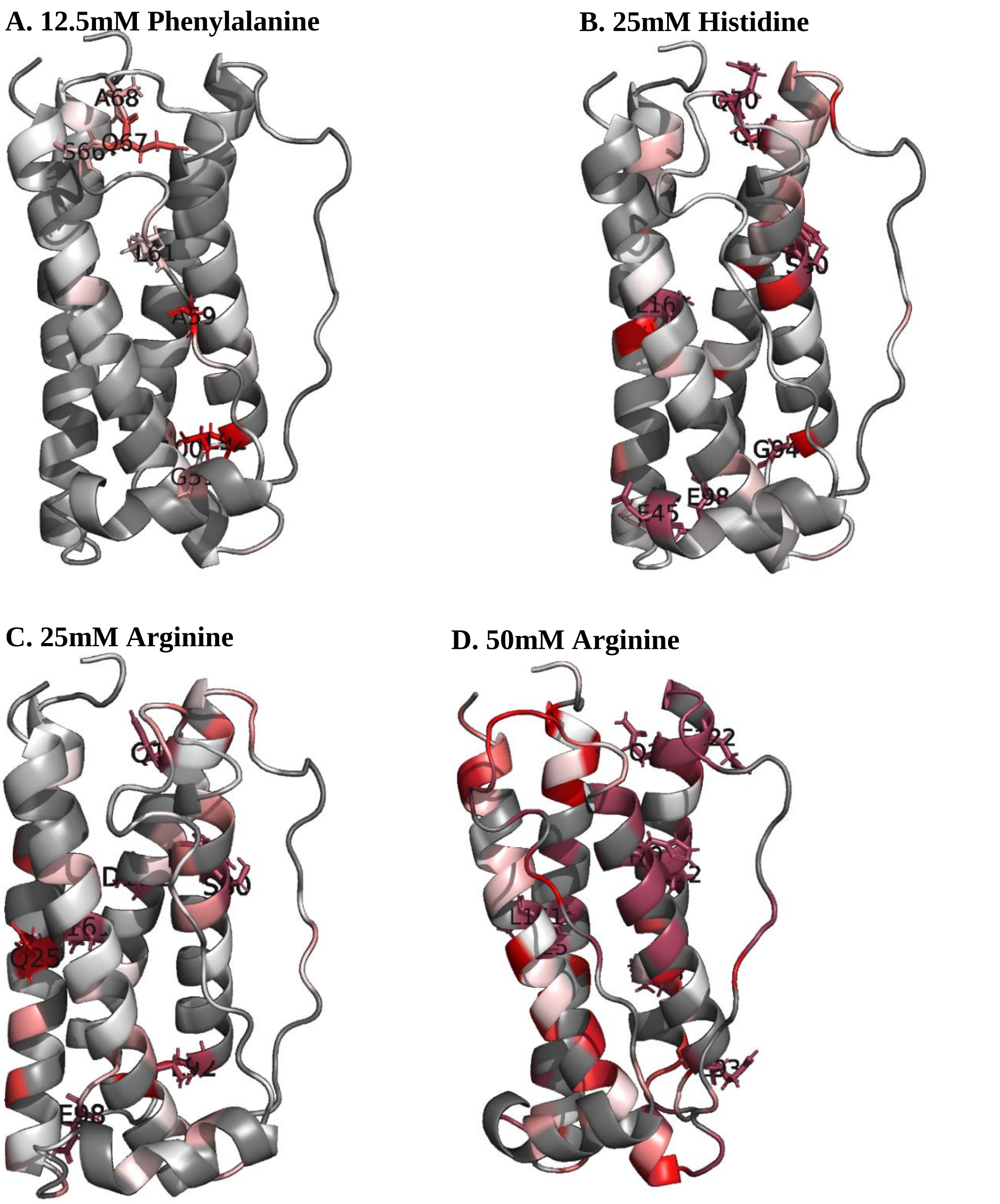
Monitoring ∑Δδ reveals that key residues in loops AB, BC and CD are the most susceptible to thermal stress, regardless of excipient condition. Nevertheless, this study shows that in a low energy state/low temperature (299K), the addition of different excipients perturbed markedly different regions (Figure 18). Excipient induced Δδ is defined in Figures 18-20 as the Δδ at each residue position between no excipient (control) conditions and respective excipient conditions at

299 K. The 90th percentile threshold for total data distribution of all control to excipient Δδ data is 0.030 Δppm and is the colour scale upper limit in Figure 18. Residues with a Δδ above this value are coloured raspberry, with residues displaying weaker Δδ from red to white to grey. Residues highlighted as sticks and labelled are in the 95th percentile of WT to excipient Δδ data for that individual excipient data set. Grey typically represents unassigned residues.

However, given that Δδ was generally very low for 12.5 mM phenylalanine (Figure 18a), most residues appeared as grey even though they are assigned. This condition caused no residues to be significantly affected (no Δδ above 0.030 Δppm), whereas all other conditions did have an impact. This could be because 12.5 mM phenylalanine had only a very weak/transient interaction (if any at all) with G-CSF. Areas that were affected under 12.5 mM phenylalanine were localised to loop AB and partially in loop BC, which could explain its stabilising mechanism given that loop AB is the least thermally resistant region (Raso *et al*., 2005; Wood *et al*., 2020). As will be shown in section 5.1.4, stabilising mutations made in loop AB when compounded with 12.5 mM phenylalanine notably increase the melting point.

Residues that were significantly affected in 25 mM histidine and 25/50 mM arginine were more spread throughout the structure than those relatively affected in 12.5 mM phenylalanine (Figure 18b-d). Furthermore, the more structured helical regions were disturbed under these three conditions, potentially hinting at strong interactions. For 25 mM histidine (Figure 18b), these significantly affected residues occurred in helix B, C and D, the short helix (residue E45), N-terminus of loop AB and loop BC. Residue E45 is key to the "switch" mechanism and, as expected, bioactivity was markedly impacted for WT G-CSF in 25 mM histidine (Figure 37). Significantly impacted residues at both arginine concentrations were spread throughout helices A-D. Of note, 50 mM arginine induced the largest changes of all conditions, which were also globally spread across both loops and helices.

Given that the structural impact was widespread under 25 mM histidine and 25/50 mM arginine, it was difficult to attribute a particular area of G-CSF to the de/stabilising mechanism of these excipients. Nonetheless, to help address this point, it is important to understand what is meant when a residue is "affected" under excipient conditions. Possible explanations may result from direct excipient interaction, non-direct interaction (forcing solvent onto the structure and changing it), or excipient-induced pH changes in solvent. Hence, probing predicted excipient docking sites, residue charge and solvent accessibility could shed light on this.



A. 12.5mM Phenylalanine

B. 25mM Histidine

C. 25mM Arginine

D. 50mM Arginine

Figure 18. **The Influence of Excipients on δ at 299K (Excipient-Induced Δδ).** The colour scale is grey (unassigned) to white (low Δδ) to raspberry (high Δδ) and is used for each excipient condition (**A.-D.**). The colour scale upper limit is 0.030 Δppm.

4.1.3 Investigating Excipient Interaction with iGEMDOCK

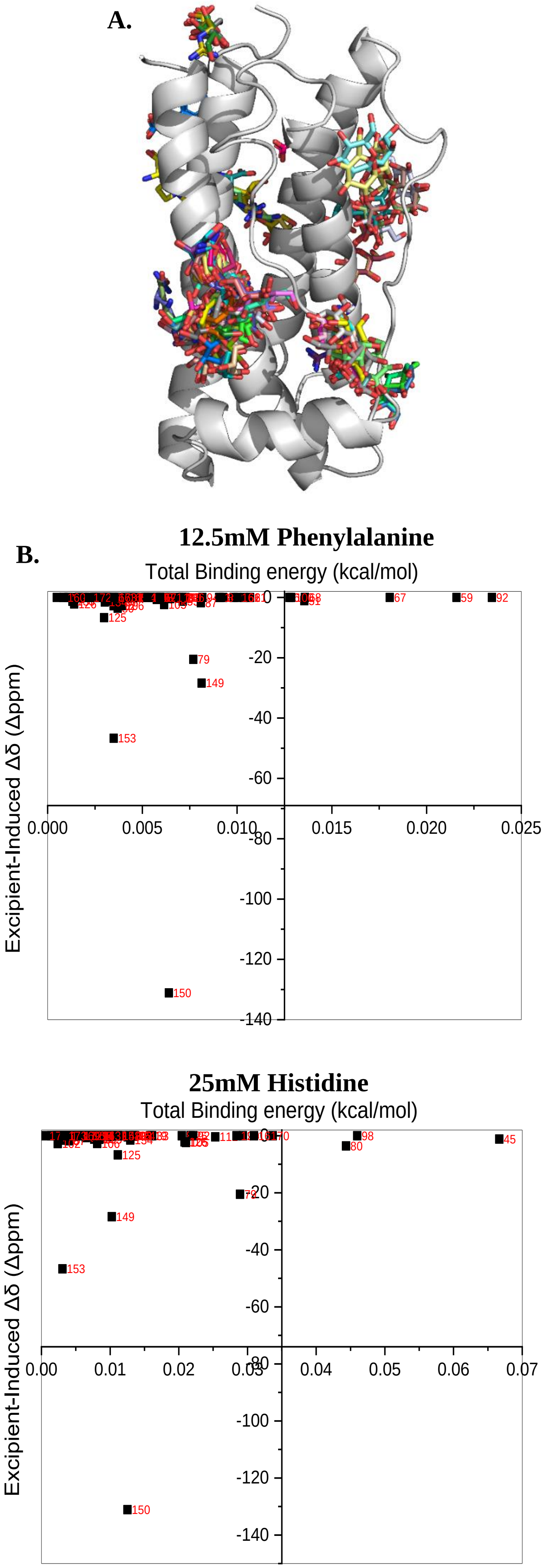
iGEMDOCK (Hsu *et al*., 2011) is used here to predict potential protein-excipient interaction sites (Figure 19). Excipients arginine, phenylalanine, sorbitol, mannitol, trehalose and sucrose were docked against a relaxed G-CSF structure (described in section 3.1.6) and found to distinctly cluster around helix A, B, D and loop AB (Figure 19a). The predicted binding energy of these excipients was calculated by summing energies for side-chain and main-chain hydrogen bonding, VDW interactions and electrostatic interactions. This “total binding energy” was compared with excipient-induced Δδ for each condition tested with NMR (Figure 19). All quadrant plots are such that the horizontal label represents the y-axis and the vertical label represents the x-axis. In Figure 19b, the lower the binding energy, the stronger the predicted binding. Although the vast majority of residues had a predicted binding energy of 0 kcal/mol, some residues like H79 (helix B), G149, G150 and V153 (helix D) were clearly predicted as interaction sites. Residue H79 was also significantly affected in 25 mM histidine and 25/50 mM arginine (Δδ at ~0.03 Δppm or above), while G150 was significantly affected in 25/50 mM arginine.

Hence, it would appear that excipient interactions can occur at H79. However, given that H79 is positively charged (and buried from solvent, Figure 20d) under these experimental conditions, histidine and arginine excipients may be interacting with the proximal E122 and indirectly causing a change to the microenvironment of H79 (Figure 20b).

Electrostatic interaction is key to many protein-ligand interactions and was therefore compared with excipient-induced Δδ in Figure 20c. Charged residues in G-CSF were almost exclusively located in its helices (aside from K40, H43, E93 and E98) as seen in Figure 20a. Residues that were both charged and significantly affected in histidine and arginine excipient conditions are illustrated with sticks and labelled in Figure 20b. All but one of these residues are negatively charged, which comports with the notion of protein-excipient interaction Figure 20b/c.

Figure 20d compares solvent accessibility with Δδ and emphasises the impact of destabilising excipient arginine on more buried residues. Here, residues with a Δδ above 0.03 Δppm skewed towards lower solvent accessibilities. These same residues skewed more towards higher solvent accessibilities (such as ~1.25 nm^2 for residue E45) for the stabilising excipient histidine. Affecting buried residues could be key to a destabilising mechanism given that hydrophobic interactions play an important role in protein folding (Newberry and Raines, 2019).

Residues E93/98 are highlighted in the green box in Figure 20b because protein-excipient interaction in 25/50 mM arginine in this region may be pivotal to the destabilising mechanism of this excipient as later described in section 4.1.4. Moreover, although E98 (circled green in Figure 20d) was affected to roughly the same extent with histidine and arginine excipients (~0.05 Δppm), E93 went from a Δδ of below 0.02 Δppm with histidine to above 0.08 Δppm in 50 mM arginine. The proximal and buried residue L92 followed this same trend. Of note, this region (L92-I95) was highly sensitive to temperature (Figure 17).


A.
B.
12.5mM Phenylalanine
Total Binding energy (kcal/mol)
Excipient-Induced Δδ (Δppm)
0.000
0.005
0.010
0.015
0.020
0.025
0
-20
-40
-60
-80
-100
-120
-140
160
172
128
125
109
87
67
59
92
79
149
153
150
25mM Histidine
Total Binding energy (kcal/mol)
Excipient-Induced Δδ (Δppm)
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
-20
-40
-60
-80
-100
-120
-140
17
102
166
125
134
98
80
45
149
153
150

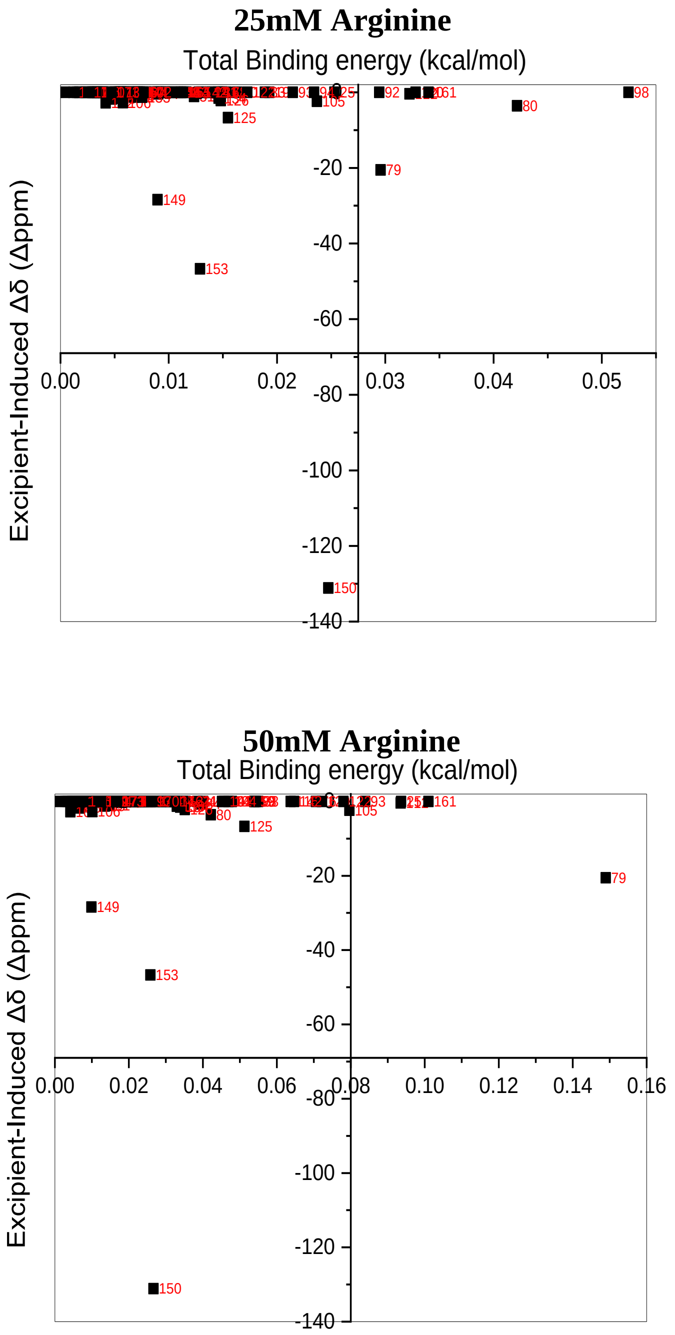
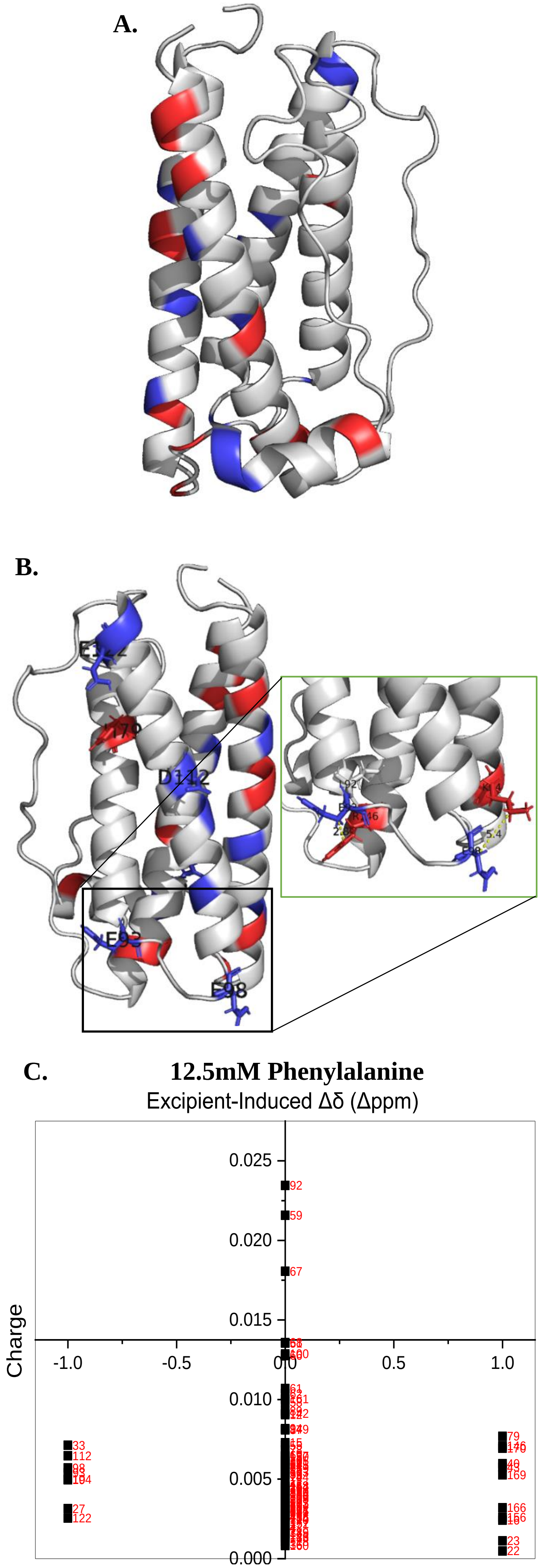
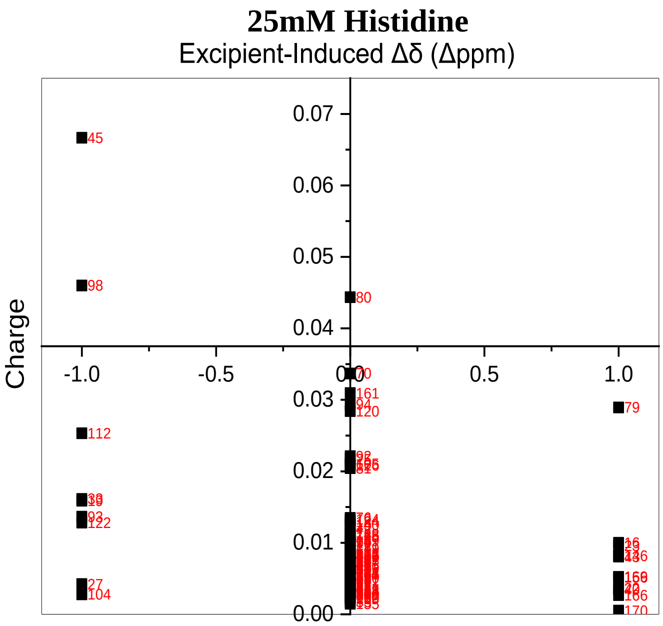
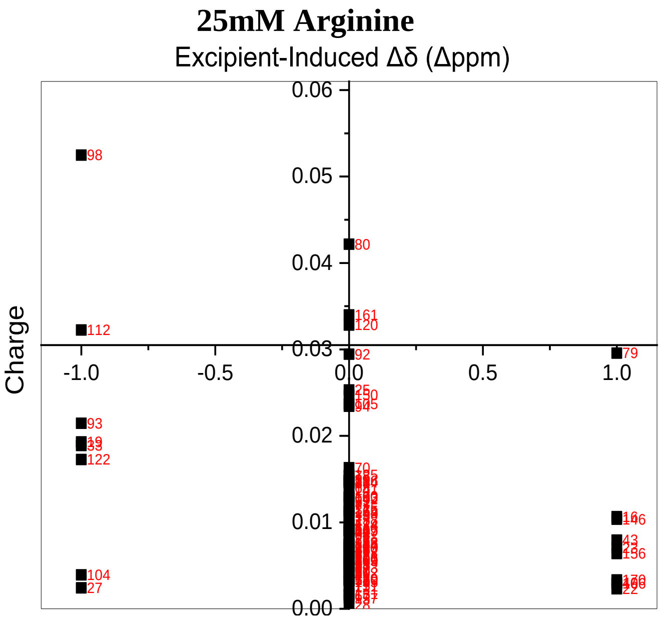


Figure 19. Comparing Excipient-Induced Δδ with In Silico Docking. A. Docking positions of excipients (arginine, phenylalanine, sorbitol, mannitol, trehalose and sucrose) from iGEMDOCK on relaxed G-CSF structure. **B.** Predicted docking energy from iGEMDOCK software (kcal/mol) vs Δδ from WT to excipient condition at 299K. Residue numbers are highlighted red.

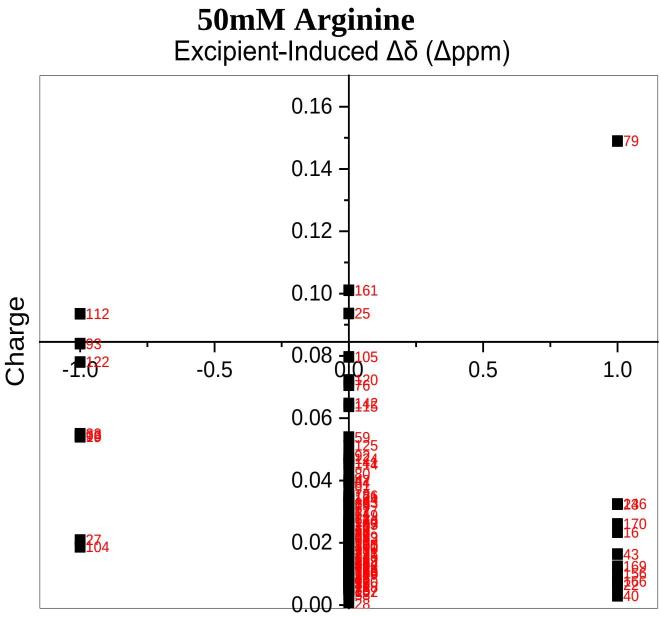

A.
B.
E122
179
D112
E93
E98
92
14
46
C.
12.5mM Phenylalanine
Excipient-Induced Δδ (Δppm)
Charge
-1.0
-0.5
0.0
0.5
1.0
0.025
0.020
0.015
0.010
0.005
0.000
92
59
67
160
61
161
142
149
33
112
98
104
27
122
79
146
170
40
169
166
156
23
22



25mM Histidine

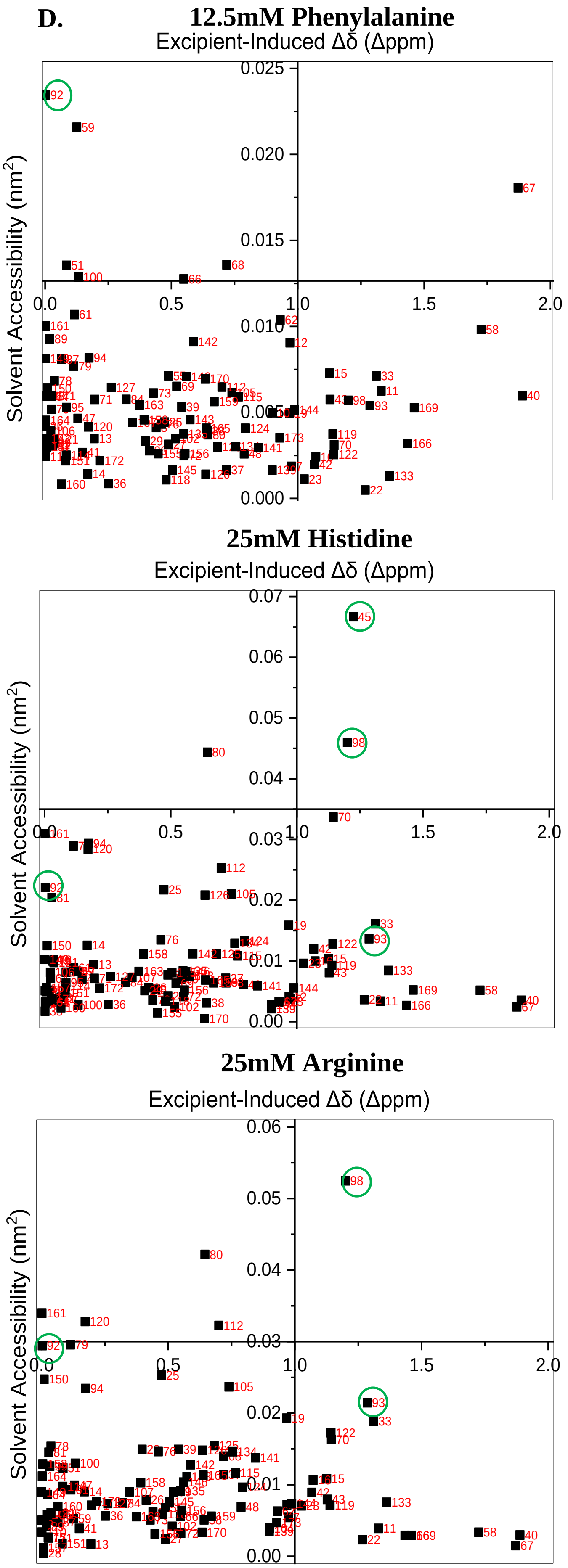


25mM Arginine



50mM Arginine

D.


12.5mM Phenylalanine
Excipient-Induced Δδ (Δppm)
Solvent Accessibility (nm^2)
0.025
0.020
0.015
0.010
0.005
0.000
0.0
0.5
1.0
1.5
2.0
92
59
67
51
100
66
68
61
62
161
89
142
12
58
94
79
15
33
40
127
73
69
170
11
169
71
163
39
144
93
98
43
164
47
143
120
124
119
13
102
173
166
70
122
141
48
172
41
151
14
145
137
139
42
133
160
36
118
23
22
25mM Histidine
Excipient-Induced Δδ (Δppm)
Solvent Accessibility (nm^2)
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.0
0.5
1.0
1.5
2.0
45
98
80
70
161
94
120
112
92
25
126
105
81
19
33
93
150
14
76
124
122
158
42
115
13
119
43
133
163
107
141
144
169
58
40
151
172
156
38
11
166
160
36
102
67
155
170
139
25mM Arginine
Excipient-Induced Δδ (Δppm)
Solvent Accessibility (nm^2)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.0
0.5
1.0
1.5
2.0
98
80
161
120
112
92
79
150
25
94
105
93
19
33
122
70
78
81
39
125
134
100
142
141
115
164
158
107
135
15
42
119
133
160
59
36
145
156
159
48
41
102
170
11
22
169
58
40
67
151
13
28

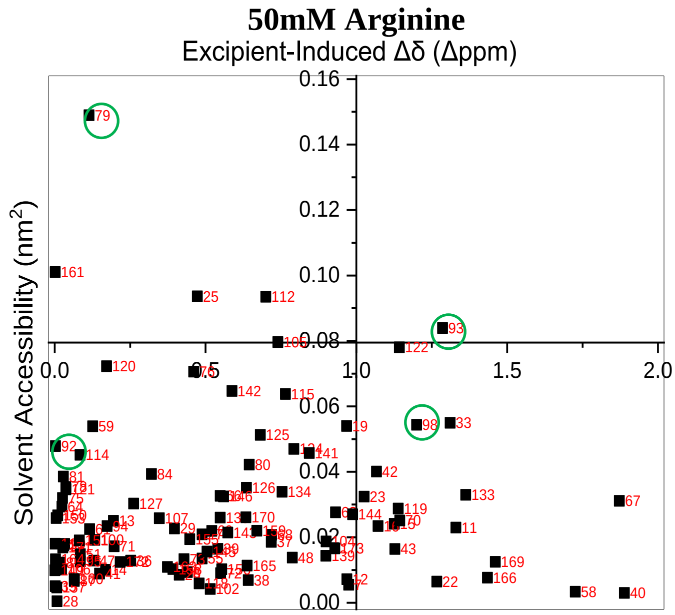
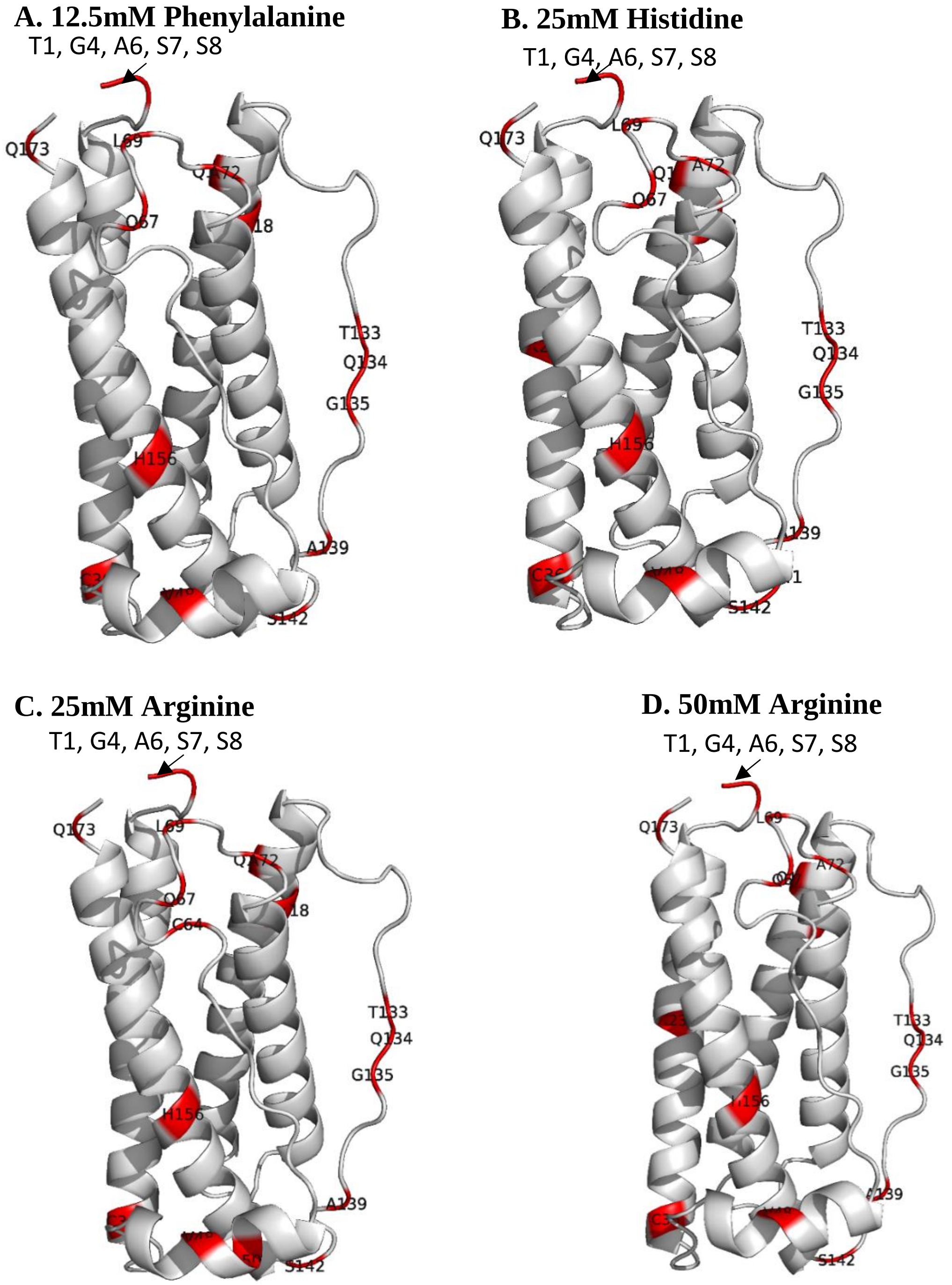


Figure 20. The Influence of Residue Charge and Solvent Accessibility on Potential Excipient Interaction. **A.** Positively charged (red) and negatively charged (blue) residues with **B.** highlighting some residues that are both charged and have high Δδ. The green box highlights an area of interest with distances between residues given on the yellow line. Charge and solvent accessibility vs Δδ from WT to excipient condition at 299K are shown respectively in **C.** and **D.** A charge value of -1.0 is negative, 1.0 is positive and 0 is no net charge. Residues of interest are circled green in **D.**

4.1.4 Probing Changes in Dynamics

PI 90th percentile residues predominantly stayed the same for the control and all excipient conditions (Figure 9b and 21a-d). The few residues that did alter between excipient conditions included K23, L47, L50 and C64. Comparing residues with the highest % increase in PI (Figure 21e-h) was more akin to the observations for the ∑Δδ 90th percentile (Figure 17). Control, 12.5 mM phenylalanine and 25 mM histidine conditions yielded similar clusters of highlighted residues, while arginine progressively affected more residues with increasing concentration. Interestingly, the majority of these additionally affected residues were in structured regions (helix A, C, D and short helix). This enforces previous observations of arginine's chaotropic behaviour, particularly in structured regions (Figure 21d and 18d). Nevertheless, since PI vastly stayed the same for all conditions, these excipients did not appear to alter previous observations of residue dynamics. On the other hand, they did alter residues with significant increases in solvent accessibility/dynamics and changes in microenvironment to some extent. Of note, 25 mM histidine resulted in a much higher concentration of residues in the "switch" mechanism region that were in the top 15 of PI % increase (C36-H43, Table 4b). This appears to play a role in its impact on bioactivity, discussed in section 5.1.5.

Although dynamics stayed the same for residues relative to each other in different excipients, global dynamics from the control to the excipients did change. In Figure 22a, ∆PI is calculated with control maximum PI – excipient maximum PI, thus, a positive differential would mean that the control PI was higher. The overwhelming positive differential for all conditions shows that the addition of excipients decreased global dynamics. Furthermore, 25 mM histidine and 25/50 mM arginine generally showed higher ∆PI than 12.5 mM phenylalanine. The main peaks in maximum PI occurred in the same areas for all excipients, such as; T1, V48 and Q120. The change in PI over the melt for these residues were compared for all conditions in Figure 22b-d. The change in dynamics with temperature for these residues was very similar for the control and 12.5 mM phenylalanine. The same cannot be said for other excipients as PI curves for these were much lower and flattened.

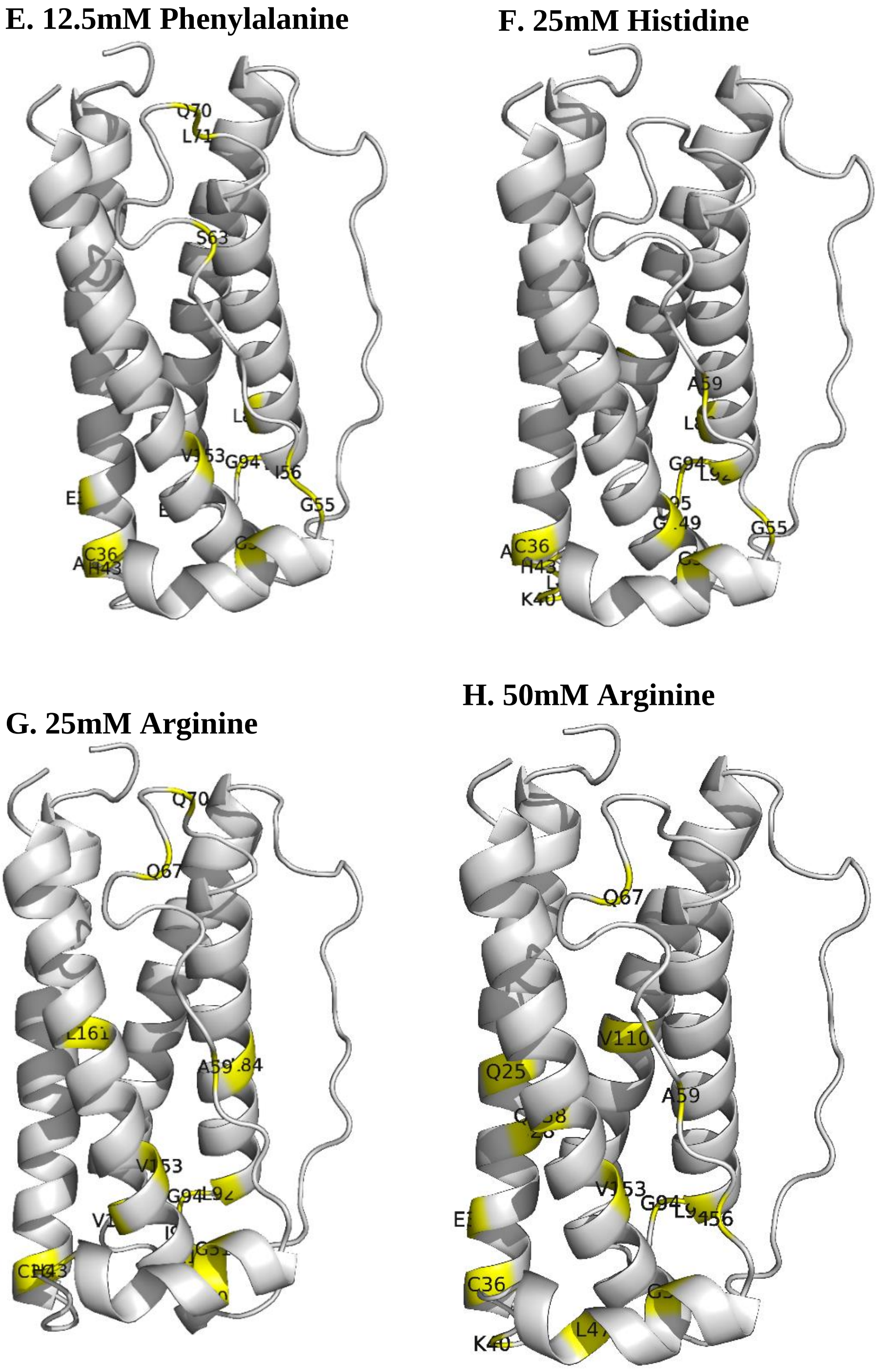


A. 12.5mM Phenylalanine

B. 25mM Histidine

C. 25mM Arginine

D. 50mM Arginine



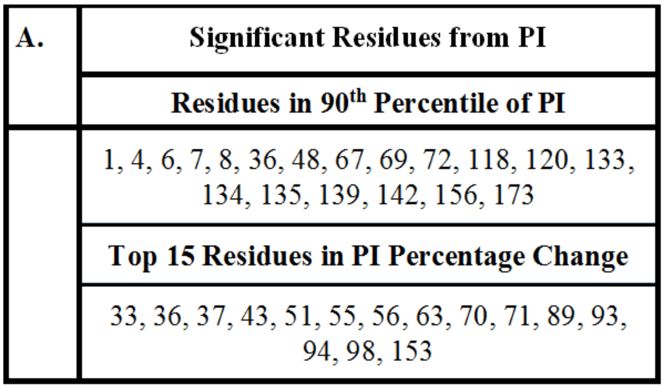
E. 12.5mM Phenylalanine

F. 25mM Histidine

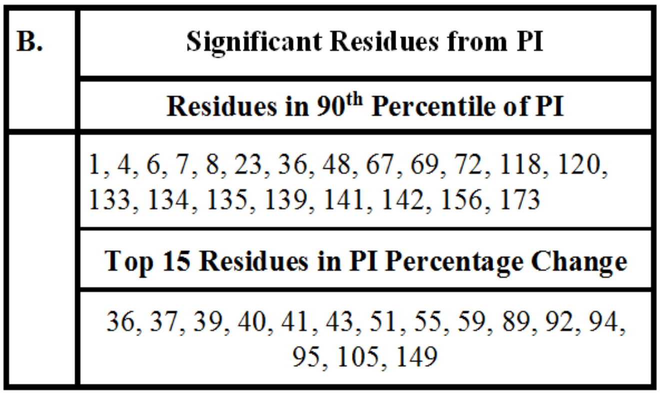
H. 50mM Arginine

G. 25mM Arginine

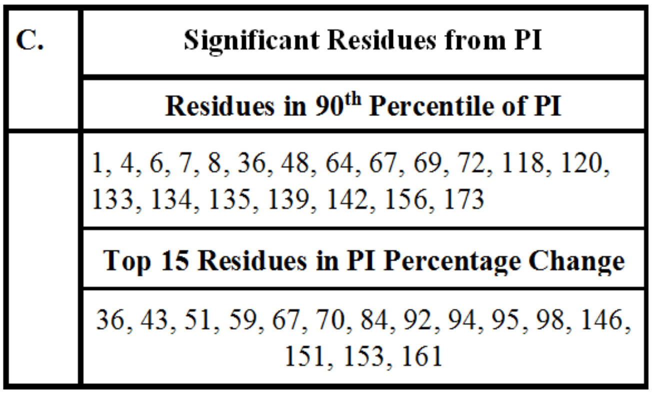
Figure 21. Mapping Regions Significant to PI onto G-CSF. A.-D. Residues in the 90th percentile for maximum PI (red) and **E.-H.** top 15 residues with the highest percentage increase in PI (yellow).



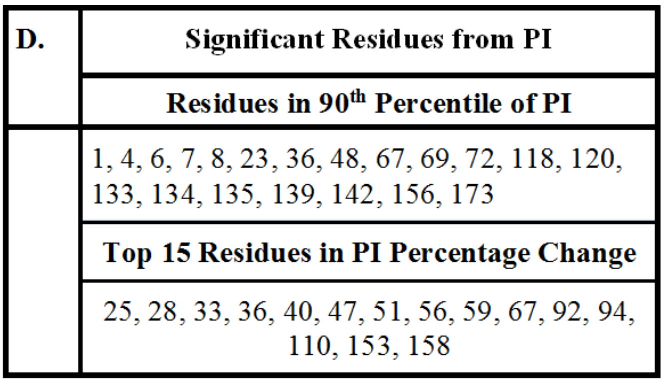
A.	Significant Residues from PI
	Residues in 90th Percentile of PI
	1, 4, 6, 7, 8, 36, 48, 67, 69, 72, 118, 120, 133, 134, 135, 139, 142, 156, 173
	Top 15 Residues in PI Percentage Change
	33, 36, 37, 43, 51, 55, 56, 63, 70, 71, 89, 93, 94, 98, 153



B.	Significant Residues from PI
	Residues in 90th Percentile of PI
	1, 4, 6, 7, 8, 23, 36, 48, 67, 69, 72, 118, 120, 133, 134, 135, 139, 141, 142, 156, 173
	Top 15 Residues in PI Percentage Change
	36, 37, 39, 40, 41, 43, 51, 55, 59, 89, 92, 94, 95, 105, 149

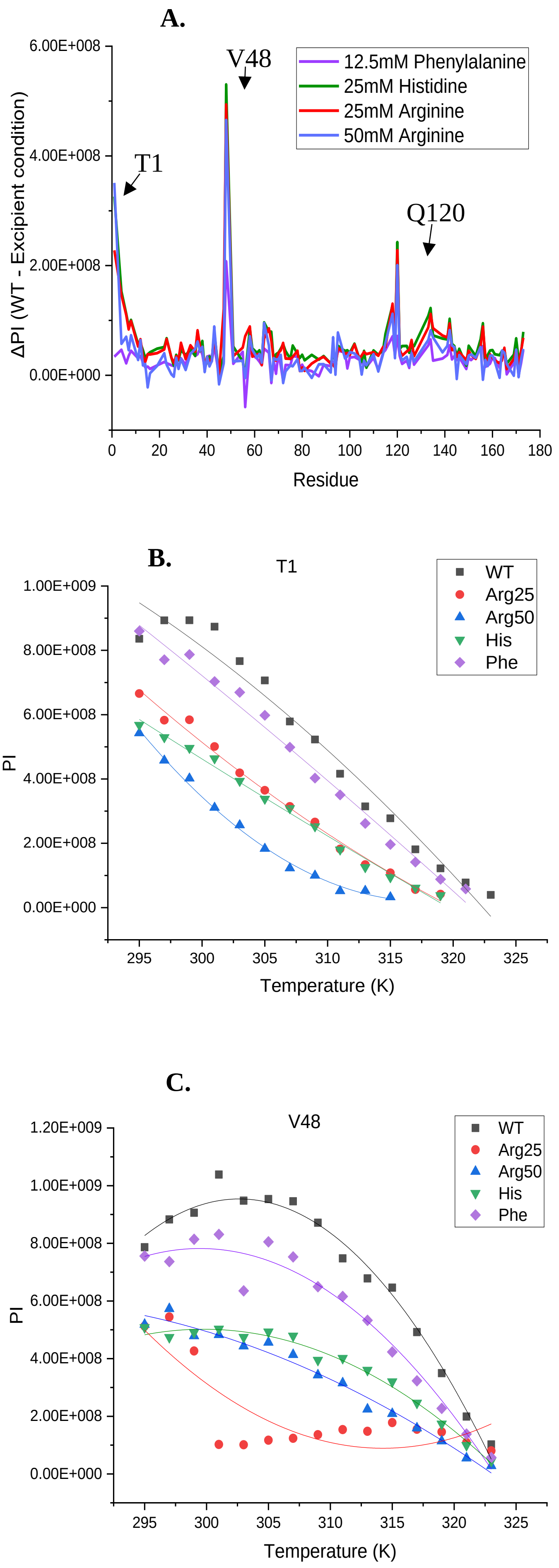


C.	Significant Residues from PI
	Residues in 90th Percentile of PI
	1, 4, 6, 7, 8, 36, 48, 64, 67, 69, 72, 118, 120, 133, 134, 135, 139, 142, 156, 173
	Top 15 Residues in PI Percentage Change
	36, 43, 51, 59, 67, 70, 84, 92, 94, 95, 98, 146, 151, 153, 161



D.	Significant Residues from PI
	Residues in 90th Percentile of PI
	1, 4, 6, 7, 8, 23, 36, 48, 67, 69, 72, 118, 120, 133, 134, 135, 139, 142, 156, 173
	Top 15 Residues in PI Percentage Change
	25, 28, 33, 36, 40, 47, 51, 56, 59, 67, 92, 94, 110, 153, 158

Table 7. Regions Significant to PI for Excipients. A.-D. represent 12.5mM phenylalanine, 25mM histidine, 25mM arginine and 50mM arginine respectively. Same table structure as with table 4b.


A.
ΔPI (WT - Excipient condition)
6.00E+008
4.00E+008
2.00E+008
0.00E+000
0
20
40
60
80
100
120
140
160
180
Residue
T1
V48
Q120
12.5mM Phenylalanine
25mM Histidine
25mM Arginine
50mM Arginine
B.
T1
PI
1.00E+009
8.00E+008
6.00E+008
4.00E+008
2.00E+008
0.00E+000
295
300
305
310
315
320
325
Temperature (K)
WT
Arg25
Arg50
His
Phe
C.
V48
PI
1.20E+009
1.00E+009
8.00E+008
6.00E+008
4.00E+008
2.00E+008
0.00E+000
295
300
305
310
315
320
325
Temperature (K)
WT
Arg25
Arg50
His
Phe

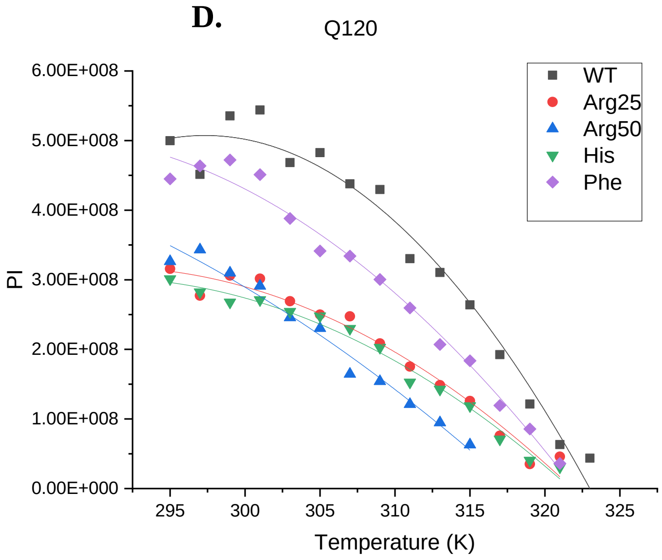


Figure 22. The Effect of Excipients on Dynamics. A. WT PI – Excipient PI calculates ΔPI. **B.-D.** Illustrates the effect of excipients on PI over the melt for residues T1, V48 and Q120 respectively.

4.1.5 Examining changes in NMR Δδ and PI Correlation in the Context of Coevolution

Figure 23 depicts a correlation between assigned residue Δδ and PI with temperature, as previously with Figure 11, for all excipient conditions. 12.5 mM phenylalanine, 25 mM histidine and 25 mM arginine all showed similar Δδ correlation clusters to the control (Figure 23a), these being; within T1-S12, within T133-A139 and between G1-S12 and T133-A139. PI CCM ((Figure 23b) echoed this same pattern but also included columns/clusters of correlation at S66-A68 (especially with T133-A139) for 25 mM histidine and 25 mM arginine. Residues E98-T102 also possessed strong correlations with T38-C42, S66-A68 and T133-A139 in 25 mM arginine.

Therefore, the consistently strong correlations between the N-terminus, the C-terminal end of loop AB and the centre of loop CD reinforced the notion that they become structurally modified in a concerted manner over the thermal melt. A suggested reasoning for this, discussed in chapter 3, may also be supported by the close coevolution relationship between these regions, as depicted in Figure 20. Here, a coevolution CCM (Pearson's correlation) weighted by phylogeny (Figure 20a) was produced in CoeViz (Baker and Porollo, 2016) and revealed how closely residues were related in G-CSF evolution. The deeper the red colour, the stronger the positive correlation and the deeper the blue, the stronger the anti-correlation. Moreover, the stronger the "self-correlation" (on the diagonal line) the more conserved a residue is. Coevolution was very sparse, while the accompanying phylogenetic tree (Figure 20b) showed a close relationship between T1-S12 and T133-A139, as highlighted by green arrows. All residues between L130 and A139 (aside from

P132 and A136) were closely related to a residue in the T1-S12 region, with correlation above 0.34 for F13 and P138 (Table 8). Hence, an epistatic nature may have been revealed between these regions, which may be conducive to better molecule fitness when jointly mutating both areas (described in section 5.1.4).

However, the correlation in 50 mM arginine was very distinct from that of the other conditions. Here, the colour gate for Δδ was widened (from -0.67-0.67 to -0.93-0.93) to reduce noise. Even after widening the colour gate there were still many regions of strong correlation, the clearest of which was for residue L92. A distinct column of strong correlation occurred between this residue and a large proportion of other regions. Some regions that showed strong correlation with L92 (such as K16 to T38 and D104 to G125) also showed clusters of correlations with other regions in the structure.

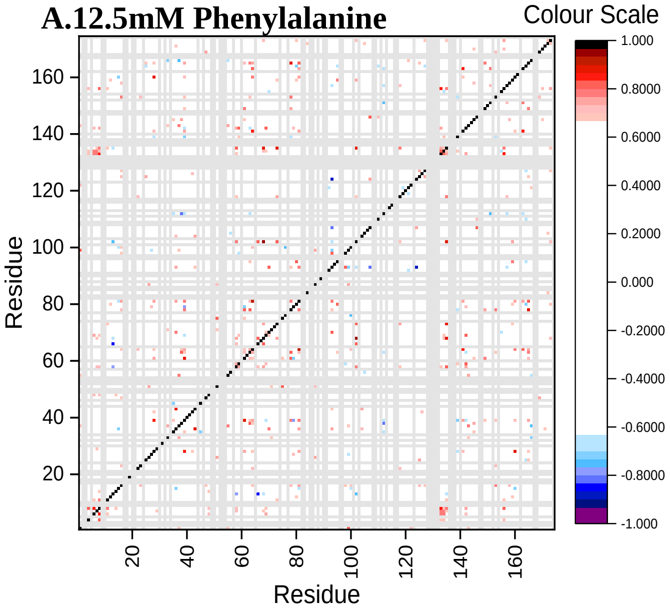
Although, some correlation occurred within T1-S12 and T133-A139 regions, correlation in general for PI was very sparse: an opposite trend to that for all other excipient and control conditions. Residue T1 also had a column of correlation but it was anti-correlation. This coincided with observations from Figure 22b where the change in PI over temperature for T1 mirrorred more of an exponential decay curve rather than the typical 2nd order polynomial.

Thus, the significance of L92 in the destabilising mechanism of arginine was emphasised by this Δδ CCM and echoed observations from Figure 20d. The mechanism being proposed here is that E93 and E98 preferentially interact with the arginine in solution (at 50 mM) instead of the neighbouring R146 and K34 (Figure 20b). Consequently, disrupting the interaction between E93-R146 and E98-K34 would disrupt electrostatic interactions that help pull all four helices together. This would render many hydrophobic residues exposed, thus explaining the dominance of buried residues affected by 50 mM arginine (Figure 20d). Interaction at E93 could explain the significance of the neighbouring L92 on structure and dynamics. Excipient-induced Δδ for L92 was not relatively high for any condition but was highest at 50 mM arginine (~0.02 Δppm for phenylalanine and histidine and ~0.05 Δppm for 50 mM arginine). Furthermore, given that L92 was one of most buried residues, this environmental change at 50 mM arginine could have disrupted significant hydrophobic pockets such as that composed of residues L92, G94 and I95 (Figure S.8), which would drive unfolding given the importance of hydrophobic interactions.

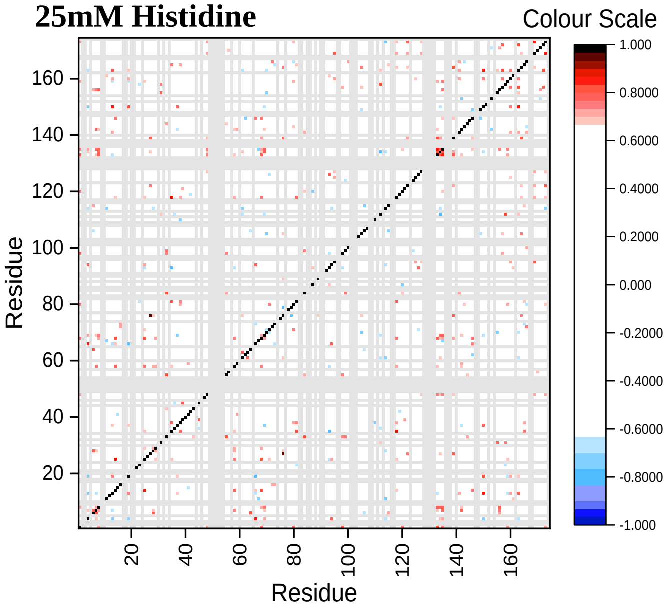
NMR observables have shown that, unlike the other structurally temperature-sensitive regions (N-terminus, loops AB/CD), L92-I95 was also not significantly dynamic (Figure 9a, 17, 21a-d and Table 4/6). Conversely, L92-I95 likely experienced a significant opening motion in all conditions, unlike the other regions (Figure 21e-h). This opening motion was accelerated by 50 mM arginine.

CCM for Δδ

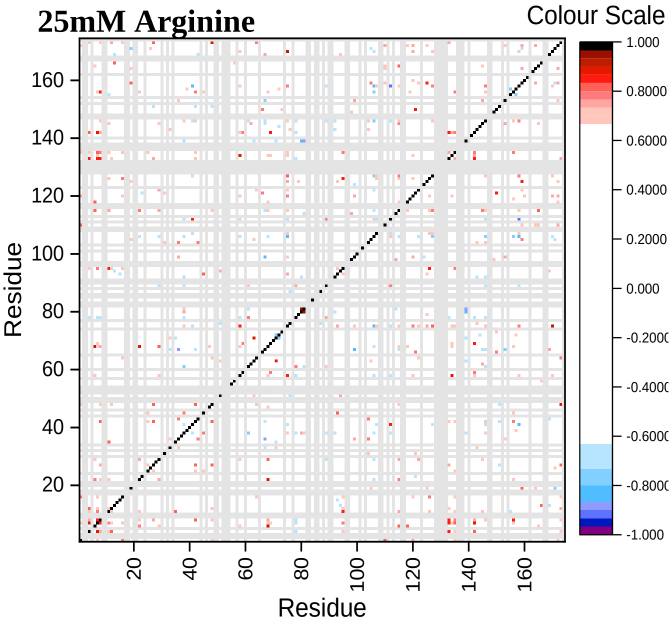
A.12.5mM Phenylalanine



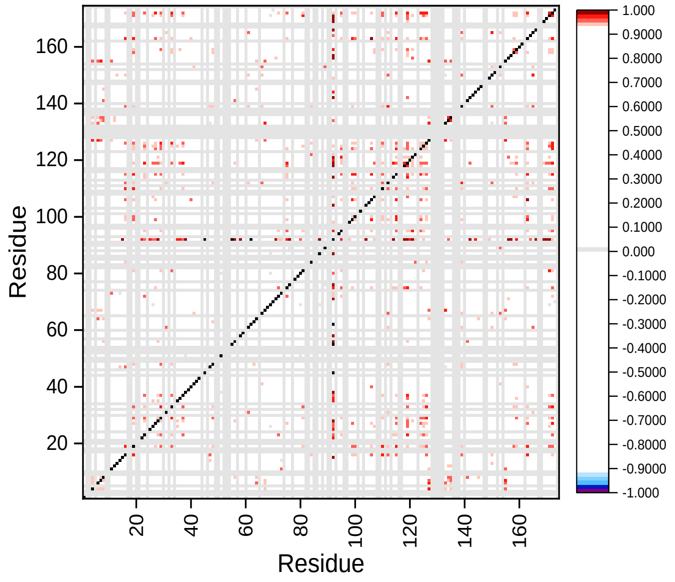
25mM Histidine



25mM Arginine

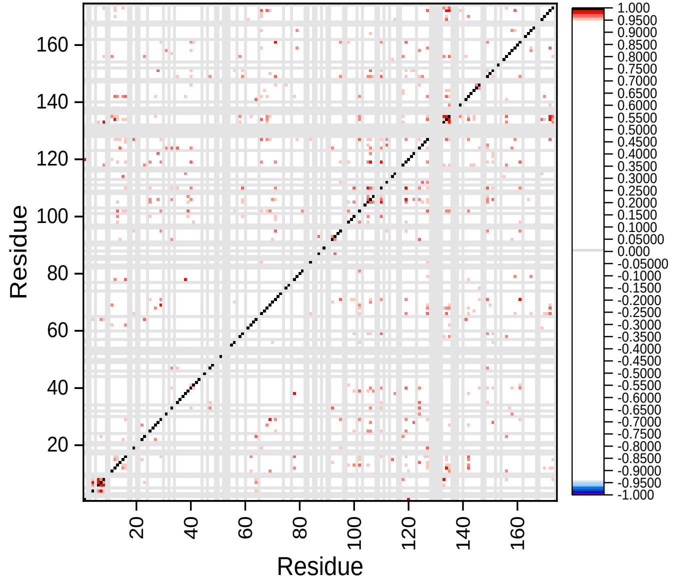


50mM Arginine

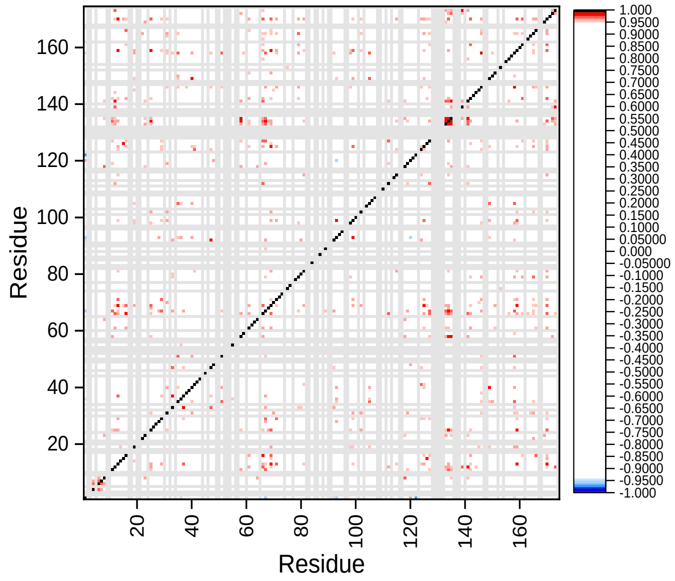


CCM for PI

B.12.5mM Phenylalanine



25mM Histidine



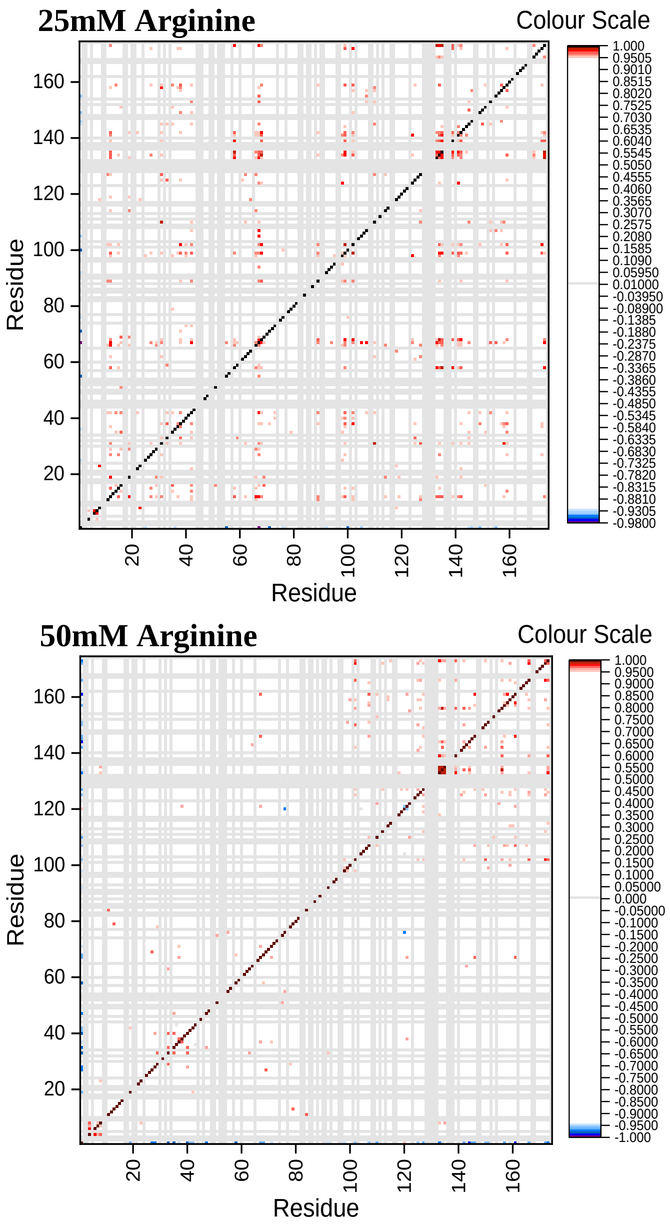


Figure 23. Cross-correlation matrices for A) Δδ and B) PI In the Presence of Excipients. Refer to Figure 11 for colour gating. The colour gate for Δδ is widened (-0.93 to 0.93) for 50mM Arginine.

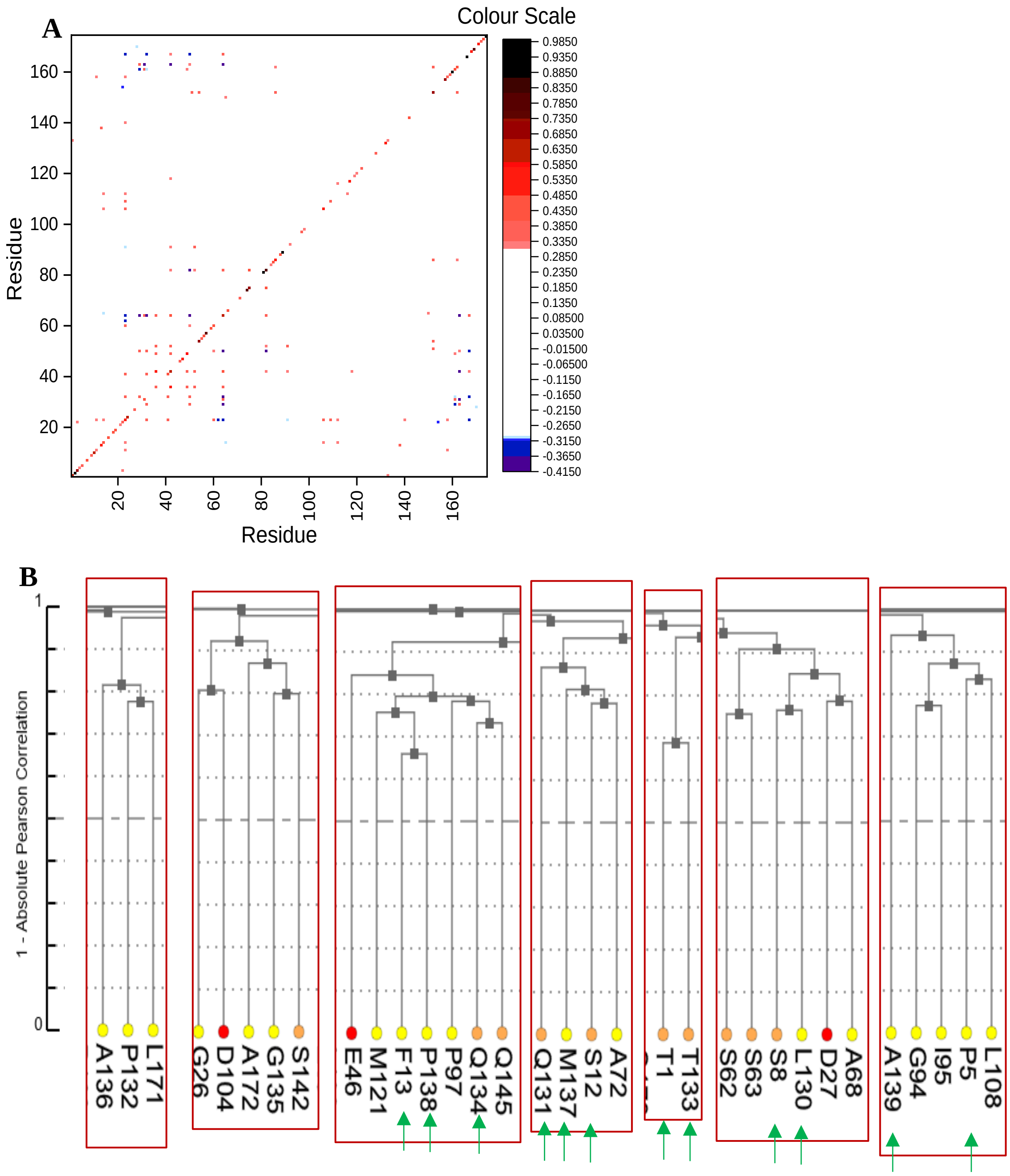


Figure 24. Coevolution of G-CSF Residues. A. Coevolution Heatmap by Phylogeny with Pearson correlation above 0.31 on the red-black scale and below -0.31 on the blue-purple scale. A white gate between these scales was added to reduce noise. **B.** Phylogenetic Tree highlighting regions T1 to F13 and L130 to A139 in red boxes and green arrows (CoeViz; Baker and Porollo, 2016).

A.	Residue										
	13	23	29	32	36	42	49	52	64	82	152
Residues with a Pearson Correlation above 0.34	138	32	32	23	42	36	36	42	36	64	51
		60	50	29	49	49	42	91	42	75	54
		106	163	41	64	52			82		86
		109				64			167		162

B.	Residue				
	23	29	50	64	163
Residues with a Pearson Correlation Below -0.34	64	64	64	23	31
	167	161	82	29	42
				32	64
				50	
				163	

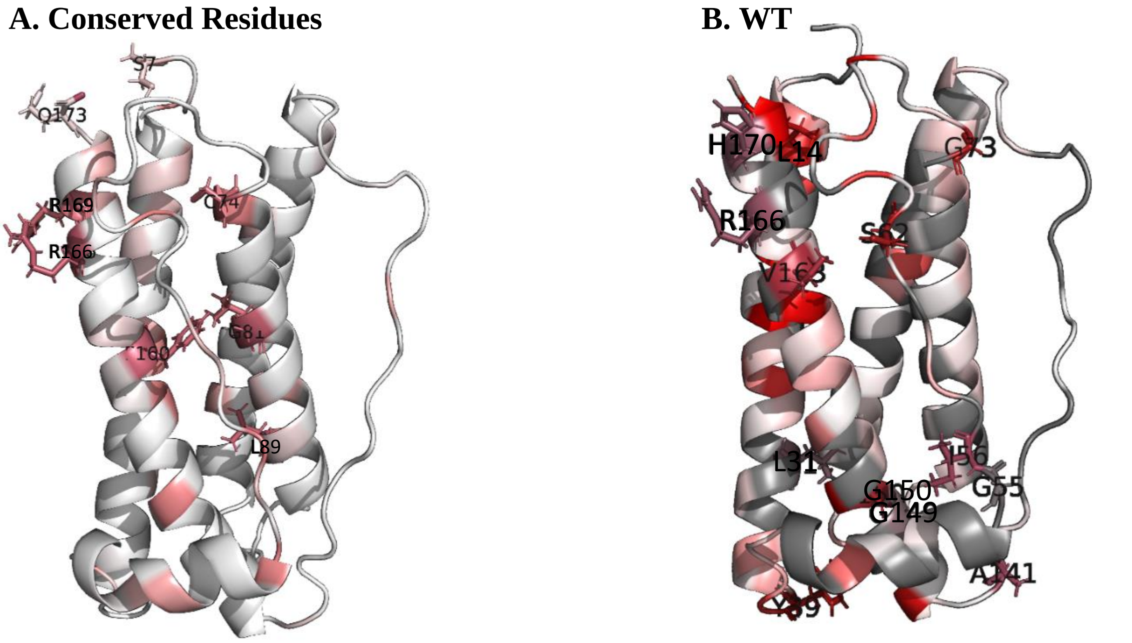
Table 8. Residues Displaying Strong Correlation in Coevolution Heatmap. Residues with a positive correlation above 0.34 **(A.)** and those with a negative correlation below -0.34 (**B.**).

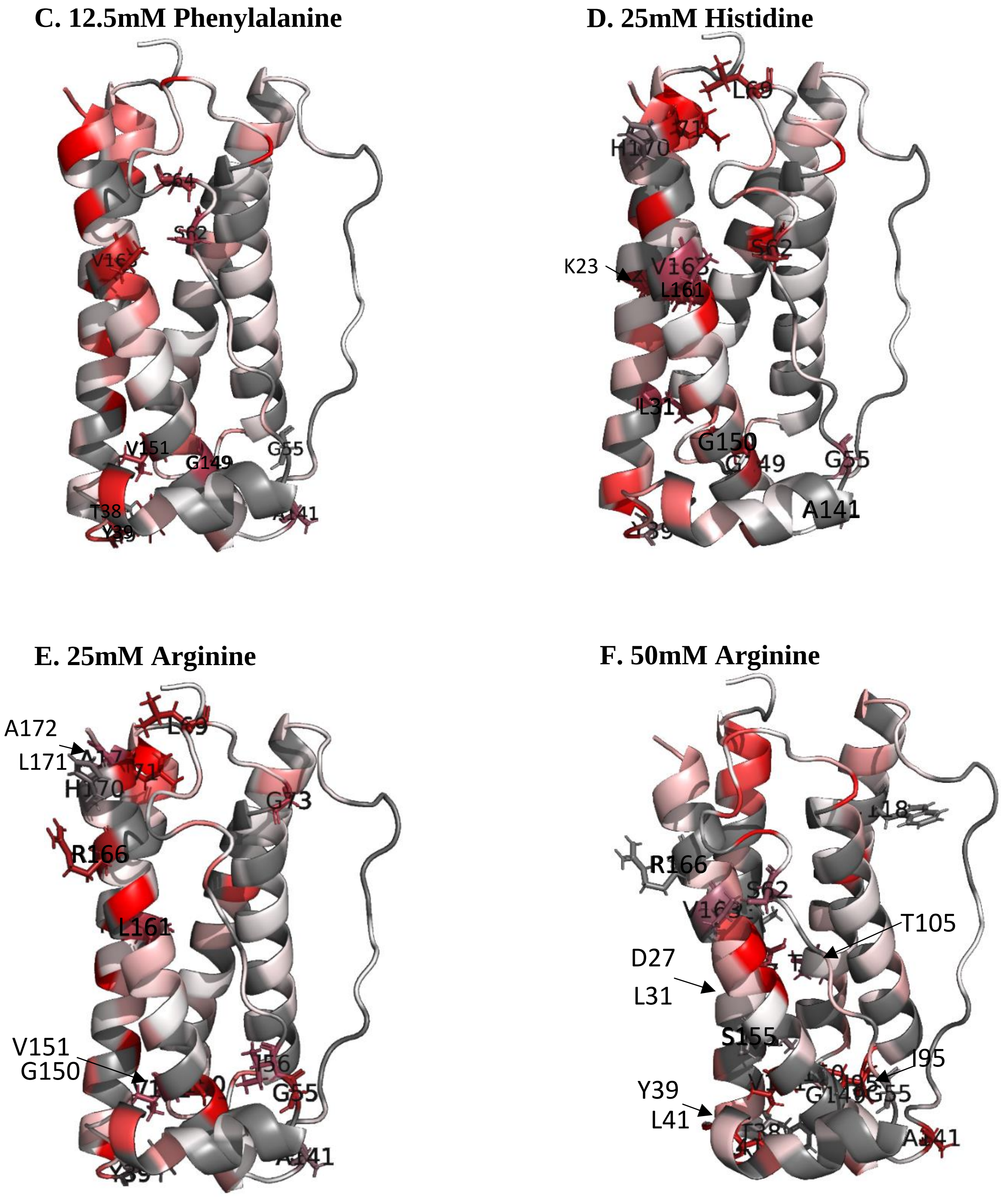
4.1.6 Linearity and $\Delta\delta_H/\Delta T$ Reveals Changes in Thermally-Induced Structure Remodelling

Conserved residues from Figure 24a were mapped on to G-CSF in Figure 25a, with those shown as sticks and coloured darker red/raspberry representing the most conserved (95th percentile for self-correlation). These highly conserved residues were concentrated around helix D and B, with moderately conserved residues populating the short helix and the rest of loop AB. These residues were not directly involved with functionality and for the most part were buried (I24, L54, L75, G81, L82, L89, L152, L157, F160). Previous studies have demonstrated the importance of conserved, hydrophobic cores in protein folding (forming “folding nuclei”) owing to their tight packing and stress resilience (Liao *et al*., 2005; Ptitsyn and Ting, 1999). Sequence conservation has also been shown to delineate functionally significant residues (Lichtarge *et al*., 1996; Ma *et al*., 2003), an observation echoed in Figure 24a with some of the mildly conserved residues overlapping with functional ones (K16, E19, K23, L41 and L49).

Peak trajectory linearity was calculated here, as previously described in Table S.1 and Figure S.5, to probe whether residues explored more/less complex pathways with different excipients as their microenvironment changed. The majority of residues were linear under all experimental conditions, however, this became less clear for increasing concentrations of arginine (Figure 25b-f and S.11). Regions of distinct non-linearity (highlighted as sticks in Figure 25b-f) concentrated around highly similar regions for all conditions, those being; loopAB and helix D (with the exception of loop BC and helix A/C for 50 mM arginine). Some of these residues overlapped with (were in close proximity to) very conserved residues in the C-terminal region of helix D (F160, R166, H170). A low linearity, ΣΔδ and maximum PI would be potential key attributes of highly conserved residues that are important to protein folding and stress resilience. This is because there would not be a simple exchange between populations over the melt and so total microenvironmental change and dynamics would be kept minimal to protect the less resilient areas. Although there was poor correlation between linearity and conservation (Figure S.11), residues K23, I56, R166 and L171 could act as cores of high stress resilience because they were conserved with a low linearity, ΣΔδ (aside from I56 which is moderate) and maximum PI (Figure S.12). Hydrogen bonding was also strong in these regions ($\Delta\delta_H/\Delta T$ values > -3.5 ppb/K).

Linearity varied between excipient conditions and was calculated as Δlinearity (control linearity – excipient linearity) in Figure 26. Therefore, a positive differential, represented by the red arrow, signified a more non-linear trajectory when the respective excipient was added, while the green arrow signified the opposite. ΔLinearity was generally moderate for 25 mM histidine and arginine (i.e. between -0.5 and 0.5) but was larger for 12.5 mM phenylalanine and 50 mM arginine. More residues had a large increase in linearity (L31, I56, R166 and H170) than a large decrease (T38) for 12.5 mM phenylalanine. The opposite was the case for 50 mM arginine where residues K23, T38, I95, T105, W118 and S155 displayed a large decrease in linearity and only residues I56 and H170 showed a large increase. This was evident in Figure 25f given that most residues were highlighted as sticks at this condition. Of note, the highly conserved residue R166 only became more non-linear at the destabilising condition of 50 mM arginine and became largely more linear at the stabilising conditions of 12.5 mM phenylalanine and 25 mM histidine.




C. 12.5mM Phenylalanine

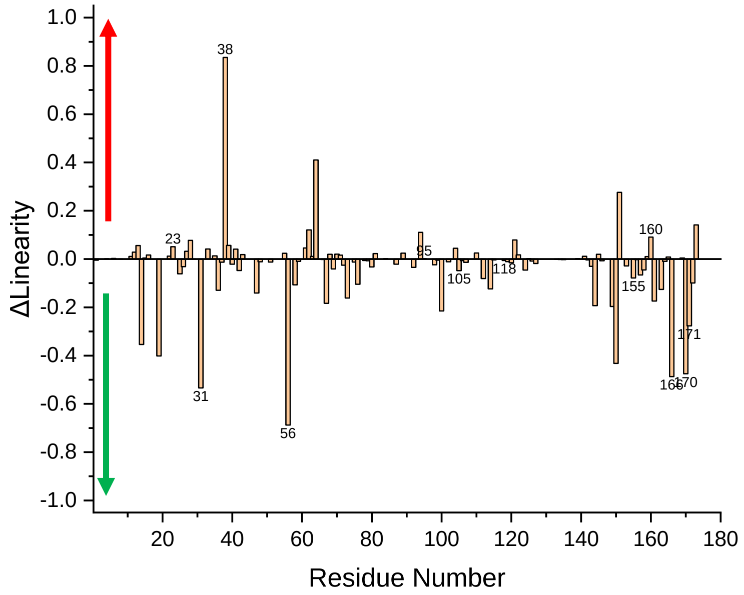
D. 25mM Histidine

E. 25mM Arginine

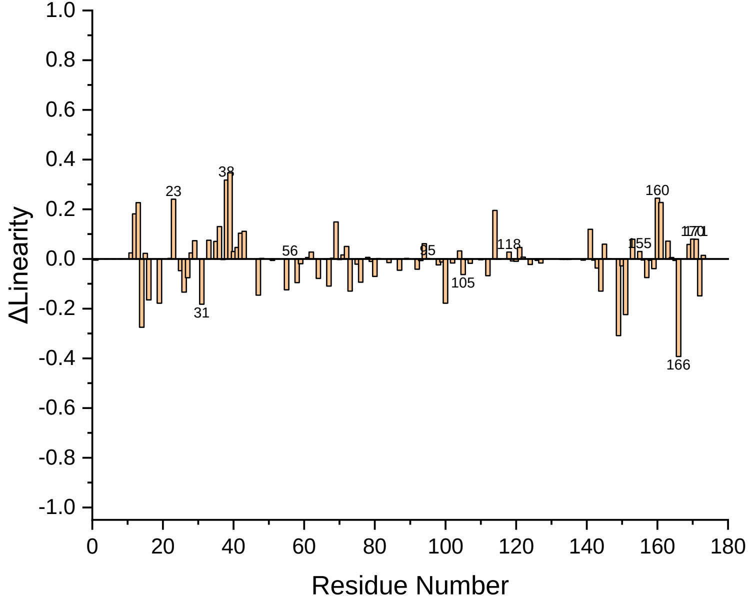
F. 50mM Arginine

Figure 25. Peak Trajectory Linearity. A. Conserved residues determined by self-correlation in CoeViz coevolution analysis are coloured on a scale of white to raspberry. Residues with highest self-correlation are highlighted as sticks and labelled. **B.-F.** NMR Peak trajectory linearity over the melt. The colour scale is white (linear) to Raspberry (non-linear) to grey (unassigned). Residues with a linearity below 0.5 are highlighted as sticks and labelled.

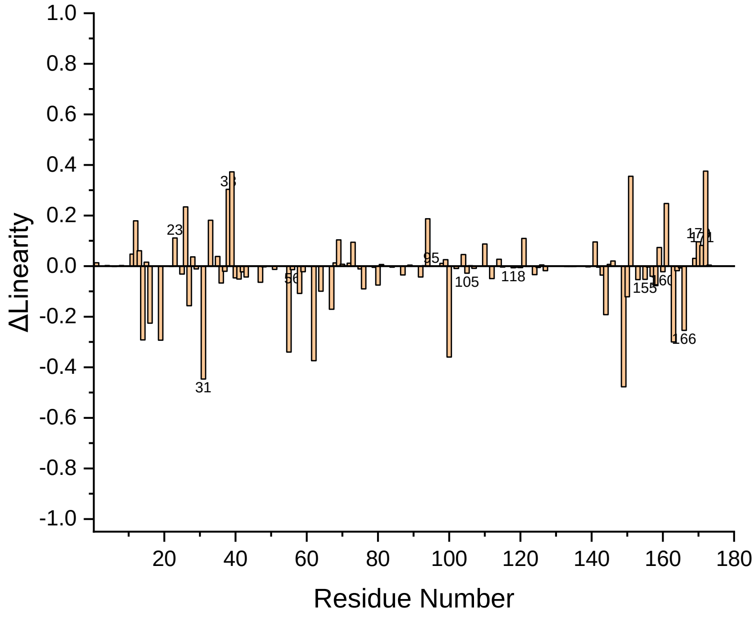
12.5mM Phenylalanine



25mM Histidine



25mM Arginine



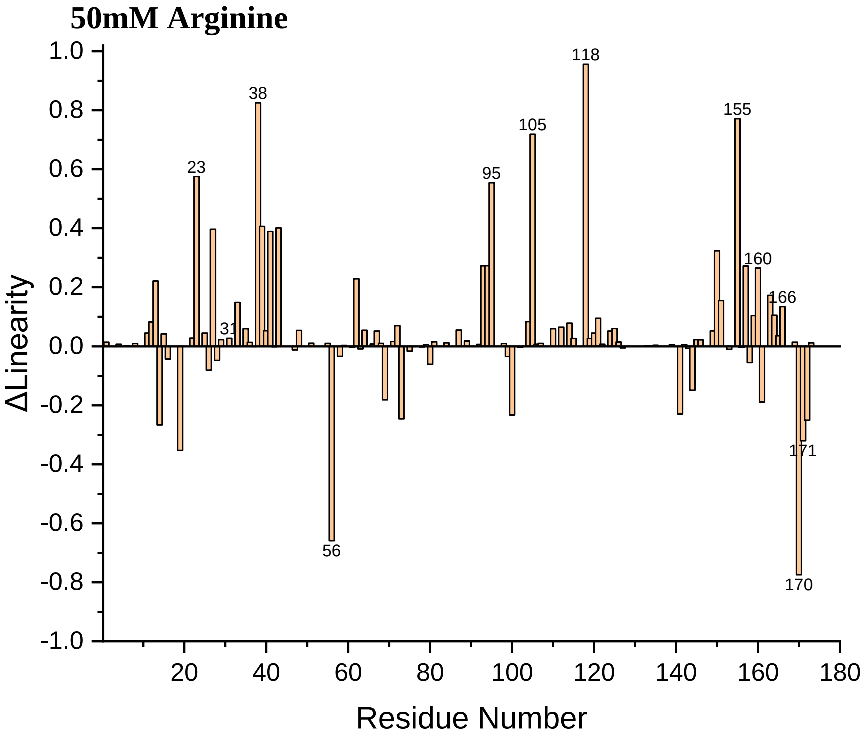


Figure 26. Linearity Changes with Excipient Condition. WT Linearity – Excipient Linearity calculates ΔLinearity. Residues are labelled above the bars. The green and red arrows represents more and less linear trajectories respectively.

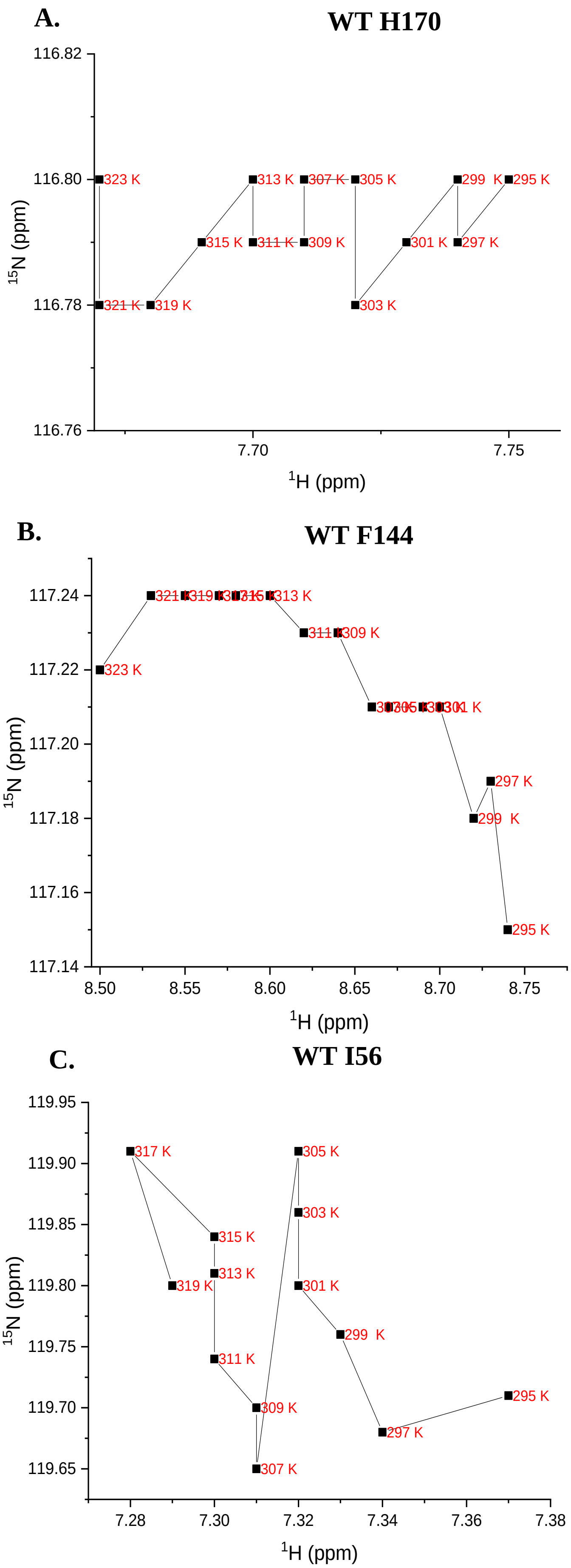
However, a caveat exists when analysing linearity because some residues were classed as having very low linearity, like R166 and H170 in the control condition, given that they have very little movement in one plane but regular movement in the other (Figure 27a and d). On the other hand, there were residues like F144 (Figure 27b) with low linearity and regular movement in both planes (and a high ΣΔδ, Table 4). Movement occurring markedly more in the ^{1}H plane instead of ^{15}N may yield information on amide hydrogen bond breaking (and reforming) to a larger extent than factors affecting ^{15}N discussed earlier. To examine this further, $\Delta\delta_H/\Delta T$ was calculated to reflect the strength of hydrogen bonding. Correlation (R^2) between $\Delta\delta_H/\Delta T$ and ΣΔδ at 323 K was strong for all conditions (Figure 28a), which may reflect the large contribution of hydrogen bond breaking to ΣΔδ. The R^2 value for 12.5 mM phenylalanine and 25 mM histidine was slightly higher than that for the control, but the R^2 for arginine was lower and decreased with higher concentrations. Conversely, $\Delta\delta_N/\Delta T$ correlated poorly with ΣΔδ at 323 K (graphs not included). Residues 92 and 106 deviate slightly from the linear trend for all conditions, lowering the R^2 value and indicating uncharacteristically high environment changes, despite strong involvement in hydrogen bonding.

Calculating $\Delta\delta_H/\Delta T$ may help circumvent the caveat of determining residues to be non-linear and not knowing whether it is due to large curvature in the ^{15}N or ^{1}H plane. This is because $\Delta\delta_H/\Delta T$ refers to curvature in Δδ in the ^{1}H plane. Therefore, where $\Delta\delta_H/\Delta T$ becomes more positive at a different excipient condition (i.e. a positive Δtempcoeff value), hydrogen bonding is stronger (Figure 28b). These Δtempcoeff values were moderate for 25 mM histidine and arginine (with the exception of H43 for 25 mM arginine). On the other hand, 12.5 mM phenylalanine and 50 mM arginine displayed relatively large Δtempcoeff in L15, H43, Q70, A72 and I95.

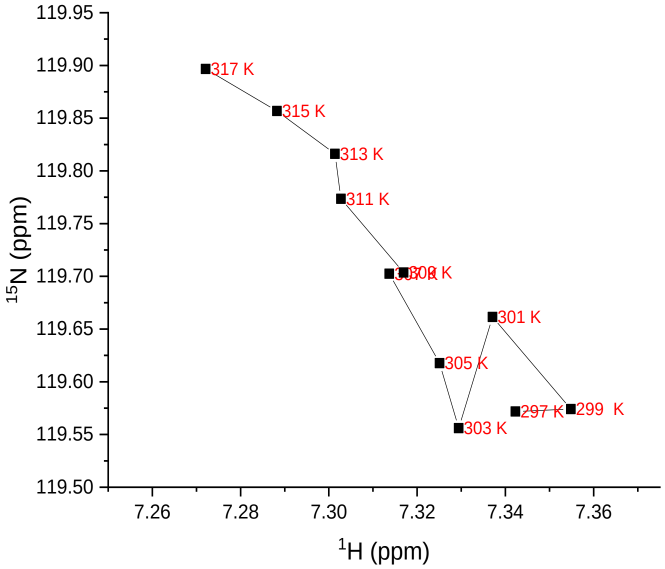
Peak position for H43 moved upfield in the ^{1}H plane for all conditions, with the control and 25 mM histidine moving to a similar extent, and the other conditions moving to a much lesser extent (Figure 27h, 28). Therefore, given that H43 was not in a structured region, this region could be more resilient to change/loss in structure in 12.5 mM phenylalanine and 25/50 mM arginine (Tomlinson and Williamson, 2012). Additionally, H43 had a general movement downfield in the ^{15}N plane under these excipient conditions, which was opposite to the control and in 25 mM histidine. With many factors influencing the evolution of magnetisation at the amide ^{15}N, the reason for this difference could be that 12.5 mM phenylalanine and 25/50 mM arginine induced alternate torsion angles and rotamer positions of H43.

Residue I95 showed the same trend between the control and 50 mM arginine where the peak moved much less upfield in the ^{1}H plane and moved in opposite fields in the ^{15}N plane (Figure 27g, 28). A potential interaction in this loop CD region at E93/98 was earlier observed in 50 mM arginine (Figure 20d), with a distinct linearity decrease in this loop, while PyMOL predicted the amide ^{1}H of I95 to hydrogen bond to L92. As a result, a protein-excipient interaction induced by 50 mM arginine could result in stronger hydrogen bonding between I95 and L92 as well as a torsion angle/ side chain rotamer change for I95. Alternatively, the amide ^{1}H on I95 could hydrogen bond with the arginine excipient, thus causing the aforementioned structural changes. Both mechanisms could open up the hydrophobic region in loop CD and contribute to the destabilising mechanism of 50 mM arginine.

The Δtempcoeff for residues Q70 and A72 showed clear positive and negative differentials for 12.5 mM phenylalanine, respectively (Figure 28). The diminished upfield ^{1}H plane movement and positive Δtempcoeff differential of residue Q70 in 12.5 mM phenylalanine compared to the control (Figure 27i, 28) hinted at stronger resilience to change/loss in structure, possibly due to stronger hydrogen bonding between S8 and Q70 (Figure S.9a). The opposite was the case for A72, where there was a large increase in upfield ^{1}H plane movement and negative Δtempcoeff differential (Figure 27j, 28). Residue A72 is just at the beginning of helix B and could, therefore, experience greater backbone hydrogen bond stretching/breaking and structural change at 12.5 mM phenylalanine compared to for the control. This may work cooperatively with Q70 to allow stronger hydrogen bonding. Additionally, there was a large increase in linearity for residue I56 at 12.5 mM phenylalanine (Figure 26) due to diminished variation in the ^{15}N plane (Figure 27c). Hence, by stabilising I56 conformational diffusion and strengthening S8-Q70 hydrogen bonding, the stabilising mechanism of 12.5 mM phenylalanine could be due to its highly concentrated impact around the C-terminus of loop AB.

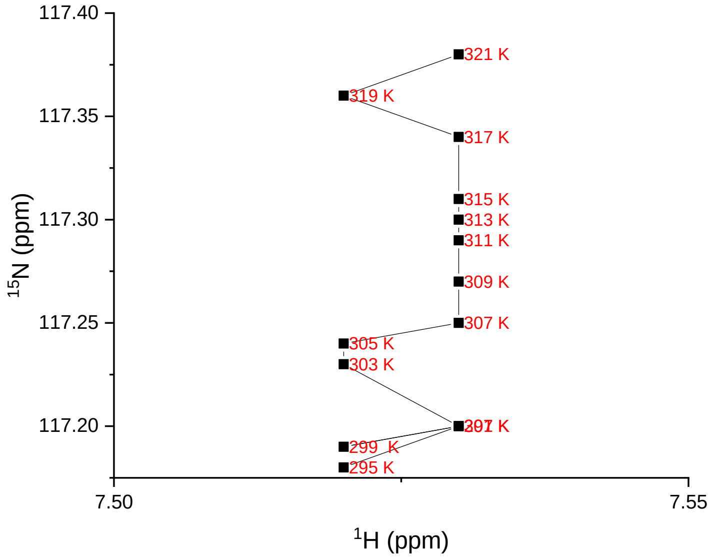


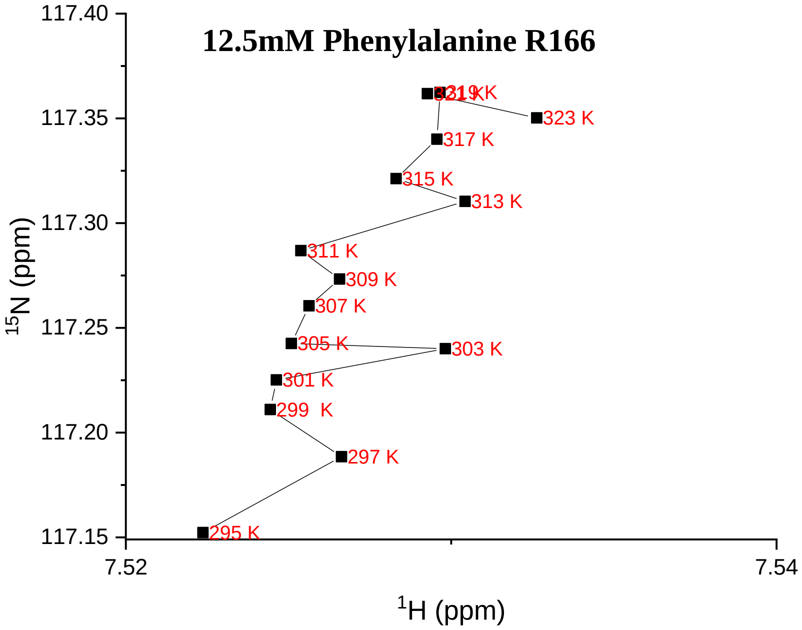
12.5mM Phenylalanine I56



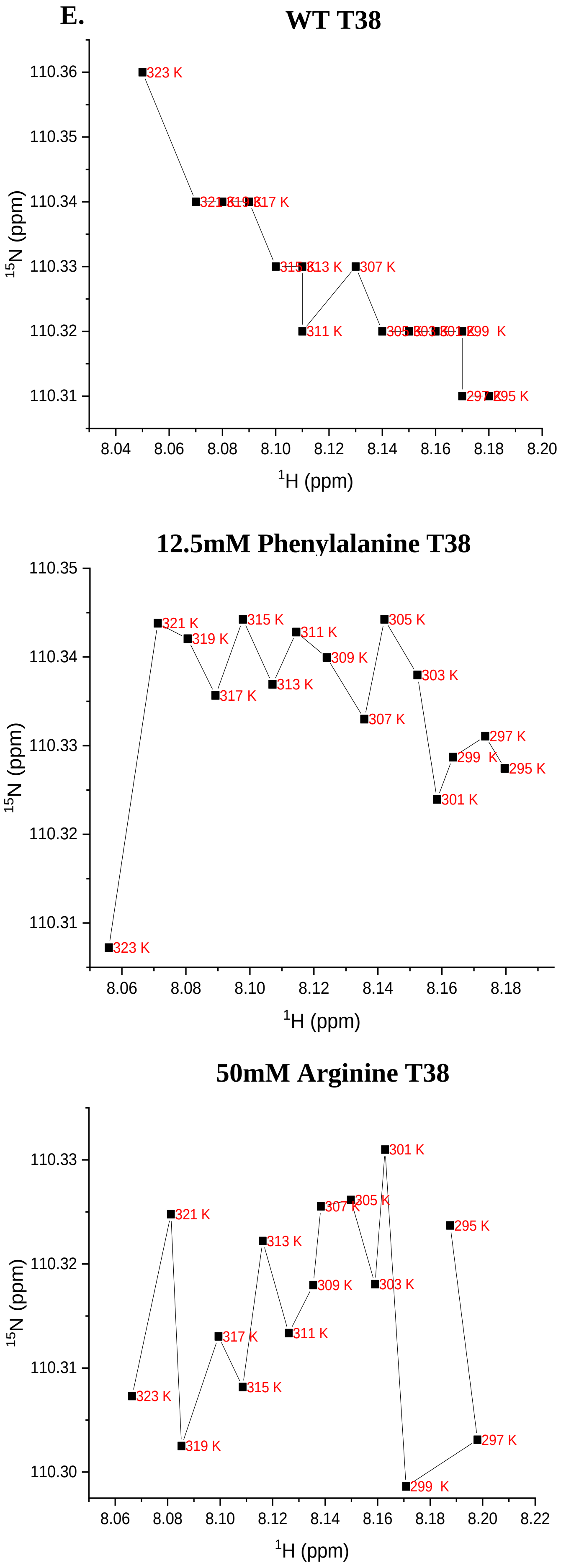
D.

WT R166





E.

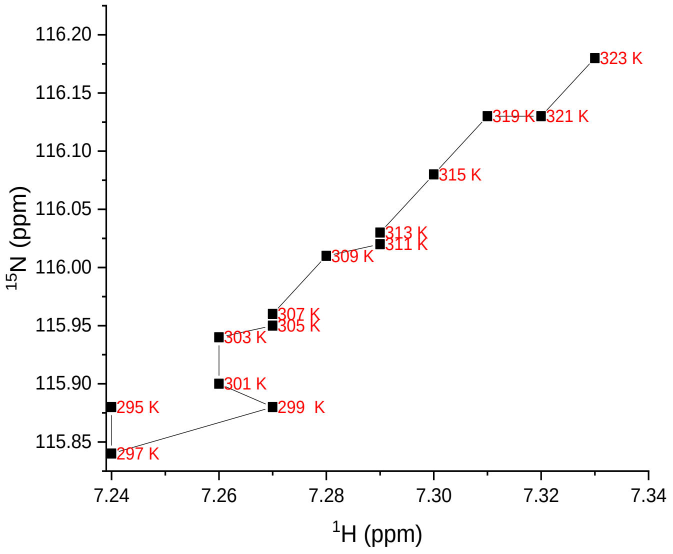


WT T38

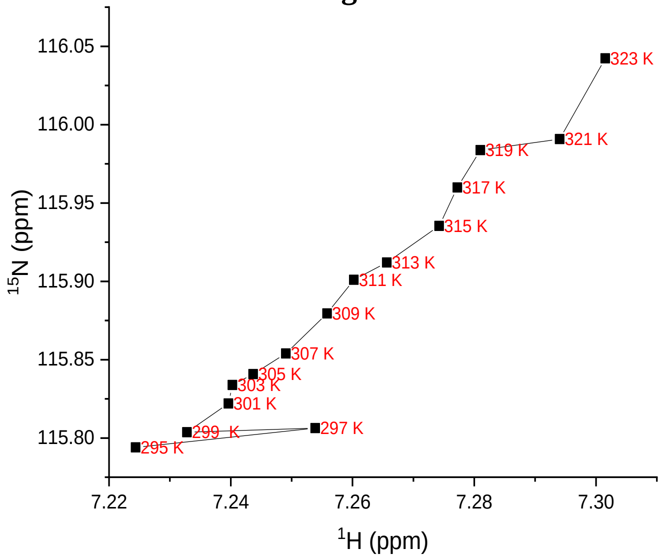
12.5mM Phenylalanine T38

50mM Arginine T38

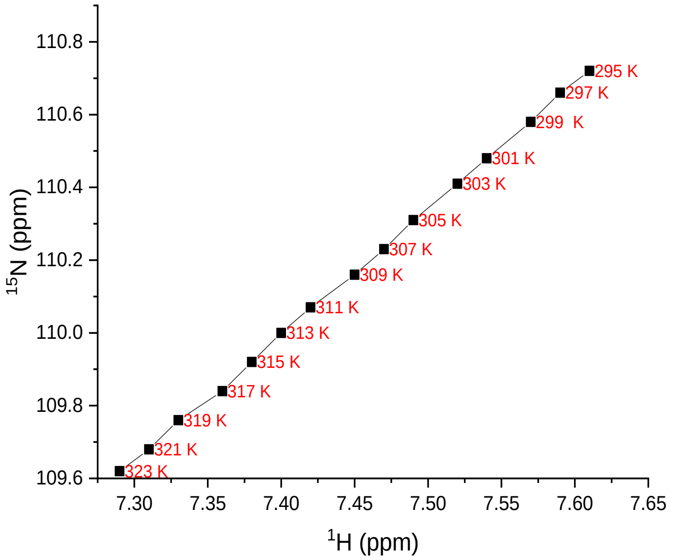
F. **WT L92**



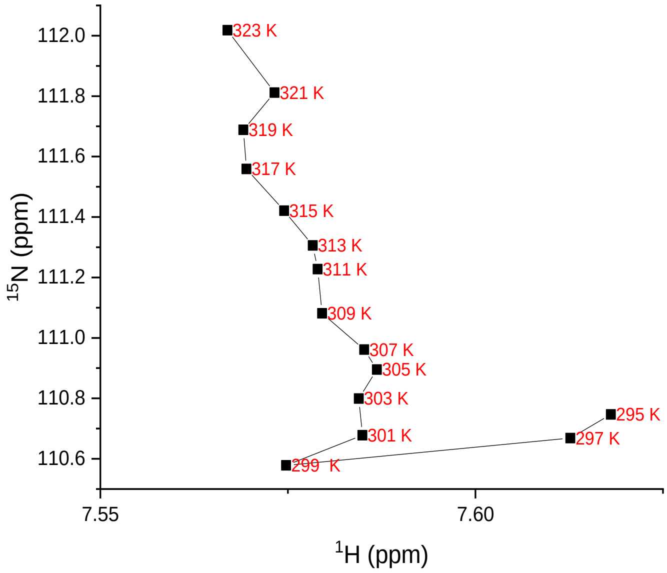
50mM Arginine L92



G. **WT I95**

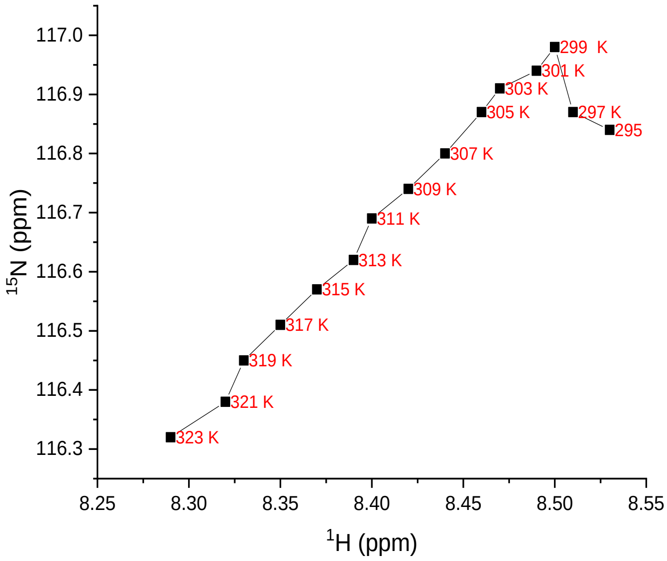


50mM Arginine I95

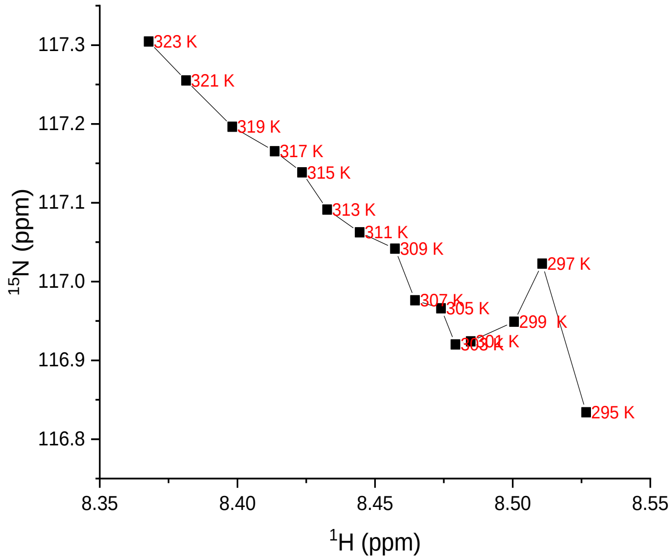


H.

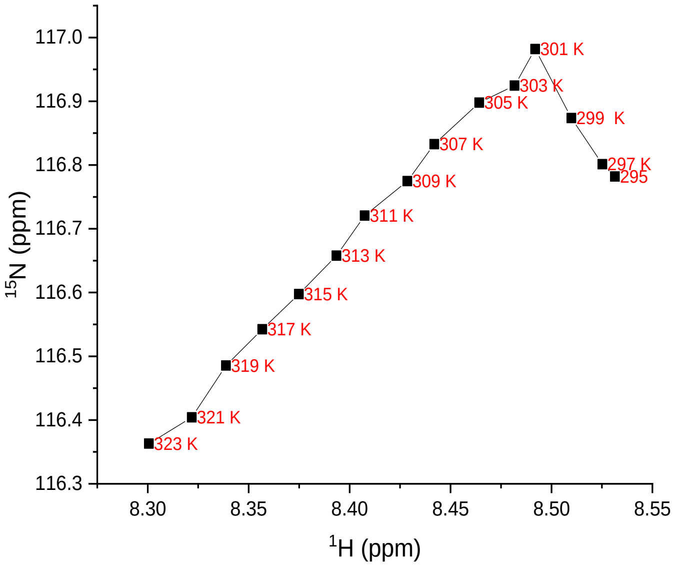
WT H43



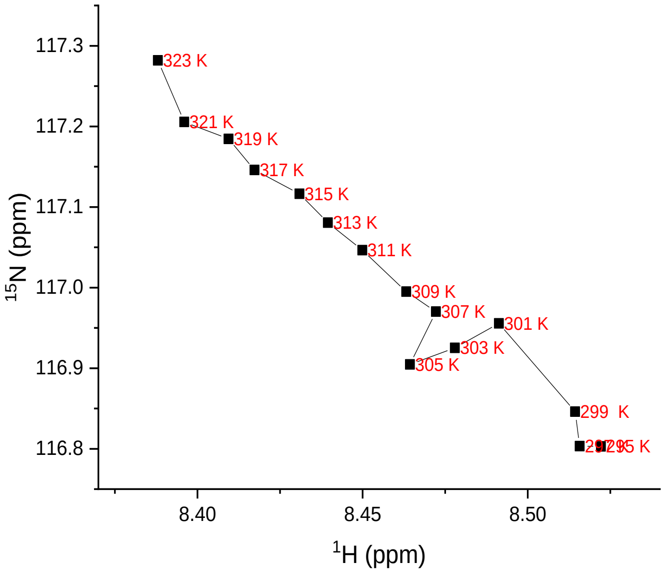
12.5mM Phenylalanine H43



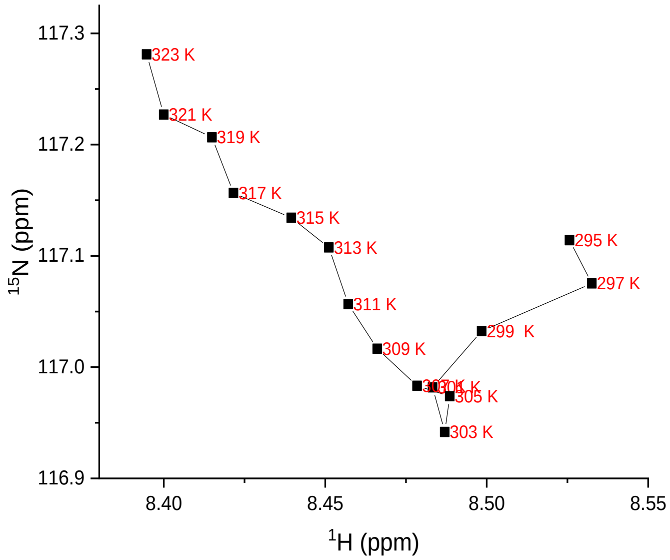
25mM Histidine H43



25mM Arginine H43

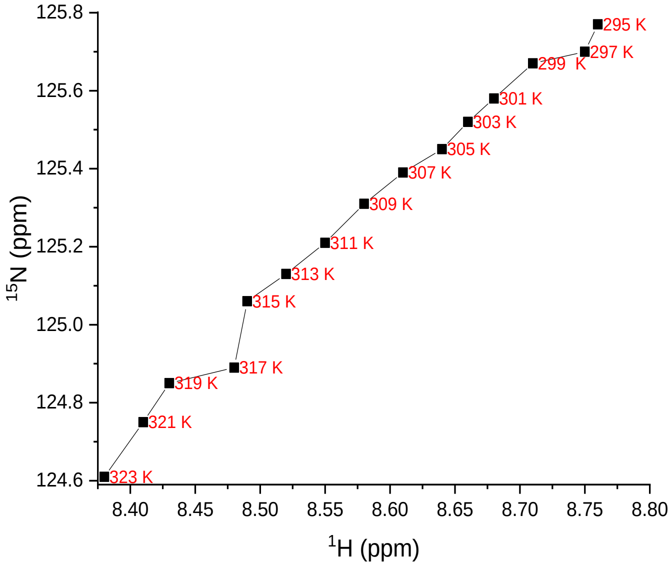


50mM Arginine H43

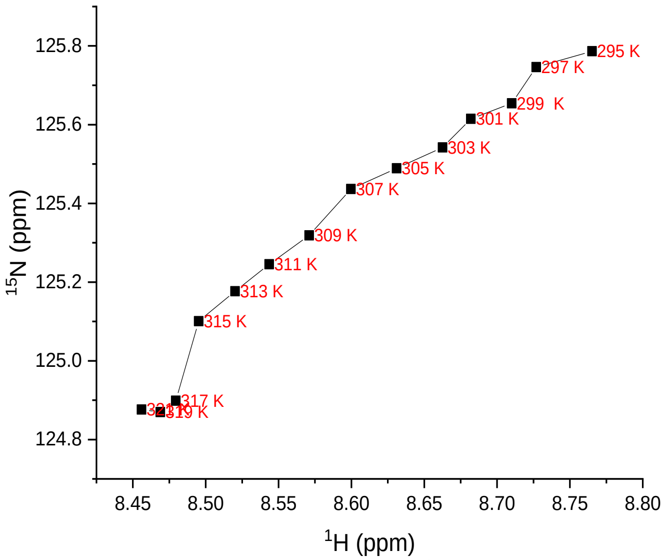


I.

WT Q70

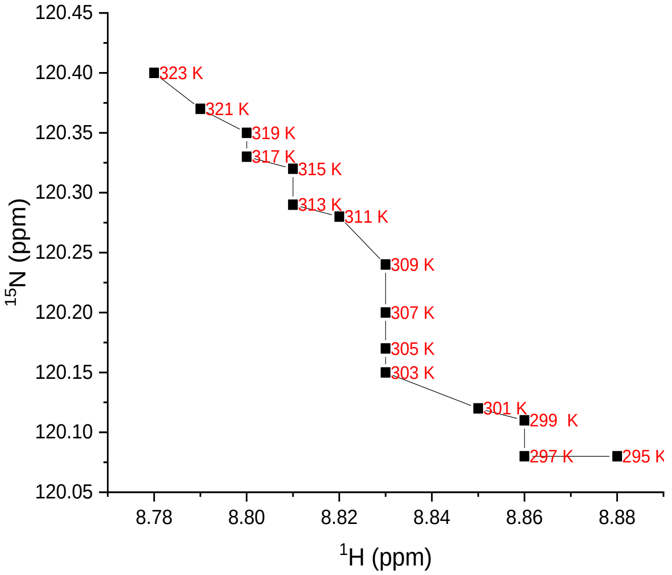


12.5mM Phenylalanine Q70



J.

WT A72



12.5mM Phenylalanine A72

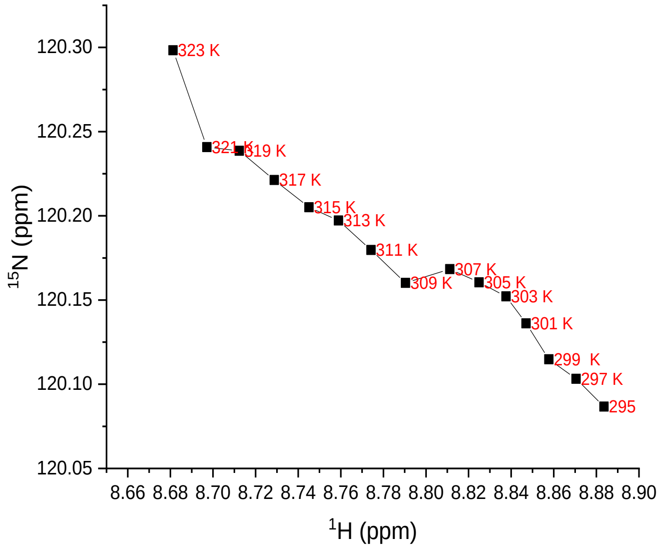
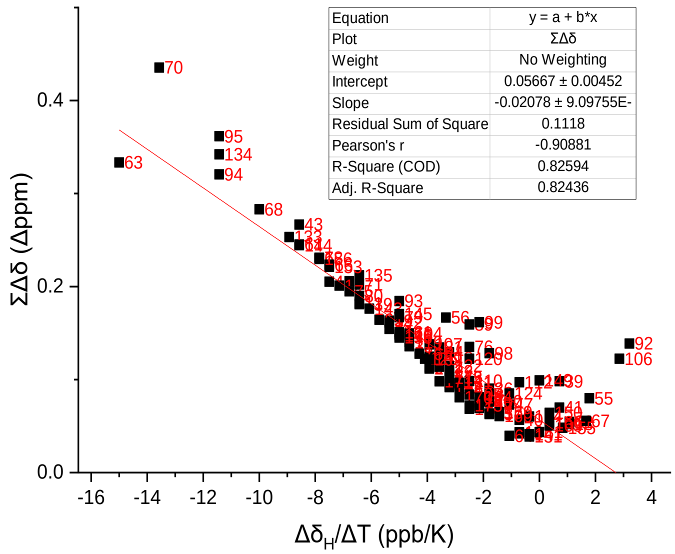


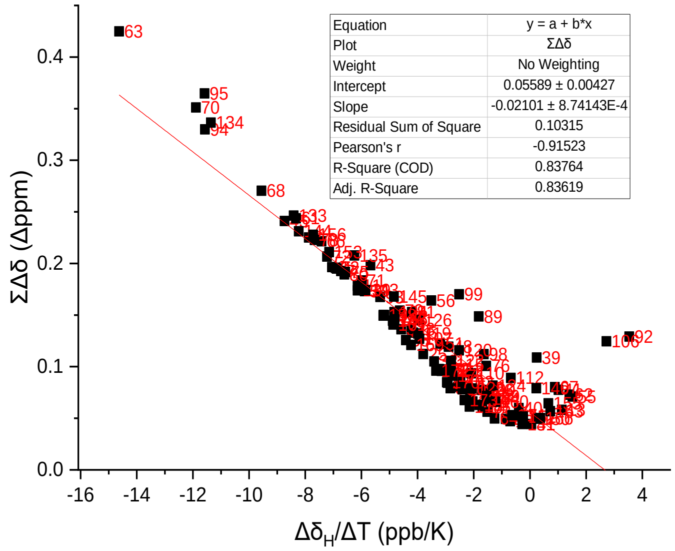
Figure 27. Residue Movement in the ^{1}H and ^{15}N Plane. A-J. Residue condition is indicated above each plot and temperature is labelled red.

A.WT

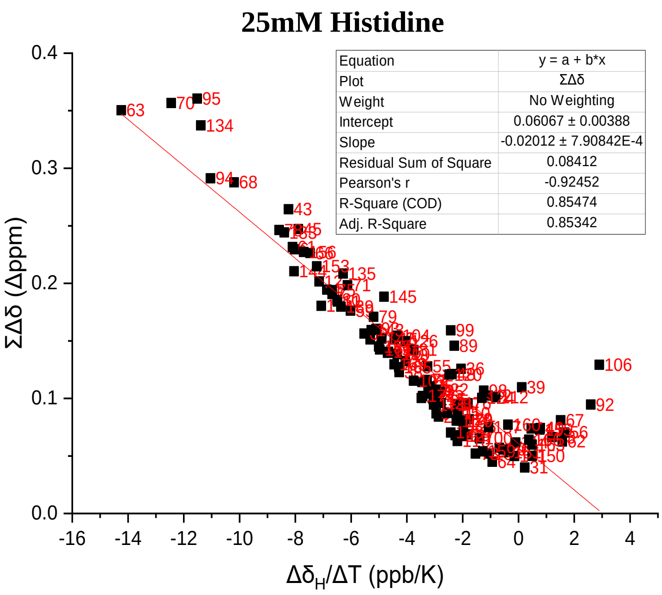


Equation	y = a + b*x
Plot	ΣΔδ
Weight	No Weighting
Intercept	0.05667 ± 0.00452
Slope	-0.02078 ± 9.09755E-
Residual Sum of Square	0.1118
Pearson's r	-0.90881
R-Square (COD)	0.82594
Adj. R-Square	0.82436

12.5mM Phenylalanine

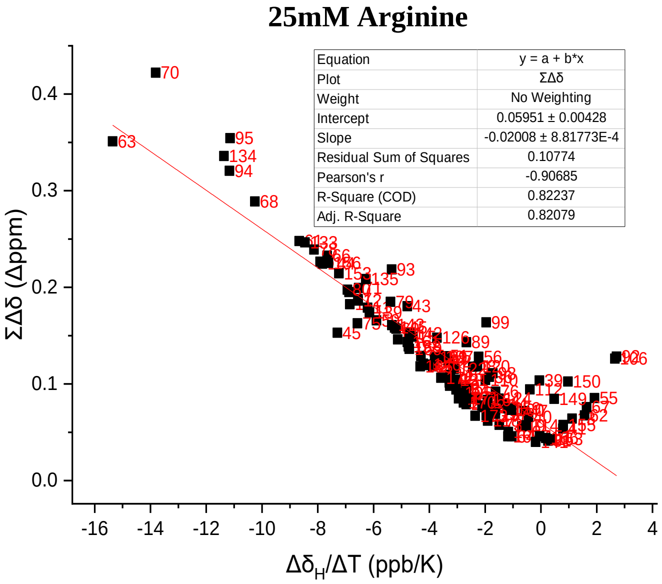


Equation	y = a + b*x
Plot	ΣΔδ
Weight	No Weighting
Intercept	0.05589 ± 0.00427
Slope	-0.02101 ± 8.74143E-4
Residual Sum of Square	0.10315
Pearson's r	-0.91523
R-Square (COD)	0.83764
Adj. R-Square	0.83619



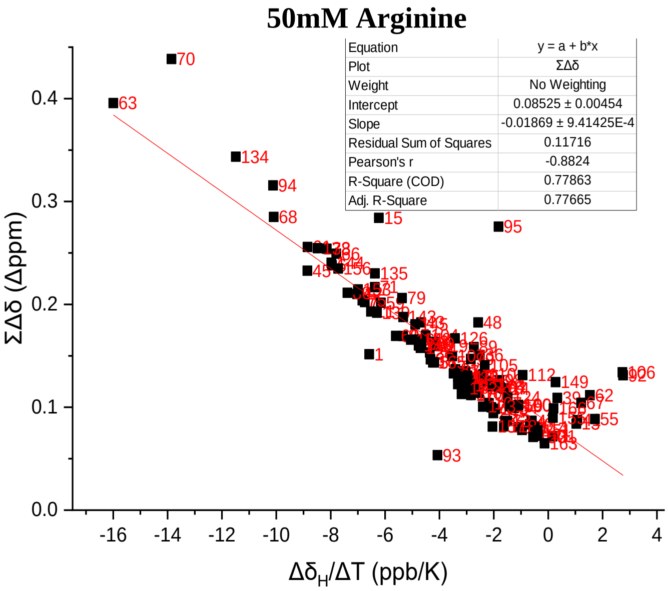
25mM Histidine

Equation	y = a + b*x
Plot	ΣΔδ
Weight	No Weighting
Intercept	0.06067 ± 0.00388
Slope	-0.02012 ± 7.90842E-4
Residual Sum of Square	0.08412
Pearson's r	-0.92452
R-Square (COD)	0.85474
Adj. R-Square	0.85342



25mM Arginine

Equation	y = a + b*x
Plot	ΣΔδ
Weight	No Weighting
Intercept	0.05951 ± 0.00428
Slope	-0.02008 ± 8.81773E-4
Residual Sum of Squares	0.10774
Pearson's r	-0.90685
R-Square (COD)	0.82237
Adj. R-Square	0.82079



50mM Arginine

Equation	y = a + b*x
Plot	ΣΔδ
Weight	No Weighting
Intercept	0.08525 ± 0.00454
Slope	-0.01869 ± 9.41425E-4
Residual Sum of Squares	0.11716
Pearson's r	-0.8824
R-Square (COD)	0.77863
Adj. R-Square	0.77665

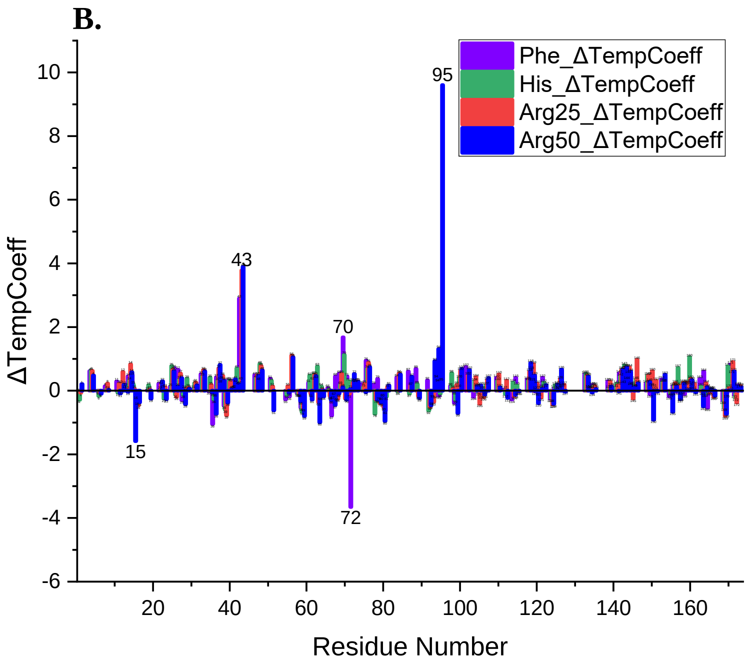


Figure 28. Excipient-induced Changes in $\Delta\delta_H/\Delta T$. A. Residues are labelled red and R^2 is indicated in the box above the plot. **B.** Positive differential indicates stronger hydrogen bonding and negative indicates weaker bonding.

4.1.7 Examining changes in the "switch" mechanism

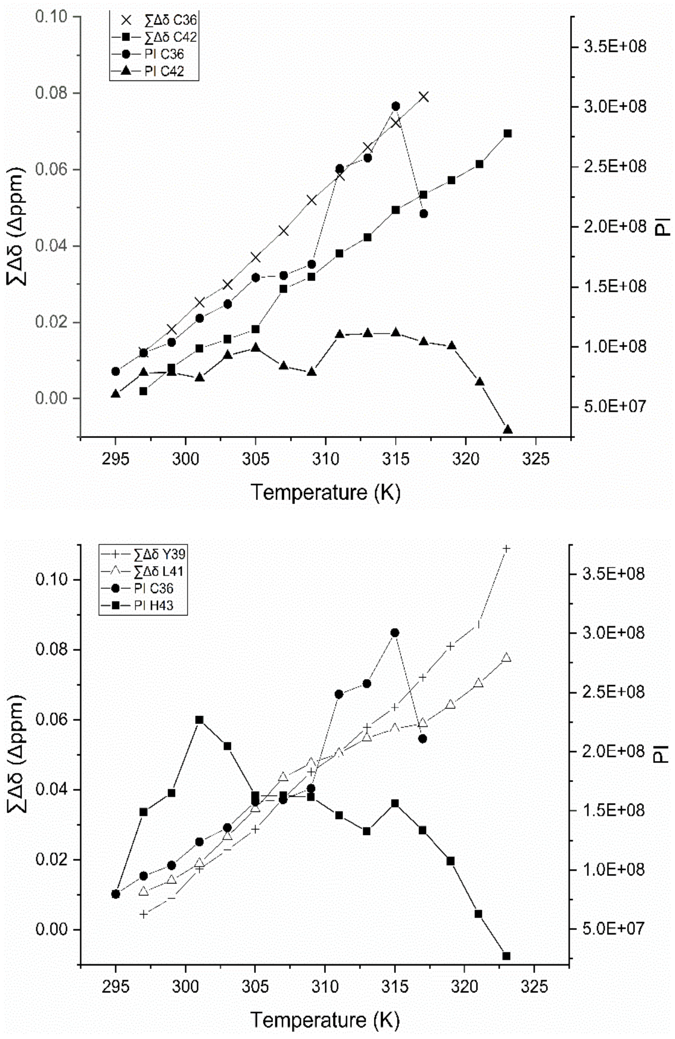
Structural remodelling around the "switch" mechanism region is important to bioactivity. The excipients used in this study impacted protein stability and may influence bioactivity consequently. Therefore, Figure 29 compares ΣΔδ with PI in the same way as Figure 12b and c to probe how excipients may affect residues significant to the "switch" mechanism. Both observables generally displayed the same pattern as the control for residues C36, Y39, L41, C42 and H43. This emphasises the significance of the structure remodelling occurring in this region to the molecule, given that all excipients appeared to alter structure (and stability) without affecting the general pattern of changes in the "switch" mechanism. The main differences reported in the NMR observables in this region was at H43 for the excipient-induced Δδ and Δtempcoeff differential, as well as a large decrease in linearity for residues T38-H43 (Figure 26, 27 and 28).

At 12.5 mM phenylalanine, the greater structural resilience of H43 in addition to potential conformational/rotamer change (Figure 27h and 28) points to potential stabilisation around the "switch" mechanism area. This could be at the cost of residue T38 as it showed increased conformational diffusion Figure 27e). The same pattern here for residues T38 and H43 was seen in 50 mM arginine, hinting at a shared mechanism of action in this region. Furthermore, the change in PI with temperature for C36 in both 12.5 mM phenylalanine and 50 mM arginine was similar (Figure 29a and d). Here, PI showed a sharp increase at ~310 K, as also seen in the other conditions, but then reached a maximum at ~3.0E+08 before sharply decreasing. This differred from the remaining conditions where C36 PI increased and then gradually decreased, resembling more of a curve than a sharp point. The reason for this could be that the increased conformational diffusion of T38 permits C36 to reach higher strain (PI), brought on by the "switch", before loss of sample causes PI to globally decrease. This could explain why C36 reached a higher PI of 2.8-3.0E+08 for 2.5 mM phenylalanine and 50 mM arginine compared to 2.4-2.6E+08 for the other conditions.

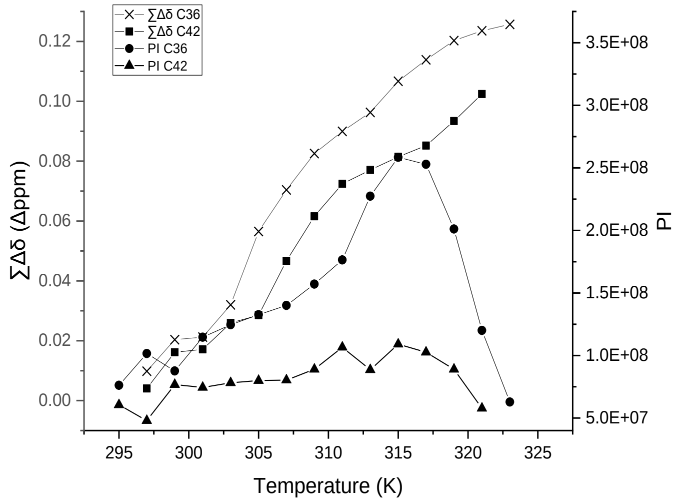
The behaviour of residue H43 was only comparable to the control at 25 mM histidine (Figure 27h and 28). Additionally, ΣΔδ at this condition was also most comparable with the control, and so minimal impact was observed on H43. Nevertheless, 25 mM histidine was the only condition to elicit potential protein-excipient interaction within the "switch" mechanism region at residue E45. A mild decrease in linearity occurred in this region at residues T38 and Y39 (Figure 26) as it did in 25/50 mM arginine. Therefore, whilst 25 mM histidine may not affect the ability of H43 to "switch", it may impact the surrounding environment at T38, Y39 and E45. This impact resembled H43 choosing to "switch" towards the free E46 instead of the excipient-occupied E45. Residue E46 is further away and would therefore put further strain on C36. Nonetheless, strain on C36 may be offset onto T38 and Y39 causing them to undergo larger conformational diffusion (Figure 26). This would explain why C36 experienced a larger overall value of (and steep incline) in ΣΔδ compared to other conditions (except 25 mM arginine) without its PI being much higher (Figure 29b). An impact on bioactivity would be expected if this scenario was the case, given than E46 is part of the binding site-III.

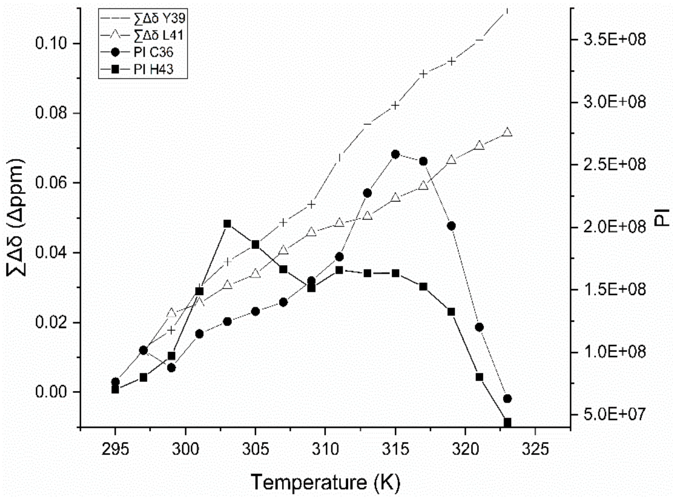
Increasing concentrations of arginine yielded more residues in the "switch" mechanism region that became less linear (Figure 26), namely: T38, Y39, L41 and H43. The microenvironment of C42 was also significantly affected upon addition of 50 mM arginine (Figure 18d and 20d). Therefore, like with 25 mM histidine, increased conformational diffusion for these residues permitted a larger ΣΔδ for C36. The larger cluster of residues (T38-H43) for 25 mM arginine would explain the larger increase in ΣΔδ for C36 compared with 25 mM histidine (Figure 29d). Given that two residues discussed here (Y39 and L41) are in binding site-III, we would also expect bioactivity to be affected at 50 mM arginine.

A. 12.5mM Phenylalanine

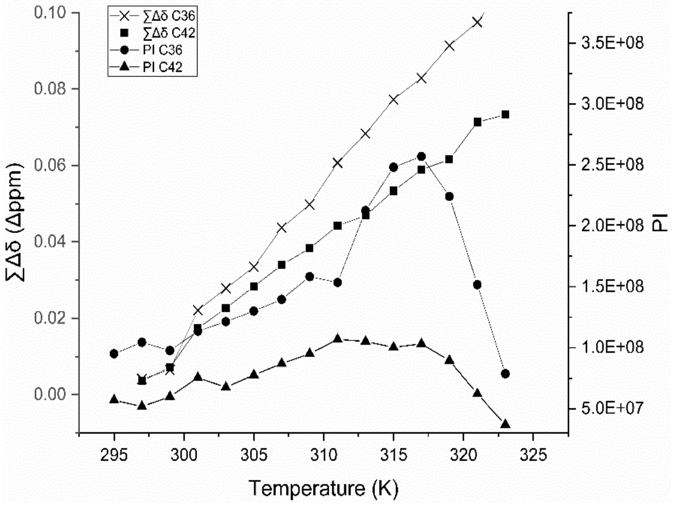


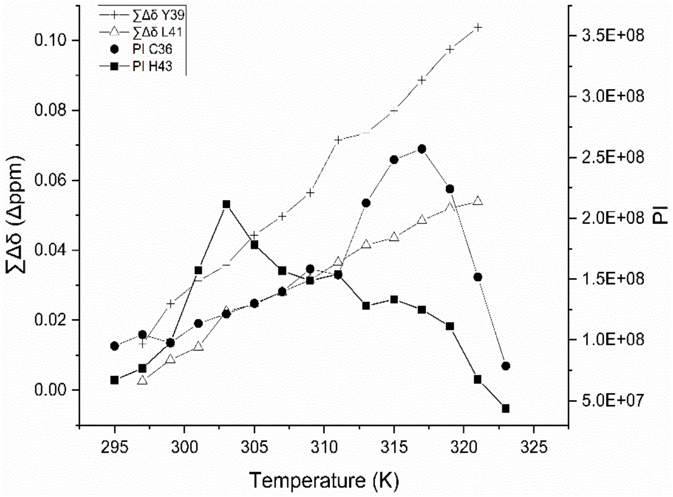
B. 25mM Histidine

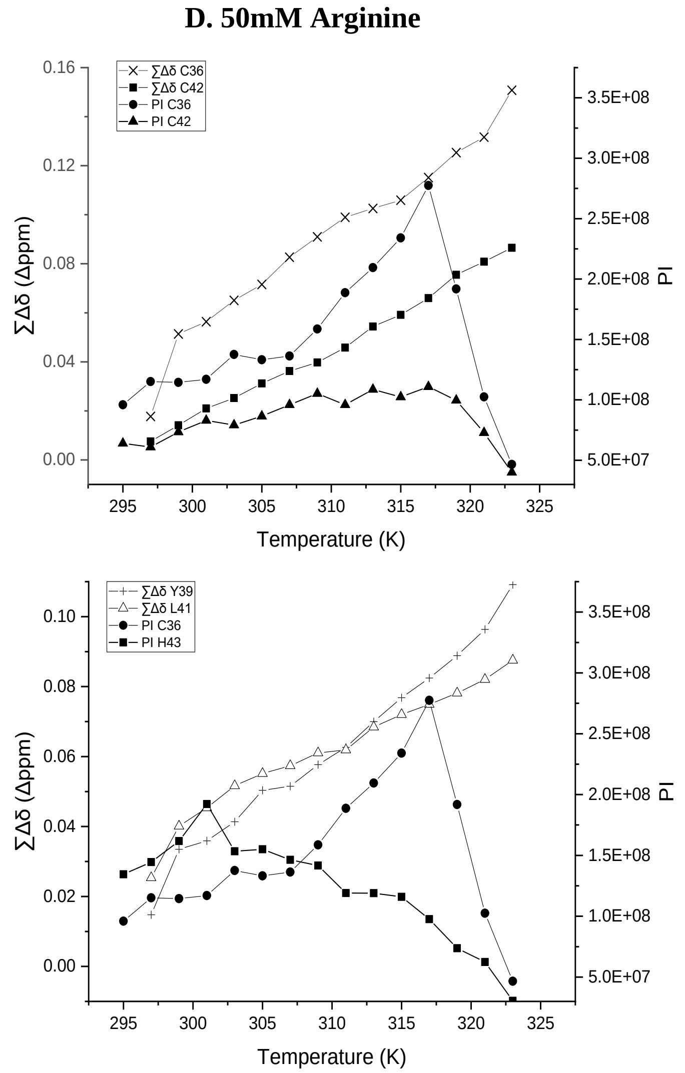




C. 25mM Arginine







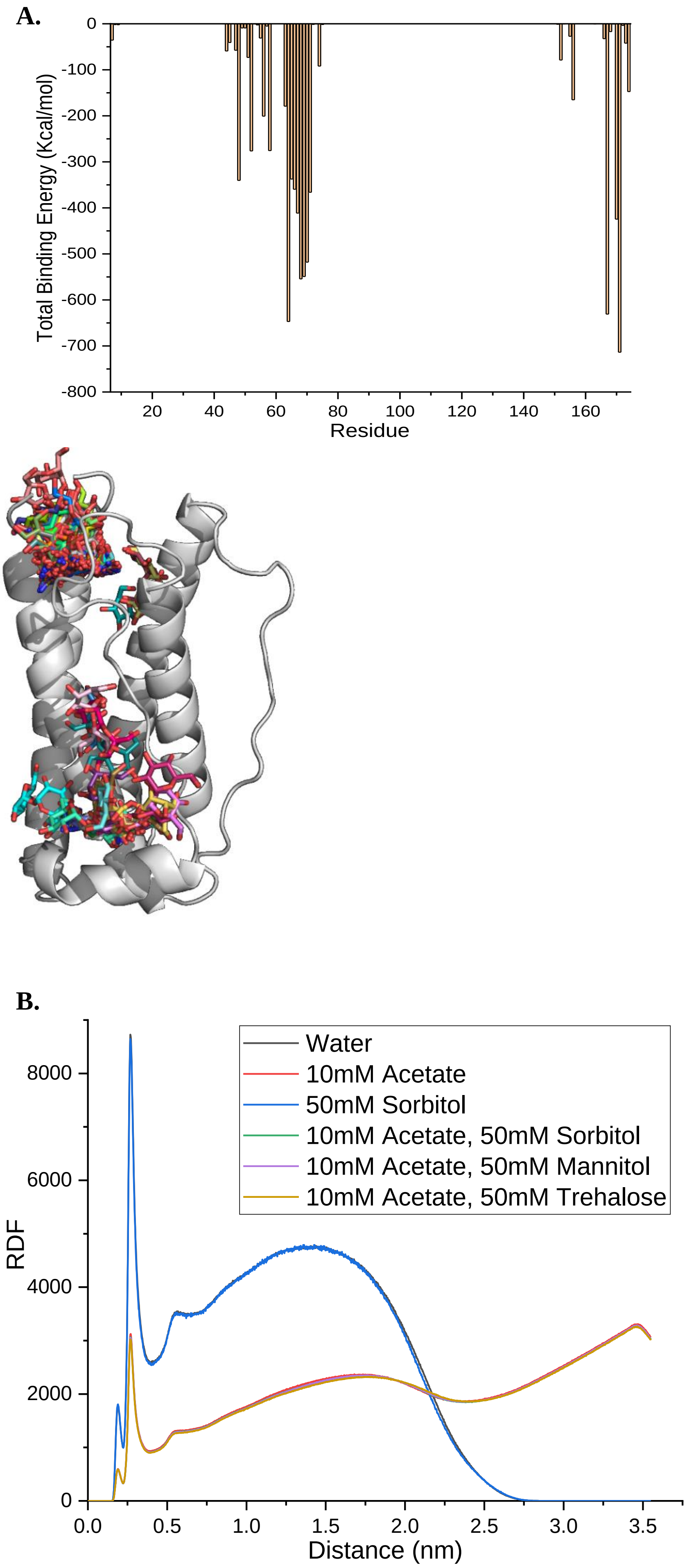
D. 50mM Arginine

Figure 29. The Influence of Excipients on the "Switch Mechanism". A.-D. Figure 12b and c are reproduced for all excipient conditions.

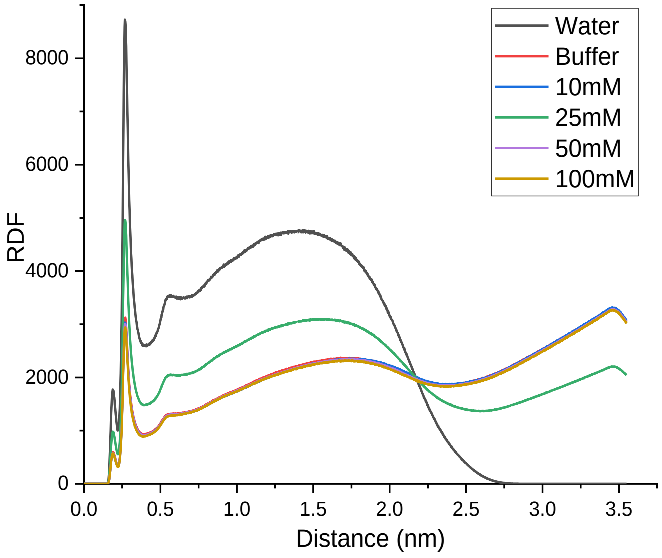
4.1.8 Can In Silico Modelling of Excipients and Hydration Shells Explain NMR Observables?

Although NMR observables suggested changes to G-CSF structure and conformational remodelling occurred upon addition of excipients, concluding that excipient binding was the cause of these changes can be challenging. Computational modelling can complement NMR in revealing energetically favourable protein-excipient interactions. iGEMDOCK was used (Figure 30a) to yield information on the sum of hydrogen bond, VDW and electrostatic free binding energies between unrelaxed G-CSF (accompanying the graph) and the excipients arginine, phenylalanine, mannitol, sorbitol, sucrose and trehalose. The unrelaxed structure was used instead of the relaxed (used in Figure 19b) to investigate whether conformational changes between the two structures influenced changes in excipient interaction sites. Excipients clustered with interactions to the unrelaxed structure along the short helix (N-terminus of loop AB) and C-terminus of loop AB (C-terminus of helix D), and was more clustered than with the relaxed structure. Strong excipient interactions were focused around the C-terminus of loop AB and helix D with the unrelaxed structure, particularly residues C64-Q70 and V167-L171. Thus, excipient docking with the unrelaxed structure reflected the targeted mechanism of action for 12.5 mM phenylalanine (Figure 31a). On the other hand, docking with the relaxed structure highlighted the potential interaction site on the N-terminus of helix B (around H79): reflecting more the mechanism of action for the remaining excipient conditions (Figure 31). The contribution of each excipient to overall free binding energy (not shown) revealed that the sugar excipients had the largest contribution, which did not mirror NMR observations where sugars did not interact with G-CSF (data not shown).

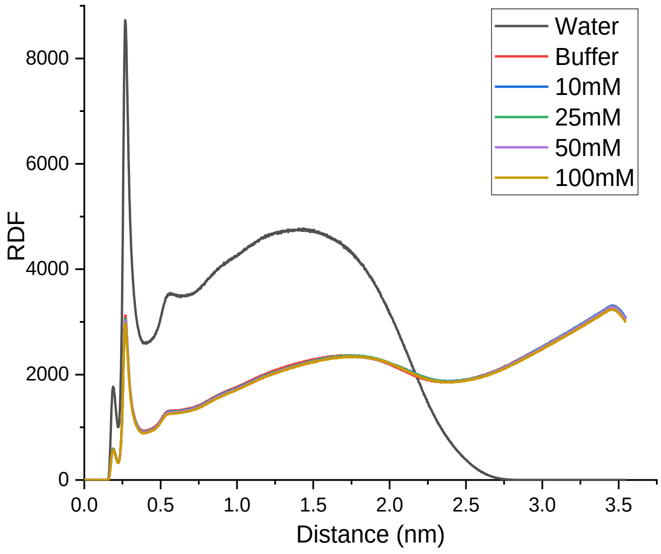
MD simulations were conducted with the same excipients used for docking (Figure 30a) using the CHARMM36 force field (TIP3P water model). Radial distribution function (RDF; Chen, Weber and Harrison, 2008) was calculated for water with all of these excipients and averaged from at least three repeats (Figure 30b). The amino acid excipients phenylalanine and arginine were simulated at various concentrations because of their palpable concentration-dependent effect on thermal stability stability (Wood *et al.*, 2020). The peaks at ~2.75 Å and 5.8 Å corresponding to the hydration shell layers would need to be disturbed for excipient interaction. However, their sharpness and position remained unperturbed indicating that protein-excipient interaction was unlikely at the tested conditions. Nevertheless, the excipient-protein distance distribution for phenylalanine and arginine decreased with increasing concentration (Figure 30c). This suggested that, although average distribution did not fall within the distance of hydrogen bonding or VDW interactions, interaction became more likely as more co-solvent was added to the solution. Of note, sites of potential interaction, as concentration increased, overlapped with NMR observables (Figure 30d, 31a and d). Figure 30d shows the probability of phenylalanine and arginine excipients (at 25 mM and 100 mM respectively) coming within the range of G-CSF residues needed for hydrogen bonding and VDW interaction. Both NMR and MD showed interaction and stabilisation most likely occurred around the C-terminus of loop AB (S62-L69) and helix D (V163-L171) with phenylalanine (Figure 30d and 31a). Moreover, both approaches showed excipient interaction near residues H79/S80 and A111-W118 (the red cluster; Figure 31d) and that arginine is less selective for areas of preferential interaction (Figure 30d). However, MD predicted there to be a higher probability of protein-excipient interaction for phenylalanine than for arginine, which did not reflect the NMR observations (Figure 30c and d). In addition, changes in the most dynamic protein regions from MD (data not shown) do not mirror NMR observations, which show dynamics to decrease with increasing co-solvent concentrations.


A.
Total Binding Energy (Kcal/mol)
0
-100
-200
-300
-400
-500
-600
-700
-800
20
40
60
80
100
120
140
160
Residue
B.
RDF
8000
6000
4000
2000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Distance (nm)
Water
10mM Acetate
50mM Sorbitol
10mM Acetate, 50mM Sorbitol
10mM Acetate, 50mM Mannitol
10mM Acetate, 50mM Trehalose

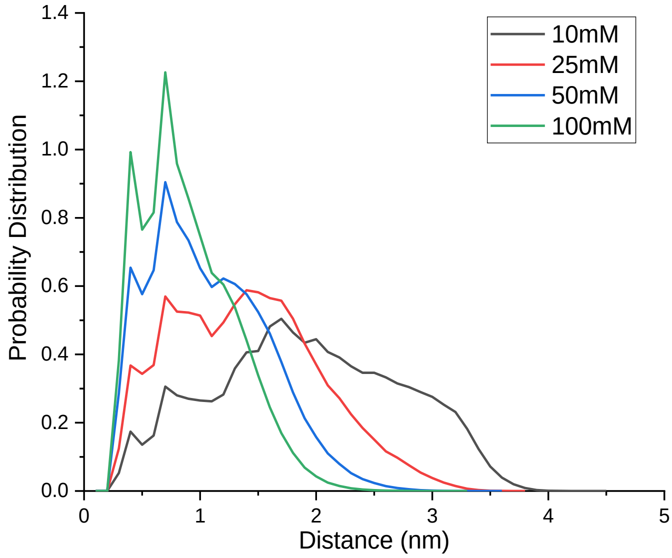
Phenylalanine



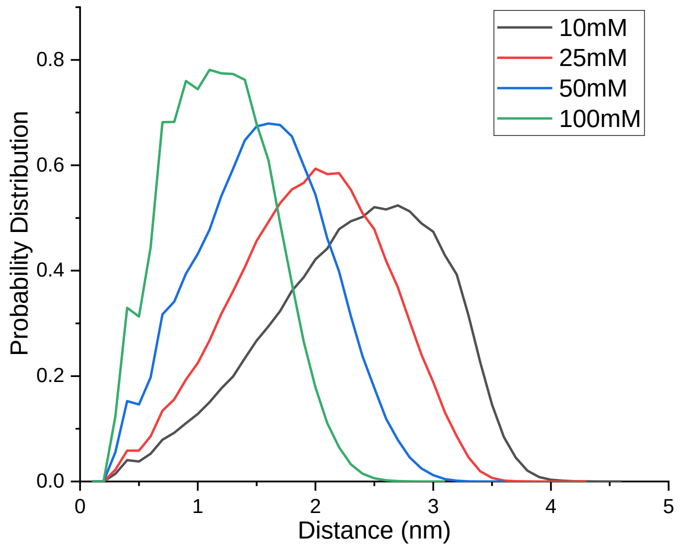
Arginine



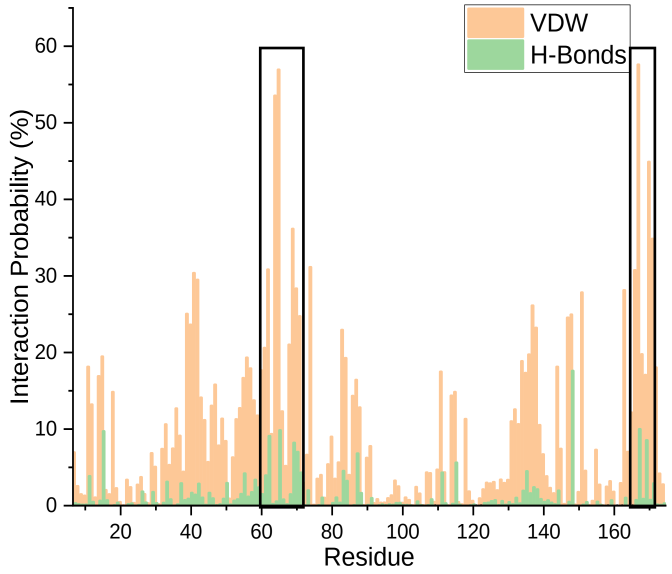
C. Phenylalanine



Arginine



D. 25mM Phenylalanine



100mM Arginine

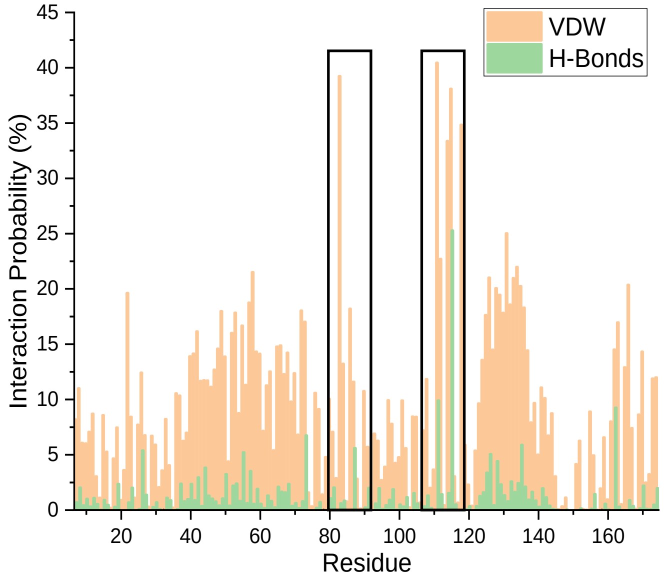


Figure 30. Probing Protein-excipient Interaction with iGEMDOCK and MD. A. Free binding energy, for interactions outlined in the legend, for excipients arginine, phenylalanine, mannitol, sorbitol, sucrose and trehalose calculated in iGEMDOCK. **B.** Water RDF. Co-solvents and concentrations are defined in the legend. **C.** Distribution of distances between varying concentrations of phenylalanine and arginine and G-CSF. **D.** Probability of interaction via hydrogen bonding or VDW forces between G-CSF and 25mM phenylalanine and 100mM arginine.

4.2 Discussion

The four structural clusters identified in Figure 9d vastly remained the same for all excipient conditions (Figure 17 and 21). This emphasises the importance of conformational remodelling in these regions, which are concentrated around loop AB and proximal regions (with the exception of loop CD). Protein function instils selective pressure on structural changes significant to bioactivity, which comports with the observation of consistent remodelling around loop AB, as this region is significant to protein function (chapter 3). Remodelling in this region also consistently translated to the general dynamic and environmental changes needed for the "switch" mechanism to occur (Figure 29).

Also unperturbed by co-solvent (aside from 20 mM arginine) were the concerted structural and dynamic changes occurring in between regions T1-S12 and T133-A139 (Figure 23). This suggests these two regions to be epistatic via dynamic correlation with W118, as described in section 3.1.4. Coevolution analysis in Figure 24 supported this hypothesis, showing strong coevolution between residues in the two regions. Implementing point mutations in either of these regions has proven to detrimental to thermal stability (Wood *et al*., 2022), an expected observation when both regions show minimal thermal resistance (Figure 17). G-CSF seems to have found a way to stabilise this region with post-translation modification, given that T133 is the site of O-glycosylation in hG-CSF (Souza *et al*., 1986). Therefore, stabilising mutations implemented at both sites simultaneously should improve thermal stability, perhaps in an additive manner.

On the other hand, co-solvent had an impact on G-CSF structure to varying extents (Figure 18). These excipient-induced structural changes hinted at interactions for the positively charged excipients histidine and arginine, because many of the affected residues were negatively charged (E45, E93, E98, D112 and E122: Figure 20c). Previous pH studies on G-CSF with ^{15}N-^{1}H NMR suggested that the structural changes induced by excipients in this study were not solely due to a change in pH (Aubin *et al*., 2015). The only assigned residues that overlapped as experiencing a significant environmental change due to pH and excipient conditions were H79, S80, E93 and E98.

A strong correlation existed between $\Delta\delta_H/\Delta T$ and $\Sigma\Delta\delta$, suggesting that hydrogen bond breaking was the largest contributor to loss of structure over the thermal melt for all co-solvent conditions (Figure 28). The high concentration of non-linear, very conserved residues in the C-terminal region of helix D (F160, R166, H170) may serve as a region of structural resilience or even folding nuclei, given their extremely low environmental and dynamic changes across all co-solvents (Figure 25b-f, S.11 and S.12). The strong hydrogen bonding in this region (Figure 28), particularly at R166, appears to have been stabilised, where all conditions (apart from 50 mM arginine) diminished hydrogen bond stretching/breaking and reforming (Figure 26 and 27d). General protein dynamics was also dampened at each excipient condition (Figure 22a), supporting previous observations that addition of any co-solvent lowers the activity of the protein through translational and rotational entropy loss (Timasheff, 2002; Irudayam, and Henchman, 2009). Nevertheless, each excipient seemed to have its own additional mechanism(s) of action. Therefore, areas displaying significant changes in structure or conformational rearrangement at their respective excipient condition are highlighted and clustered in Figure 31.

4.2.1 Phenylalanine

Phenylalanine had the lowest impact on protein structure but elicited the largest influence on structure remodelling and response to thermal perturbation along with 50 mM arginine. The largest excipient-induced structural change observed at 299 K with 12.5 mM phenylalanine was clustered in the C-terminus of loop AB and slightly in loop BC, coloured blue and green respectively in Figure 31a. Loop BC was affected across all excipient conditions. However, pH-sensitive residues E93/98 (Aubin *et al*., 2015) were not affected at 12.5 mM phenylalanine, unlike with the other excipients (discussed below). Therefore, what makes phenylalanine's mechanism of action unique was its concentrated impact on loop AB. Residues I56, Q70 and A72 revealed changes in remodelling in this loop region, where hydrogen bonding was strengthened around Q70 (and weakened around A72: Figure 28) and large changes in torsion angle/side chain rotamer around I56 were diminished (Figure 27c).

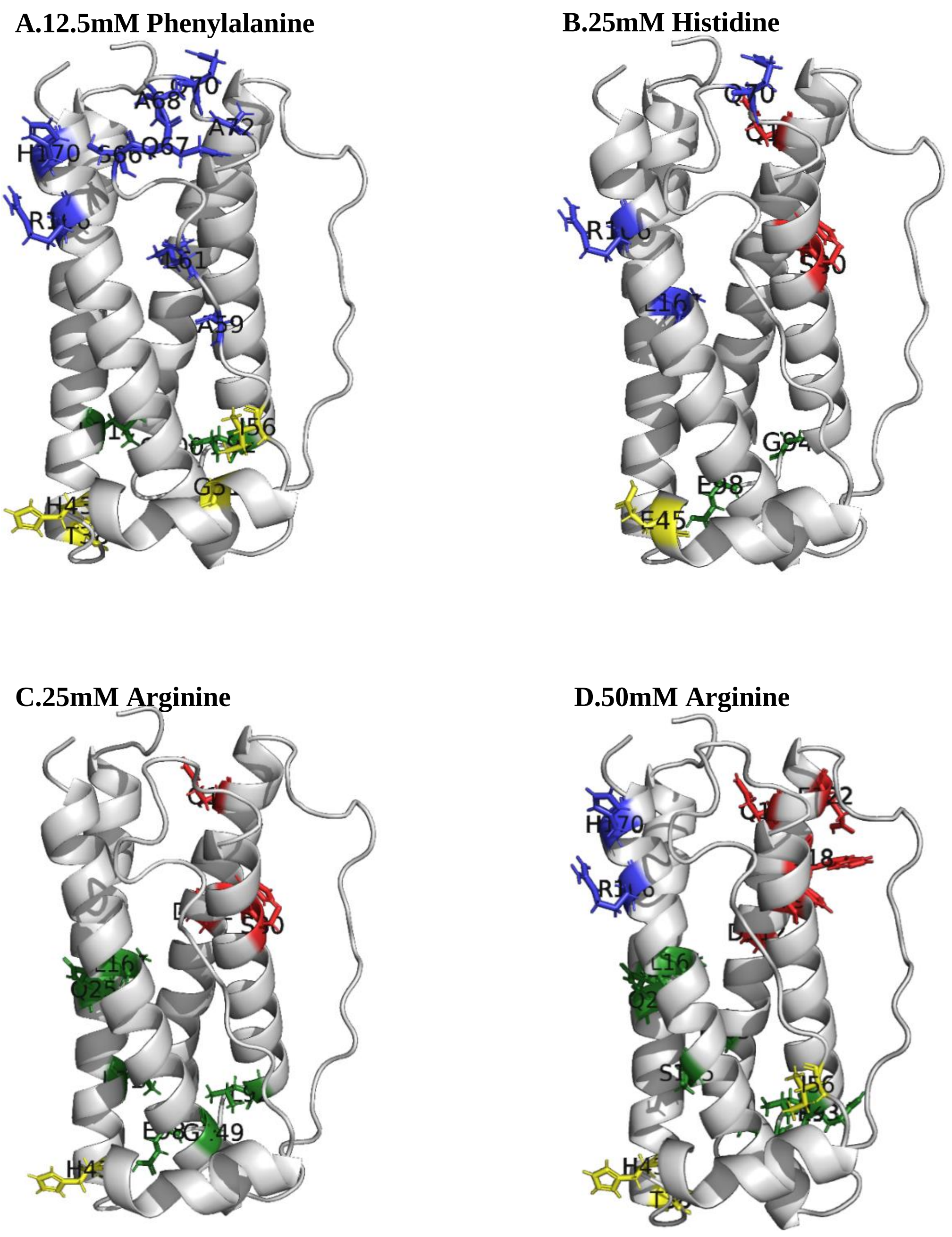
Similar to I56, residues T38 and H43 likely experienced profound differences in torsion angle/side chain rotamer changes compared to the control (Figure 27h) and were thus clustered in yellow (Figure 31a). In addition, 12.5 mM phenylalanine reduced hydrogen bond stretching/breaking at R166 and rotamer switching at H170 (Figure 26, 27a and d): a region proximal to the C-terminus of loop AB. As a result, 12.5 mM phenylalanine stabilised two regions significant to structural stability and thermal resilience, namely the C-terminus of loop AB and helix D. This region in helix D may be stabilised by charge shielding of proximal, positively charged residues R166, R169 and H170. Loop AB could be stabilised by phenylalanine through resilience of this loop in a similar relaxed conformation to that in Figure S.9a. This could take place via an increase in osmolarity upon excipient addition, effectively pushing the hydration shell towards the protein surface, the least structurally resilient of which is the loop AB C-terminus and loop BC. Consequently, Q70 was kept near S8 (explaining the increased hydrogen bonding strength) and Q67 may be kept near H170 for longer, thus resulting in the delayed formation of the aggregation prone conformation. This excipient condition serves to back the previous observation that binding strength does not dictate the ability of a co-solvent to stabilise the protein (Zalar, Svilenov and Golovanov, 2020).

4.2.2 Histidine

In general, 25 mM histidine exhibited the opposite features to 12.5 mM phenylalanine, where significant structural changes were observed (albeit sparse), while structure remodelling changes were minimal (Figure 26, 27h, 28 and 30b). Protein interaction likely occurred at the solvent exposed, charged/polar residues E45, Q70 and E98 (Figure 20c and d), the strongest of which was at E45 (coloured yellow). The importance of this residue to the "switch" mechanism (Figure 12) appears to have had a knock-on effect, causing greater structural changes (and conformational diffusion) in nearby residues (C36, T38 and Y39) which could affect bioactivity (Figure 29b). Moreover, 25 mM histidine may have affected the pH of the solvent, thus explaining the structural impact on pH-sensitive residues (Aubin et al., 2015) H79 and S80 in helix B and E98 in loop BC (Figure 31b). Therefore, histidine did not have a concentrated impact on a specific structural region like phenylalanine, but instead had a targeted effect on individual/small groups of residues. The targeted stabilising effect occurred via interaction at Q70 and reduced hydrogen bond breaking/stretching at R166, both in areas key to structural resilience (Figure 18b, 26 and 30b). Although, mildly destabilising events may occur at E45, H79, S80 and E98, 25 mM histidine reduced global dynamics more than for 12.5 mM phenylalanine (and drastically more at very dynamic residues: Figure 22). Hence, having a moderate impact on protein structure with a large impact on global dynamics, may be pivotal to histidine's mechanism of action.

4.2.3 Arginine

Arginine exhibited chaotropic behaviour in disrupting many hydrophobic residues (Figure 20d, 30c and d). Chaotropic behaviour resulted from disruption of bulk and hydration shell water structure, destabilising the hydrophobic effect and leading to protein-excipient interaction (Salvi, De Los Rios and Vendruscolo, 2005). Interaction may occur at the solvent exposed, negatively charged/polar residues E93, E98, D112 and E122 (Figure 20c and d). Disruption of hydrophobic clusters was more striking with arginine given that green and red clusters grew larger with increasing concentration (Figure 31c and d). The growth of the red cluster also signalled the pH-sensitive structural change in this region to be more pronounced. At the heart of this change was H79 (given the PI of histidine), showing a predicted interaction from iGEMDOCK and the largest excipient-induced change in microenvironment (Figure 20c and d). Packing of loop CD may be hindered due to this structural change, promoting instability. Nevertheless, the most important disrupted hydrophobic region was in loop BC (Figure 19 and 20b), at residues L92-I95. Structural instability in this region came in the form of increased conformational diffusion (Figure 26) and significantly increased hydrogen bond strength at I95 (Figure 27g and 28). These changes culminated in L92 leading to an effective global remodelling in structure (Figure 23a). Moreover, 50 mM arginine overwhelmingly increased the conformational diffusion for many residues, such as; K23, T38, T105, W118, S155 and R166. This excipient condition was the only one to have this effect on the structurally resilient R166. Therefore, arginine promoted instability through disruption of loop BC and CD packing, disrupting hydrophobic pockets, and by destabilising residues key to structural resilience.


A.12.5mM Phenylalanine

B.25mM Histidine

C.25mM Arginine

D.50mM Arginine

Figure 31. Excipient Mechanisms of Action. Clusters important to excipient mechanisms of action are coloured yellow, green, blue and red.

MD and iGEMDOCK only predicted the potential stabilising mechanism of action for phenylalanine (Figure 30a-d, 31a), showing that interaction was unlikely and that the closest protein-excipient distances occurred at the N-terminus of loop AB and helix D. NMR did not support the *in silico* prediction that sugars bound G-CSF stronger than amino acid excipients and that arginine interaction was less likely than phenylalanine interactions. Perhaps forcing excipients on to the structure of G-CSF in iGEMDOCK (not accounting for the hydration shell) means that sugars appear as stronger candidates for interaction, and so this should be considered in modelling. Furthermore, MD could still be correct in its prediction of arginine/phenylalanine interaction likelihood. This is because arginine may just disturb the water structures in the hydration shell, thus explaining why hydrophobic residues are affected significantly more with arginine. However, this does not explain why dynamics wesre reduced more with arginine (Figure 22) as this would be expected with closer excipient-protein distances, thus reducing translational and rotational entropy for water molecules and surface residues.

Chapter 5: Semi-rational Mutant Design using NMR and *In Silico* Modelling

The previous chapter found that the examined excipients affected stability by influencing both common and distinct clusters of residues. These observations along with the structural clusters from chapter 3 convey a pathway to implement a protein engineering strategy that considers multiple factors affecting protein behaviour. This chapter will compare *in silico* modelling with NMR observations from the previous chapters to provide rationale to mutant selection and evaluate the ability of *in silico* modelling to reflect experimental data. The structure-function relationship refers to a trade-off where an increase in stability typically comes at a loss to bioactivity. Therefore, this chapter aims to leverage knowledge of functionally significant regions to construct ideal mutant candidates that improve both function and stability.

Constructing and screening large libraries of mutants is a popular experimental approach to protein engineering because it explores a large mutational space (Cravens *et al*., 2021; Sohrabi, Foster and Tavassoli, 2020), which is particularly important when investigating epistasis (Poelwijk, Socolich and Ranganathan, 2019). Nevertheless, construction and screening can be expensive and time consuming, leading some to look for ways to guide construction of these libraries (Wu *et al*., 2019). This chapter shows that *in silico* modelling supports key conclusions from NMR regarding structural and functional changes. Furthermore, excipient-mutation interactions are observed in regions of overlapping structural clusters. Therefore, the potential for NMR to guide semi-rational mutagenesis is explored here.

5.1 Results

5.1.1 How well does In Silico Modelling Reflect NMR Observables?

The comparability of all-atom MD simulation with experimental observations is first inspected with simulations at pH 4.4 and 7.4. These simulations were carried out in Gromacs v.19 in triplicate at 295 K (22 ºC) for both pH conditions with the CHARMM27 force field. Root-mean-square fluctuation (RMSF) is calculated in Figure 32a/b and measures the standard deviation of Cα atom positions throughout the simulation (Sinha and Wang, 2020). Thus, RMSF measures the flexibility of residues. This value is averaged for the triplicates at their respective pH (Figure 32b) and subtracted (pH 7.4 – pH 4.4) to give ΔRMSF (Figure 32a). A negative differential in ΔRMSF, signifying greater flexibility at pH 4.4, occurs for the vast majority of residues, reflecting previous HDX analysis at these pH conditions (Wood *et al*., 2020). This may present as counterintuitive because Filgrastim products are formulated at pH 4 (Lipiäinen *et al*., 2015) and higher flexibility is typically associated with instability. However, the high electrostatic repulsion of G-CSF at low pH combined with the proximity of its histidines to loop regions, discussed in section 1.5.1.3 and Figure 14, would provide the molecule with enhanced colloidal and conformational stability. Supporting the stipulation where histidines make the structure of G-CSF more compact, are observations with radius of gyration (Rg), a measure of G-CSF compactness during simulation (Figure 32c). Here, Rg mainly overlaps for both pHs but is higher at ~20 to 35 ns for pH 7.4.

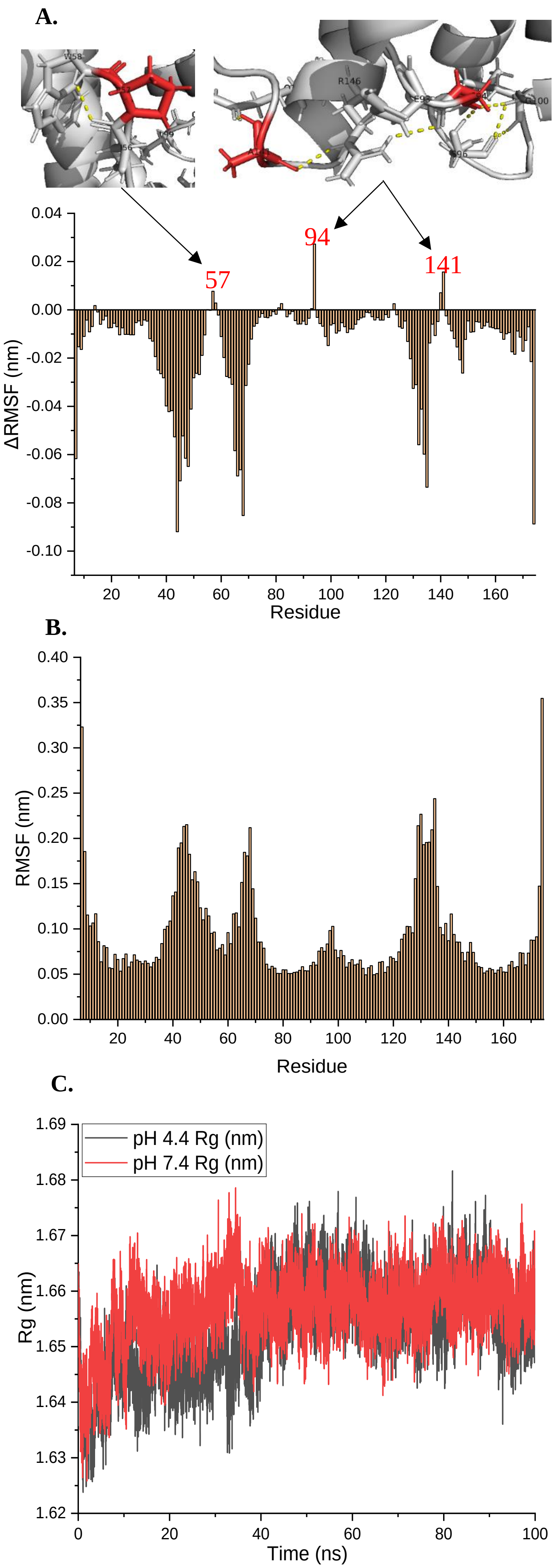
Residues P57, G94 and A141 deviate from the pattern of increased flexibility at pH 4.4 (Figure 32a). P57 mediates backbone hydrogen bonding between W58 and I56, while the backbone carbonyl and amide hydrogen of A141 hydrogen bond Q145 and R146 side chains. The backbone carbonyl of G94 can either hydrogen bond G100 or S96. HDX analysis also shows higher flexibility at pH 7.4 closer to the C-terminal region of loops AB (near P57) and CD (near A141) but not in loop BC where G94 resides. This could be because higher flexibility at pH 4.4 is compensated for by P57, G94 and A141 forming preferential/stronger hydrogen bonding.

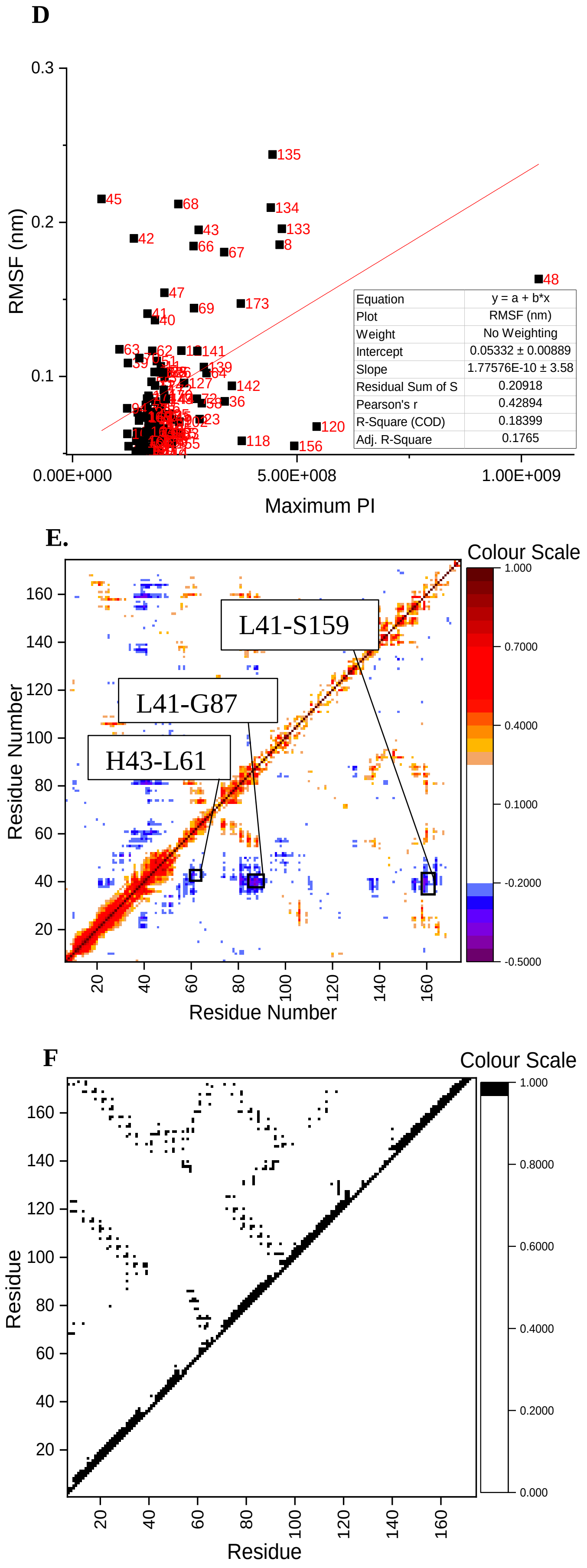
Further MD analysis in Figure 32b-i to 35 was carried out at pH 4.4 (aside from Figure 34) given the comparability of general dynamics at different pH conditions with HDX. RMSF moderately correlates with maximum PI for WT G-CSF (Figure 32d), displaying a Pearson's r-value of 0.429. Most residues in the 90th percentile of maximum PI (such as S8, V48, Q67, T133-G135 and Q173) also have high RMSF values. Yet residues like C36, W118, Q120, S141 and H156 are in the 90th percentile for maximum PI but have low RMSF values, while residues C42 and E45 have a high RMSF but a low maximum PI. Reasoning for the former could be that these residues are in structured regions and MD is measuring dynamics on a nanosecond timescale. Therefore, MD shows all of each loop region to be distinctly dynamic from the α-helices (Figure 32b). MD analysis here would not pick up longer timescale dynamics that NMR reflects. Reasoning for the latter could be that in nature G-CSF is more selective with which residues can be dynamic so that global stability is not compromised by whole regions being very dynamic. Residues that are permitted to be very dynamic must pose functional importance: such as V48, which is in the middle of binding-site III. An exception to this is the whole region of T133-G135, which is dynamic in both MD and NMR, because this region does not show any obvious functional importance. However, the selection of O-glycosylation in this region by hG-CSF could explain this (Souza *et al*., 1986).

The relationship between different residue fluctuations in simulation is examined in the dynamic cross-correlation map (DCCM) in Figure 32e in a similar manner to Figure 11a and 23a. The extent to which all Cα atom fluctuations are positively- or negatively-correlated with another atom in simulation is given by pairwise cross-correlation coefficient (C_{ij}) values, calculated in the R studio package: Bio3D (Grant, Skjærven and Yao, 2021). C_{ij} values above 0.2 indicate positive correlation and below -0.2 indicate anti-correlation. This threshold was also used for DCCM analysis on Transketolase (Yu and Dalby, 2018b). Figure 32f is a map, also produced in Bio3D, showing contacting residues in PDB:2D9Q. Contacting residues are defined as such (and coloured black) when any non-hydrogen atom is within 4.5 Å, an approach used by Sethi *et al*. Areas of positive correlation in the DCCM, which represents the majority of Cij values, overlaps with contacting residues. This emphasises the point that MD conducted on a nanosecond timescale is not sufficient to pick up the correlating conformational changes highlighted with NMR (Figure 11a and 23a).

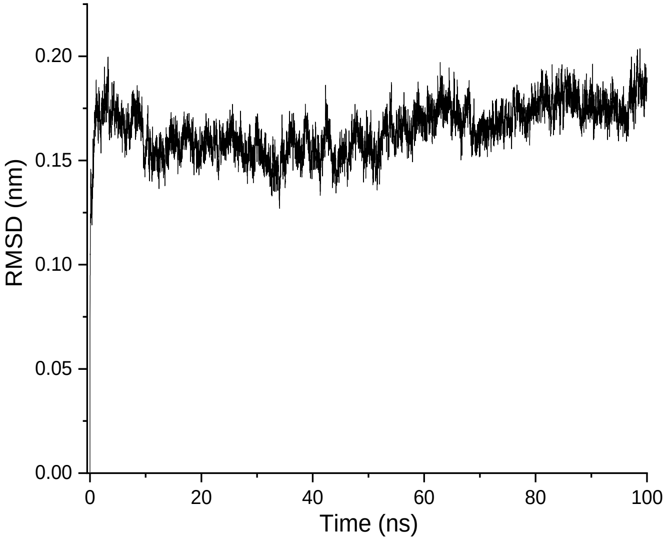
Areas of negative correlation deviate from contacting residues and vastly occurs in the region of residues ~A30 to E45 (Figure 32e). Strong negative correlation occurs between residues L41-G87, L41-S159 and H43-L61, which are in and proximal to regions showing large structural change in simulation (Figure 33c and d). Therefore, DCCM is picking up the conformational rearrangement in loop AB confirmed by NMR (Figure 9d), suggesting that it is a coordinated movement.

Root-mean-square deviation (RMSD) measures how much atoms deviate from their starting position as the simulation progresses. This measure is compared at 295 K (22^{o}C) and 393 K (100^{o}C) in Figure 32g and h respectively to assess thermal-induced unfolding events that can be compared to NMR. Although RMSD is averaged from a triplicate of simulations at 295 K, just one simulation is used at 393 K as this temperature was just used to test for unfolding. At 295K, there is a steep increase in RMSD early in the simulation before it equilibrates at ~10 ns, after which it becomes the "productive region" of the simulation. The simulation at 393 K takes a longer time to reach equilibrium (at 90 ns) and when it does equilibrate, the RMSD is ~15 nm larger, as expected when the simulation has more thermal energy. Rg deviates from its steady baseline, at ~1.61 nm, between 90 ns and 110 ns (Figure 32i), which is also where RMSD begins to equilibrate. This could suggest a conformational rearrangement that is needed at a higher temperature to bring the molecule to equilibrium. Nevertheless, the steady Rg at 393 K indicates that unfolding did not occur under these conditions. A single course-grained MD (CG-MD) run was attempted with the SIRAH 2.0 force field (Machado *et al*., 2019) at 473 K (200^{o}C) for 900 ns to induce unfolding (Figure S.13). Here, RMSD took even longer to reach its higher equilibrium point (at 200 ns) and Rg appeared to oscillate between 1.6 nm and 1.85 nm. Therefore, the CG-MD presented here seemingly picked up much larger conformational perturbation than the aforementioned all-atom MD runs.

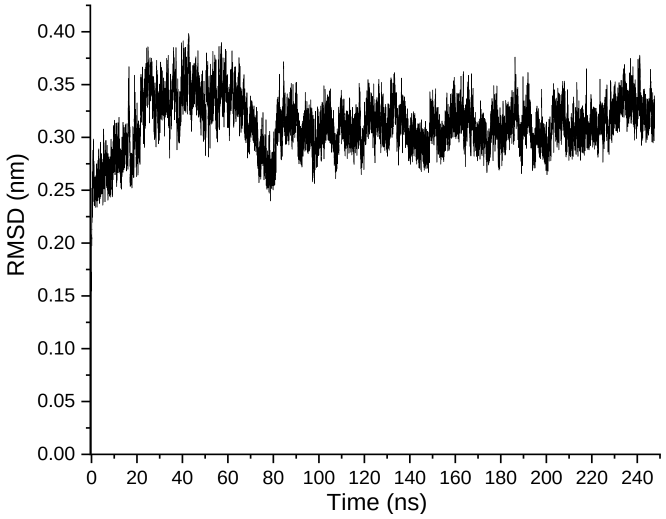

A.
W58
R146
E93
D100
57
94
141
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
-0.10
ΔRMSF (nm)
20
40
60
80
100
120
140
160
Residue
B.
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
RMSF (nm)
20
40
60
80
100
120
140
160
Residue
C.
pH 4.4 Rg (nm)
pH 7.4 Rg (nm)
1.69
1.68
1.67
1.66
1.65
1.64
1.63
1.62
Rg (nm)
0
20
40
60
80
100
Time (ns)


D
0.3
0.2
0.1
RMSF (nm)
0.00E+000
5.00E+008
1.00E+009
Maximum PI
135
45
68
134
43
133
42
66
8
67
48
47
69
173
41
40
63
62
141
139
64
127
142
36
23
120
118
156
Equation | y = a + b*x
Plot | RMSF (nm)
Weight | No Weighting
Intercept | 0.05332 ± 0.00889
Slope | 1.77576E-10 ± 3.58
Residual Sum of S | 0.20918
Pearson's r | 0.42894
R-Square (COD) | 0.18399
Adj. R-Square | 0.1765
E.
Colour Scale
1.000
0.7000
0.4000
0.1000
-0.2000
-0.5000
L41-S159
L41-G87
H43-L61
Residue Number
20
40
60
80
100
120
140
160
Residue Number
F
Colour Scale
1.000
0.8000
0.6000
0.4000
0.2000
0.000
Residue
Residue

G


RMSD (nm)
0.00
0.05
0.10
0.15
0.20
0
20
40
60
80
100
Time (ns)

H


RMSD (nm)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
20
40
60
80
100
120
140
160
180
200
220
240
Time (ns)

I.

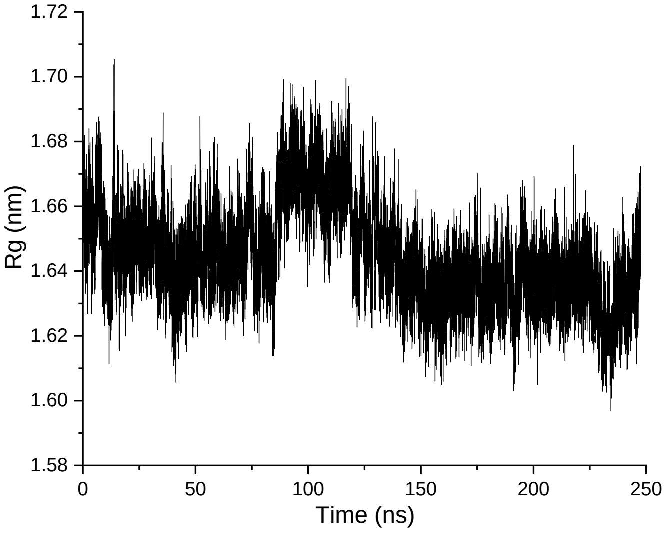
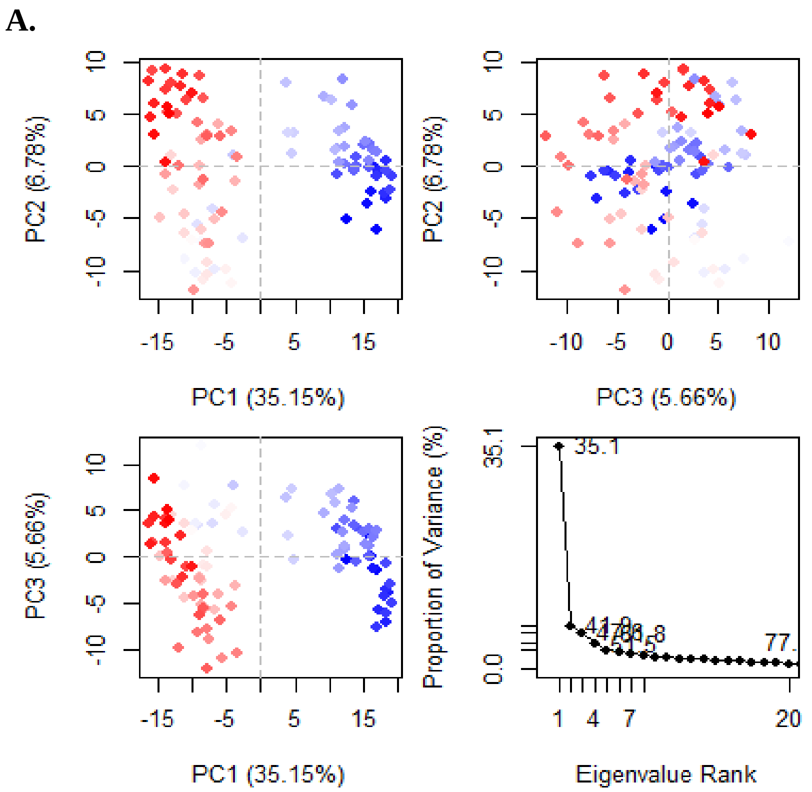
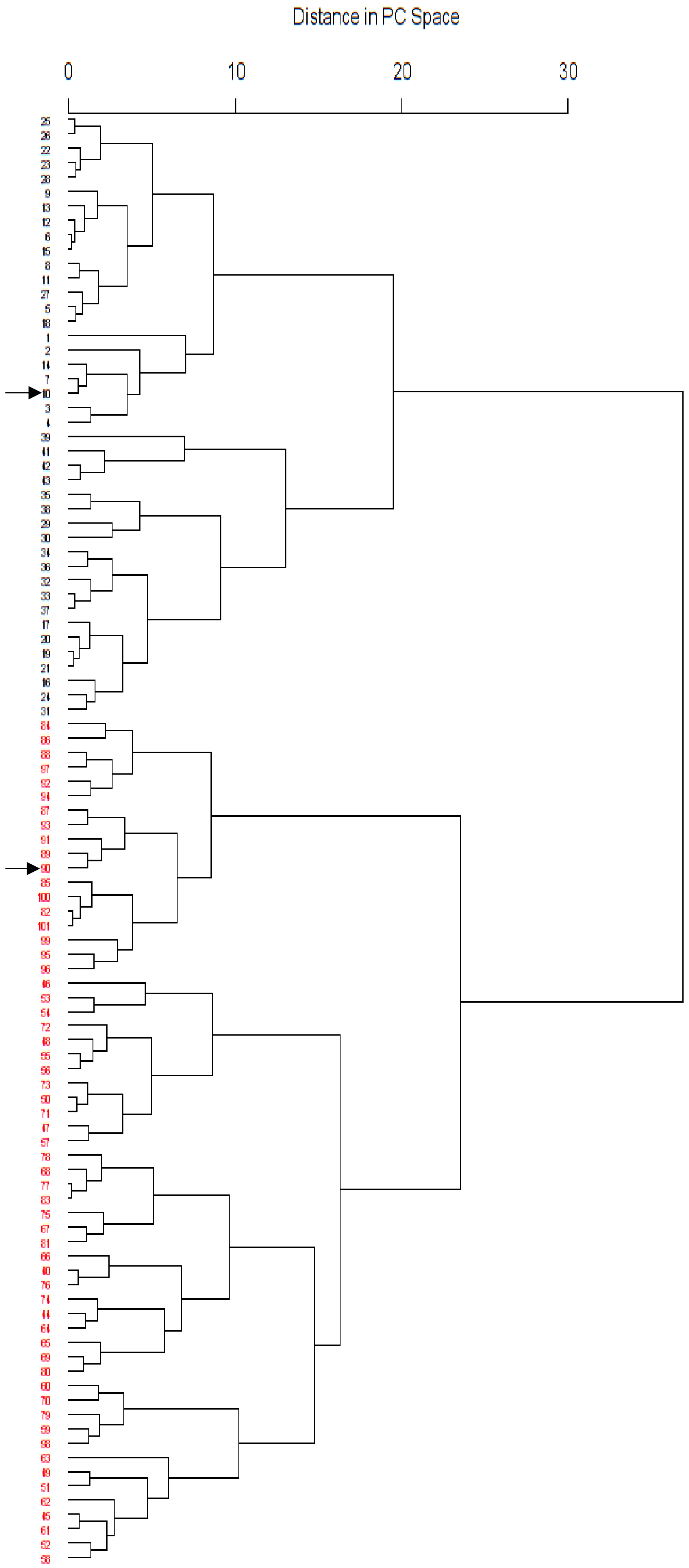

Rg (nm)
1.58
1.60
1.62
1.64
1.66
1.68
1.70
1.72
0
50
100
150
200
250
Time (ns)

Figure 32. Referencing All-atom MD with Experimental Data. All MD runs were at 295 K aside from H and I. **A.** ΔRMSF is calculated with averaged RMSF at pH 7.4 – averaged RMSF at pH 4.4. Residues showing a positive differential are labelled red and illustrated above (highlighted red and shown as sticks), where yellow dashed lines show polar interactions. **B.** Averaged RMSF at pH 4.4. **C.** Averaged Rg at respective pHs. **D.** Average RMSF at pH 4.4 vs maximum PI from WT condition (NMR). Residues are labelled red. **E.** DCCM averaged from MD runs at pH 4.4. positive correlation is indicated on the red scale (>0.2) and negative correlation is on the purple scale (<-0.2) with white in between to reduce noise. **F.** Contact map produced using PDB:2D9Q. Contacting residues are coloured black. **G.** Averaged RMSD at 295 K. **H.** RMSD at 373 K from single MD run. **I.** Rg at 373 K from single MD run.

Although the all-atom MD simulations picked up the general dynamic trends in loop regions observed with NMR, they did not pick up any unfolding events. This could hinder the ability of MD to reflect the microenvironmental changes that residues experienced with NMR. To assess this comparability, the conformational movements that accounted for the majority of motional variation in an MD run were examined with principal component analysis (PCA) in Figure 33. Figure 33a shows three score plots, for the top three principal components (PCs), and a scree plot. The score plots show that the first PC (PC1) clearly separates the data (frames from the simulation) into two clusters, unlike the other PCs. PC coverage of the variance in data, shown in the scree plot, can vary between studies (Grant *et al.*, 2006; Papaleo *et al.,* 2009). In this study, the majority of variance is not covered until PC5, however, the two distinct clusters in PC1 appears to be enough to reflect significant conformational changes. The dendrogram in Figure 33b depicts these two clusters as black and red with the two main branches (at the largest distance in PC space). The closer the branches get, and the shorter the distance between data points in PC space, the more similar the structural conformation. Therefore, two frames (indicated with arrows) were chosen at random and are compared in Figure 33c, where the deep-salmon and grey structures are from the red and black clusters respectively. Large motions are evident at both termini, loop AB, the N-terminus of helix B and loop CD. The difference between Cα positions in these frames is calculated as fluctuation in Figure 33d, and confirms these differences between frames. There is also a lot of overlap between ΣΔδ at 323K (Figure 33e) and fluctuation, particularly in loop AB and CD. The large ΣΔδ at H43 is not quite reflected by its fluctuation in simulation (with the neighbouring P44 displaying a large fluctuation instead), however this is explored in more detail blow (section 5.1.2). While the region of L92-L99 is predicted to experience a mild conformational change (Figure 33d), NMR shows this region experiences a much larger environmental change. On the other hand, NMR and MD are in agreement that this region is only moderately dynamic (Figure 32d), suggesting that it is predicted to possess thermal resilience but in reality is very susceptible. The reason for MD not predicting this unstructured loop (residues L92-L99) to be dynamic or undergo large structural change could be the many hydrophobic residues that it harbours. This supports the destabilising mechanism proposed for 50mM arginine (section 4.1.8.3).



B.


Distance in PC Space
0
10
20
30
25
26
22
23
28
9
13
12
6
15
8
11
27
5
18
1
2
14
7
10
3
4
39
41
42
43
35
38
29
30
34
36
32
33
37
17
20
19
21
16
24
31
84
86
88
97
92
94
87
93
91
89
90
85
100
82
101
99
95
96
46
53
54
72
48
55
56
73
50
71
47
57
78
68
77
83
75
67
81
66
40
76
74
44
64
65
69
80
60
70
79
59
98
63
49
51
62
45
61
52
58

PC1-2

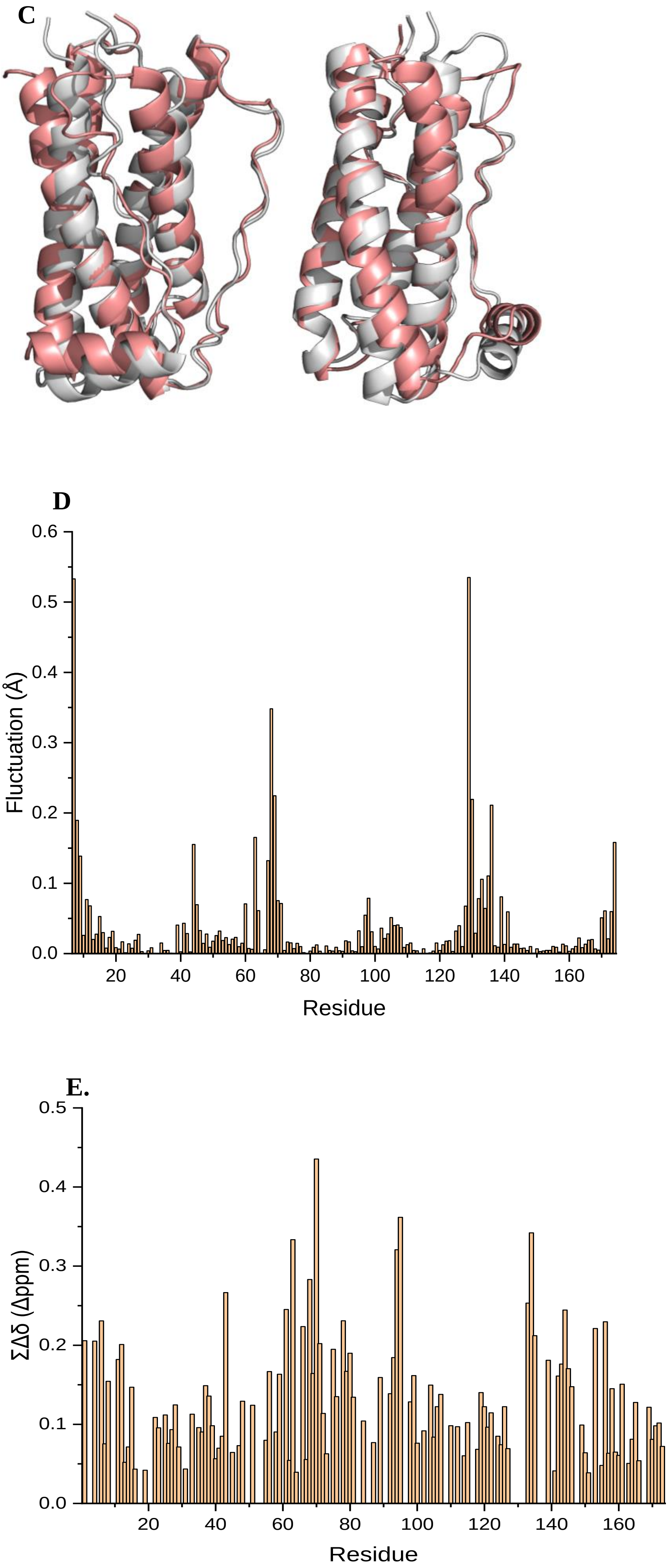
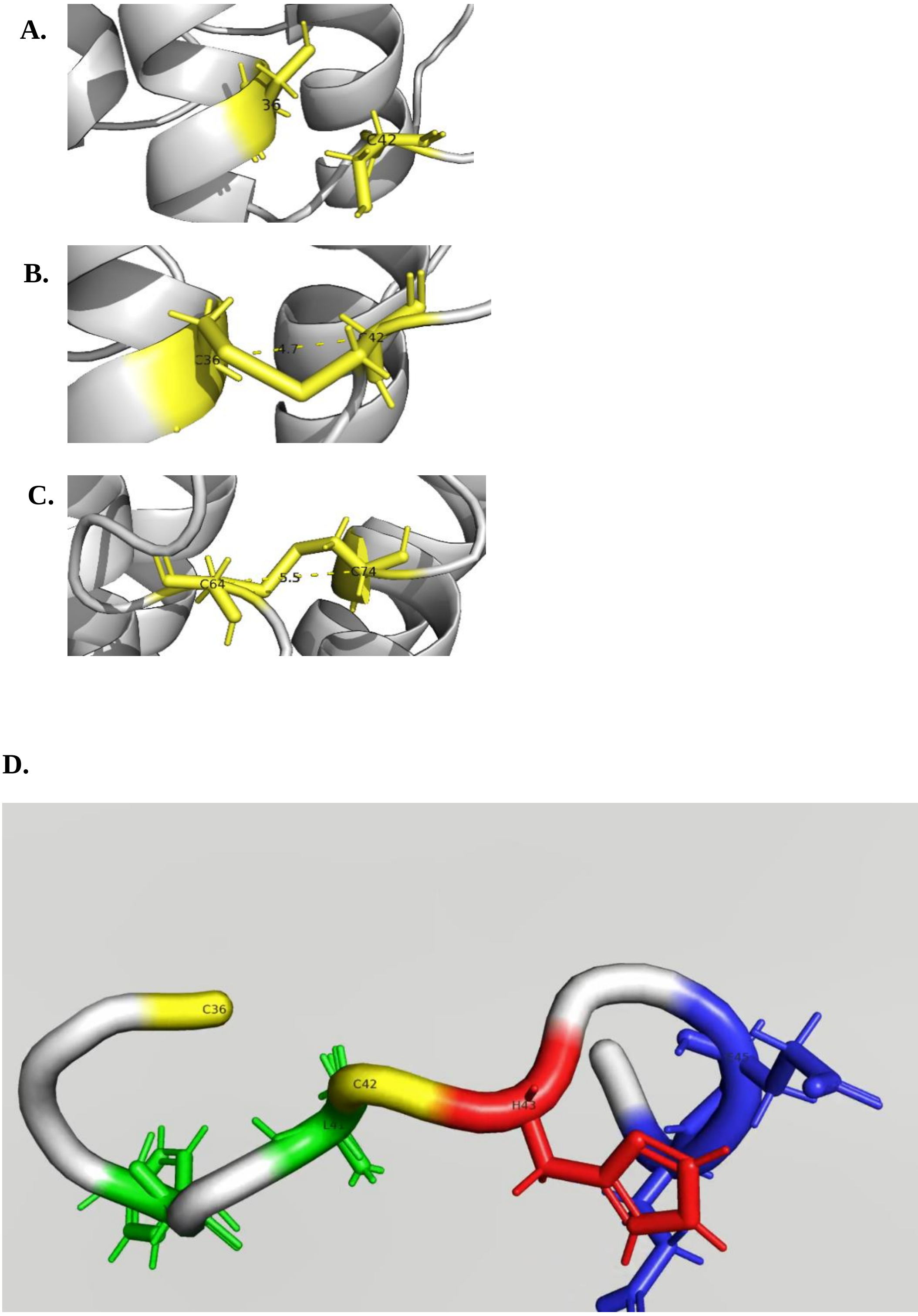

C
D
Fluctuation (Å)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
20
40
60
80
100
120
140
160
Residue
E.
ΣΔδ (Δppm)
0.0
0.1
0.2
0.3
0.4
0.5
20
40
60
80
100
120
140
160
Residue

Figure 33. PCA vs. NMR Observables. A. Scree and score plots for PCA on a single MD run (pH 4.4, 295K). **B.** Dendrogram from PC 1 and 2. **C.** Superposition of simulation frame 90 in red and 10 in grey, front and side view. **D.** Fluctuation between frame 10 and 90. **E.** ΣΔδ at 323K for WT condition.

5.1.2 C36-C42 Disulphide Bond Adopts an Allosteric Conformation Important to "Switch" Mechanism

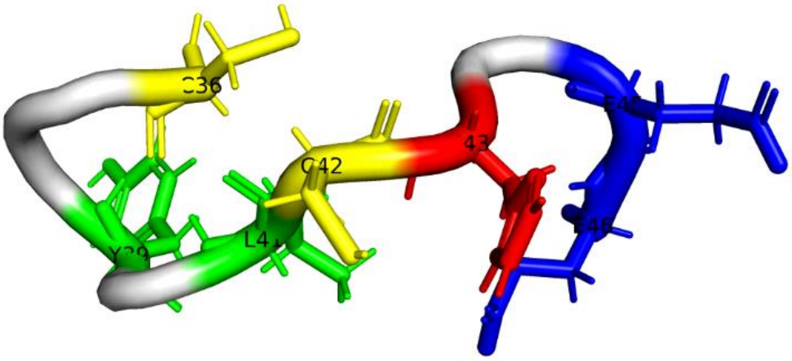
Disulphide bonds are important to protein stability and function. These functional roles are adopted by a minor population of disulphide bonds and are either catalytic, mediating thiol-disulphide exchange (Holmgren, 1989), or allosteric, where their breaking/forming induces functional changes (Schmidt, Ho and Hogg, 2006). Although, G-CSF contains two disulphide bridges, at residues C36-C42 and C64-C74 (Arvedson and Giffin, 2012), the energy minimisation step before the MD run removes the disulphide bridge at C36-C42 (Figure 34a). The purpose of this energy minimisation step is to ensure that there is no inappropriate geometry or steric clashes in the system before the simulation is run. This predicted inappropriate geometry around the C36-C42 disulphide bridge is likely because this bridge adopts a less common conformation. Although cysteine χ_3 and χ_3 angles can be tricky to determine, the φ, ψ and χ_1 angles for C36-C42 (calculated in Bio3D in Table 9) are normal (Armstrong, Kaas, and Rosengren, 2018). Disulphides fall into three basic categories (spirals, hooks and staples) and are either right- or left-handed depending on the χ_3 angle (Richardson, 1989). These definitions are extended to reflect the χ_1 angle (Schmidt, Ho and Hogg, 2006), revealing that the C36-C42 disulphide bridge resembles a –LHStaple allosteric configuration (Schmidt and Hogg, 2007). This is due to the short Ca-Ca' distance of 4.7 Å in the PDB:2D9Q for this bridge (Figure 34b) in contrast to 5.5 Å for C64-C74 (Figure 34c). These –LHStaple allosteric types typically have a high average strain energy of 36.1 kJ/mol, in contrast to 19.2 kJ/mol for a typical -LHSpiral, thus disposing them to cleavage, which would benefit the "switch" mechanism. The strain on C36 resulting from the switch (Figure 12) compounded with the innate high strain of the –LHStaple and reducing environment of bone marrow (Spencer *et al*., 2014) would likely cause breaking of the C36-C42 disulphide bridge. This would improve receptor binding by allowing further expansion of binding site-III.

MD simulations at pH 4.4 and 7.4 pick up the "switch" mechanism and show the strain that C36 would be subjected to had the disulphide bridge formed in the simulation (Figure 34d and e). The switch happens at both MD conditions where H43 attracts towards E45/46 before making a ring flip towards this region where it remains in this conformation for the remainder of the simulation. An emphasis was earlier placed on the significance of pH in the attraction of H43 to E45/46. However, given this attraction occurred at both pH conditions in MD, the significance of P44 in angling H43 toward this region may have been understated. This reflects the significance of the "switch" mechanism to function, because the success of the mechanism does not rely on one characteristic. Moreover, both simulations reinforce the observation from NMR that the movement of H43 towards E45/46 pulls on C42, which would strain the disulphide bridge. Therefore, MD alone reveals key predictive markers of structure remodelling that are functionally important by identifying the highly strained –LHStaple bond and the “switch” mechanism (accounting for a large portion of structure variation: Figure 33).



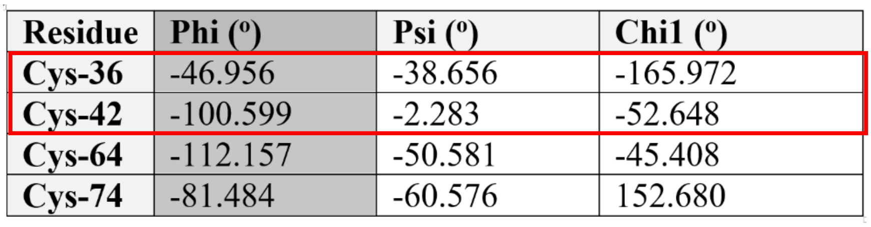
https://vimeo.com/819487033?share=copy

E.



https://vimeo.com/819489857?share=copy

Table 9.



Residue	Phi (°)	Psi (°)	Chi1 (°)
Cys-36	-46.956	-38.656	-165.972
Cys-42	-100.599	-2.283	-52.648
Cys-64	-112.157	-50.581	-45.408
Cys-74	-81.484	-60.576	152.680

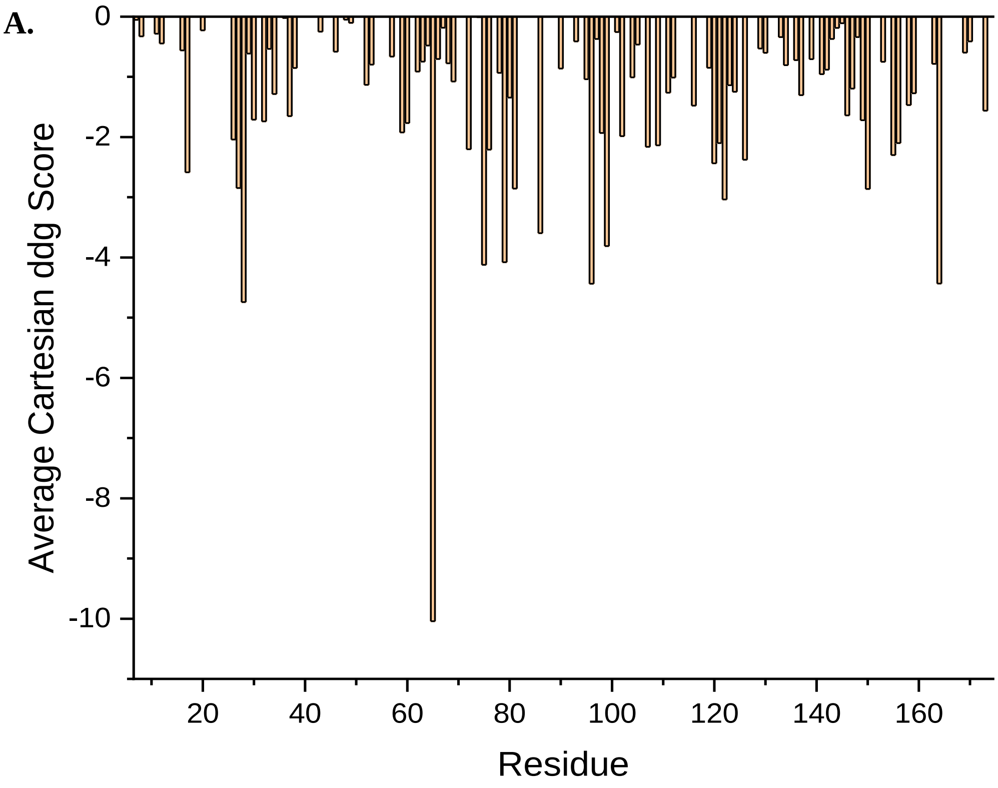
Figure 34. Allosteric Configuration of C36-C42 Disulphide Bridge. A. Unformed disulphide bridge between C36 and C42 after energy minimisation. Formed C36-C42 and C64-C74 disulphide bridges in PDB:2D9Q are shown in **B.** and **C.** respectively. **Table 9.** Dihedral angles for residues involved in disulphide bridges. **D.** and **E.** are screenshots from MD runs, at pH 4.4 and 7.4 respectively, focused on residues around the "switch" mechanism region. The simulation video can be watched with link below each image (copy and paste into browser).

5.1.3 Constructing Mutants with Rosetta

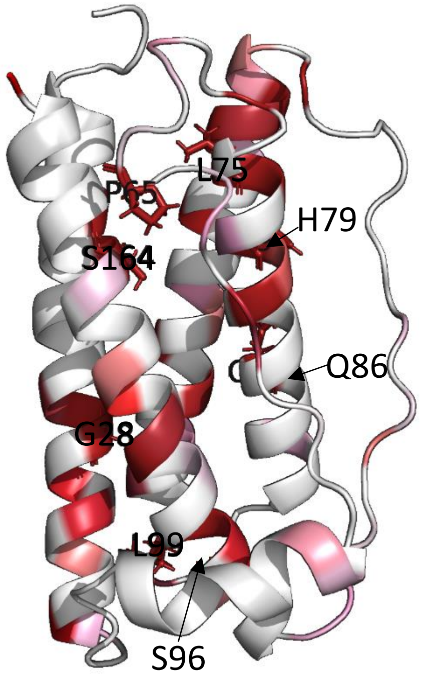
So far, NMR has identified regions subjected to significant remodelling during thermal denaturation and MD has confirmed most of these observations. Thus, semi-rational mutagenesis can be targeted to these regions. The Rosetta Cartesian_ddg application was used to identify stabilising mutations (Figure 35). This tool determines mutations to be stabilising when they better pack their own (and neighbouring residues within 6 Å) side chains and backbone (Park *et al.*, 2016; Frenz *et al.*, 2020). Figure 35a shows the average stabilising Cartesian_ddg score, where the more negative the score, the more stabilising the suggested mutations. Those with a score of 0 have no suggested stabilising mutations. These scores are mapped on to the structure of G-CSF in Figure 35b, where deeper red represents residues with a more negative score. Residues with a score below the 5th percentile (the most stabilising scores) are labelled and highlighted as sticks. The locations of these residues closely resembled those of highly conserved residues (Figure 25a) given their concentration around helix B, D and loop AB.

Moreover, although there was weak to no correlation between ΣΔδ at 323K and the average stabilising Cartesian_ddg score (Figure 35c), there is a general trend that the lower the ΣΔδ, the more conserved the residue (Figure 25a and S.12a) and stabilising the predicted mutations. Therefore, Rosetta may recognise that these conserved regions offer structural resilience. There is not a lot of exact overlap between conserved residues and those with the most stabilising mutations, although they are proximal, suggesting that backbone and side chain packing surrounding conserved residues may be sacrificed for global stability. Alternatively, this could hint that structure packing is not always equivalent to stability, which should be considered when using this Rosetta application. This is especially true for residues H156 and H79 because their side chains are close to (and interact with) residues in the neighbouring loop AB or CD (Figure 14). Hence, Cartesian_ddg suggests residues with smaller side chains to help with packing of the loops. However, these histidine side chains are likely important for stability of these loops and pH sensitive remodelling of these regions.

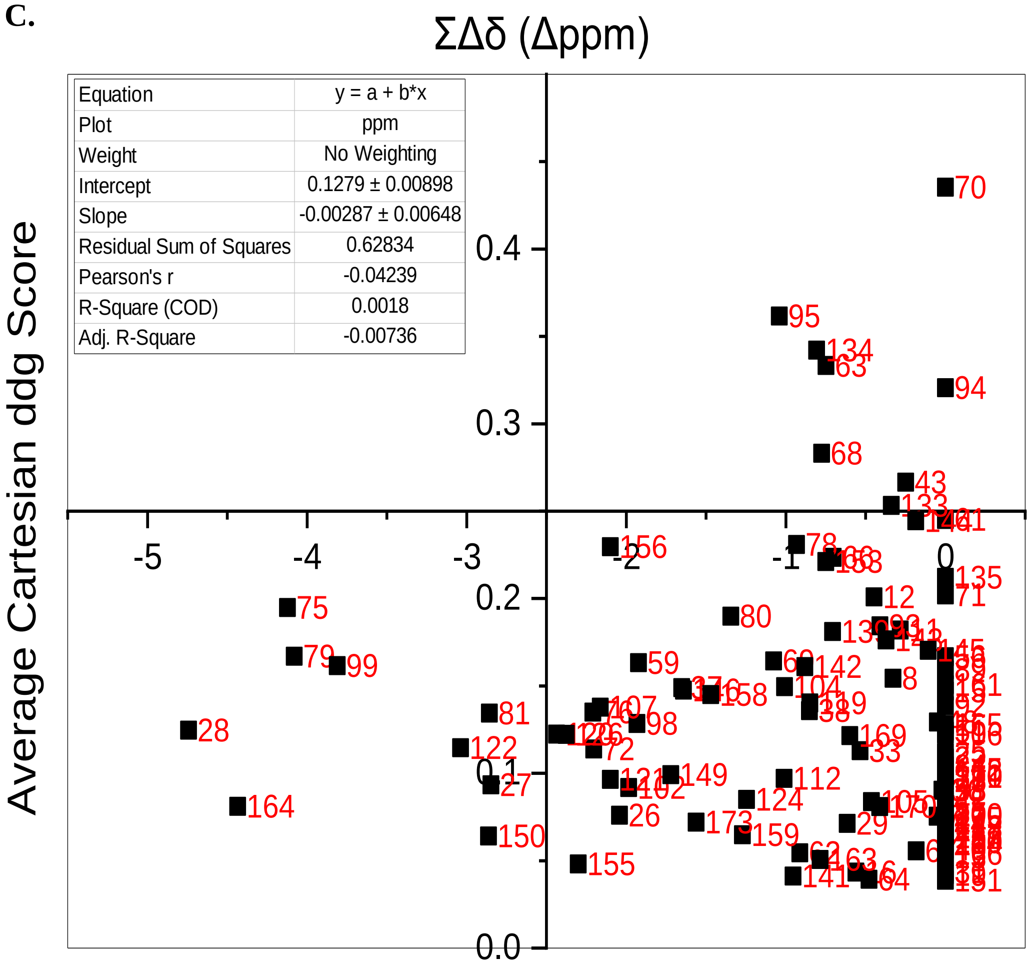
Nevertheless, the most stabilising mutations were predicted to be at residue P65, which is located in the C-terminal region of loop AB where NMR confirms there to be significant thermal-induced structural remodelling. Therefore, predicted stabilising mutations from Cartesian_ddg will be taken into consideration with NMR observations when synthesising these mutants *in vivo*.



B.



C.



Equation	y = a + b*x
Plot	ppm
Weight	No Weighting
Intercept	0.1279 ± 0.00898
Slope	-0.00287 ± 0.00648
Residual Sum of Squares	0.62834
Pearson's r	-0.04239
R-Square (COD)	0.0018
Adj. R-Square	-0.00736

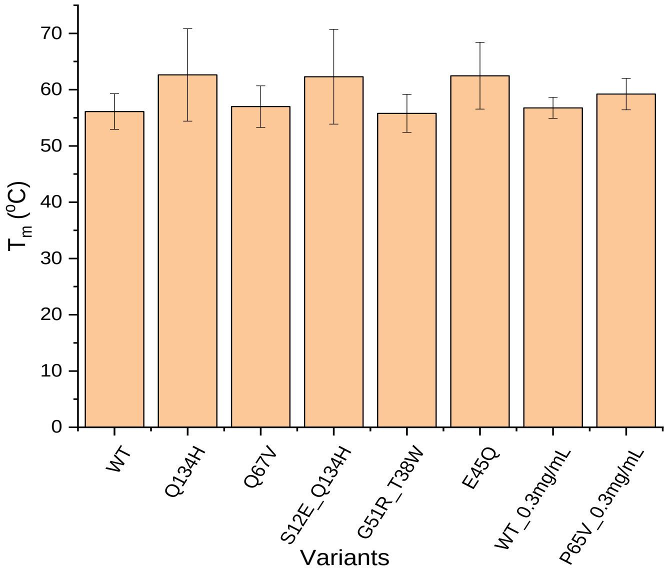
Figure 35. Identifying Stabilising Mutants with Cartesian_ddg. A. Averaged Cartesian_ddg score from all possible stabilising mutations. Negative scores are considered stabilising based on the equation Cartesian_ddg = mutant free energy – WT free energy. Residues with no stabilising mutations have a score of 0. **B.** Averaged stabilising Cartesian_ddg score mapped onto PDB:2D9Q. Deeper red colours represent more stabilising mutations and residues with stabilising mutations in the top 95th percentile are labelled and highlighted as sticks. **C.** Averaged stabilising Cartesian_ddg score vs WT ΣΔδ at 323K.

5.1.4 Mutant-Excipient Interactions Influence Thermostability

Semi-rational mutagenesis is approached here by considering NMR and *in silico* observations pertaining to the main structural and functional mechanisms of G-CSF. To this end, Rosetta's Cartesian_ddg application was used to construct stabilising mutations at these sites of interest, namely mutants; Q134H, Q67V, S12E, S12E_Q134H, P65V and H156F. Mutants E45Q and S12E_Q134H (a double mutant) were constructed based on NMR observations discussed below. The mutant G51R_T38W was constructed by combining two previous point mutations (Wood *et al.*, 2021) that MD shows to be in anti-correlating regions (Figure 32e). All mutants were synthesised as described in section 2.1 and possessed the correct molecular weight (Figure S.14). Successful expression and purification was achieved for all mutants except for S12E, S96V and H156F. The latter mutant was successfully expressed but failed at the refolding step, emphasising the structural importance of residue H156 (Figure 14).

Conformational stability is examined in Figure 36 using the van't Hoff thermal parameter: thermal unfolding temperature (T_m), where 50% of the protein population is unfolded. WT and mutant variants were formulated at 0.5 mg/mL in 50 mM sodium acetate at pH 4.25, with the exception of P65V and a comparative WT sample (concentrated to 0.3 mg/mL) given the difficulty of concentrating the mutant sample (Figure 36a). The T_m was averaged from four runs (n = 4) for all variants at 0.5 mg/mL and 0.3 mg/mL (Figure 36a) and three runs (n = 3) from variants at 0.15 mg/mL (Figure 36b-h). WT achieved a T_m similar to previously reported at all concentrations in Figure 36 (Wood *et al.*, 2021; Robinson *et al.*, 2018). Nevertheless, the mutants Q134H, S12E_Q134H, E45Q and P65V showed a different trend, with some having a difference of more than 10 ^{o}C in average T_m, perhaps hinting at a concentration effect. Figure 36b-h shows the T_m for WT and variants formulated with the same excipient conditions used in chapter 4; 12.5mM phenylalanine, 25mM histidine and 50mM arginine. The typically stabilising excipients phenylalanine and histidine improved T_m for all variants except for G51R_T38W, and Q134H, and Q67V in the case of histidine. This likely eludes to phenylalanine having little mutant-excipient interaction, supporting NMR observations that this excipient exhibits its stabilising affect by increasing the water tension of the protein surface. On the other hand, histidine had more of an impact on protein structure, explaining why this excipient has a less consistent stabilising effect. Arginine, unexpectedly, did not have a destabilising effect on WT, S21E_Q134H, E45Q and P65V. Reasoning for this could be that the concentration of protein was lower than that used for previous experiments, 0.3-0.5 mg/mL (Wood *et al.*, 2020), which could mirror the same effect seen from NMR where higher concentrations of arginine yield more structural changes. P-values from one-way ANOVA analysis indicate that a significant difference in thermostability is only seen when formulating mutants P65V, Q67V and S12E_Q134H with the different excipients. However, the large error bars in the majority of graphs indicate high run variability, suggesting that repeating this data may be necessary.

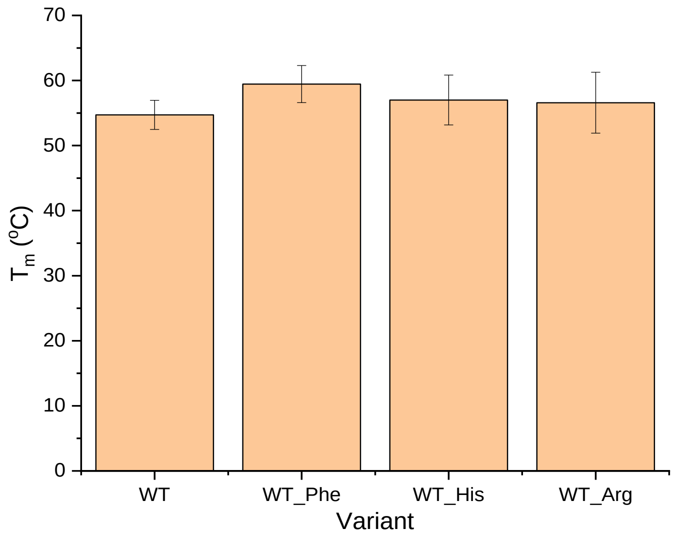
A.



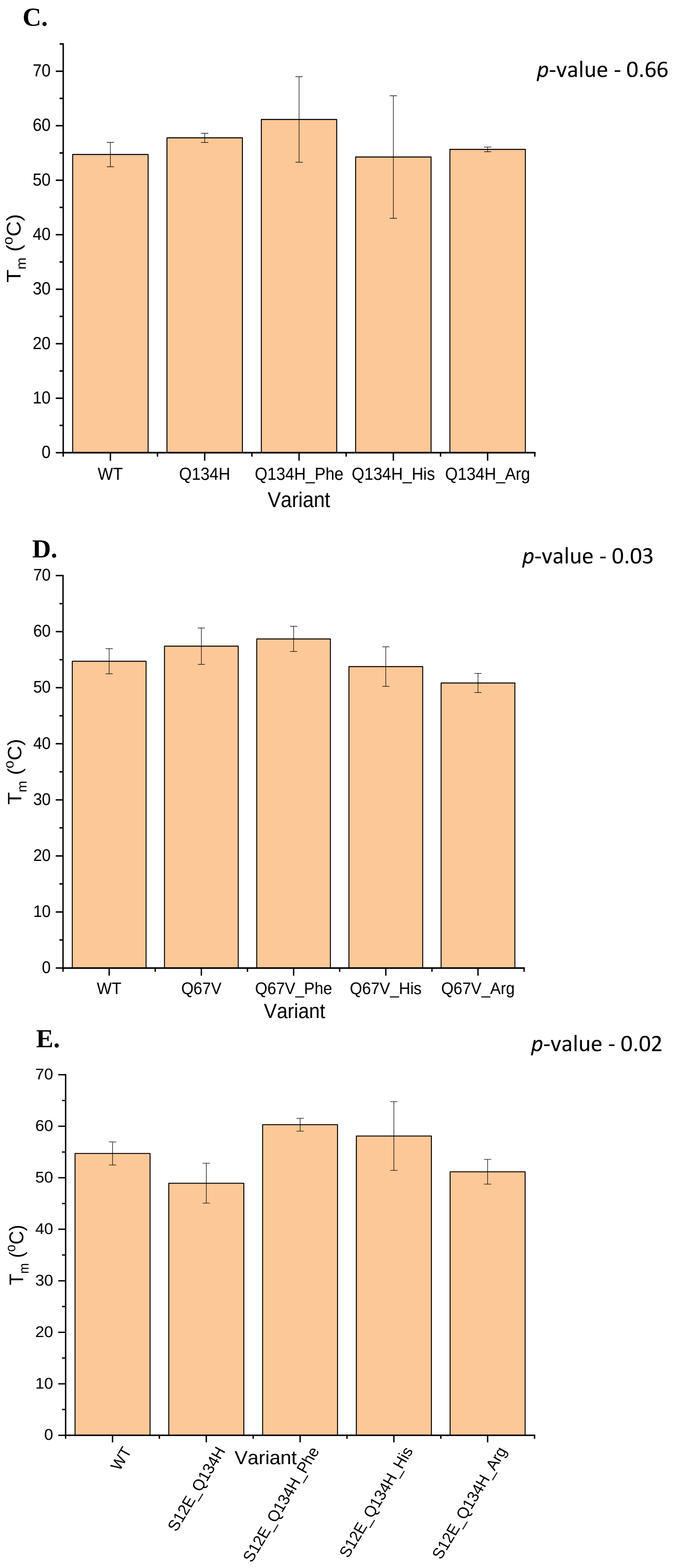
0.5mg/mL *p*-value - 0.30

0.3mg/mL *p*-value - 0.20

B.



p-value - 0.48


C.
p-value - 0.66
T m (oC)
70
60
50
40
30
20
10
0
WT
Q134H
Q134H_Phe
Q134H_His
Q134H_Arg
Variant
D.
p-value - 0.03
T m (oC)
70
60
50
40
30
20
10
0
WT
Q67V
Q67V_Phe
Q67V_His
Q67V_Arg
Variant
E.
p-value - 0.02
T m (oC)
70
60
50
40
30
20
10
0
WT
S12E_Q134H
S12E_Q134H_Phe
S12E_Q134H_His
S12E_Q134H_Arg
Variant

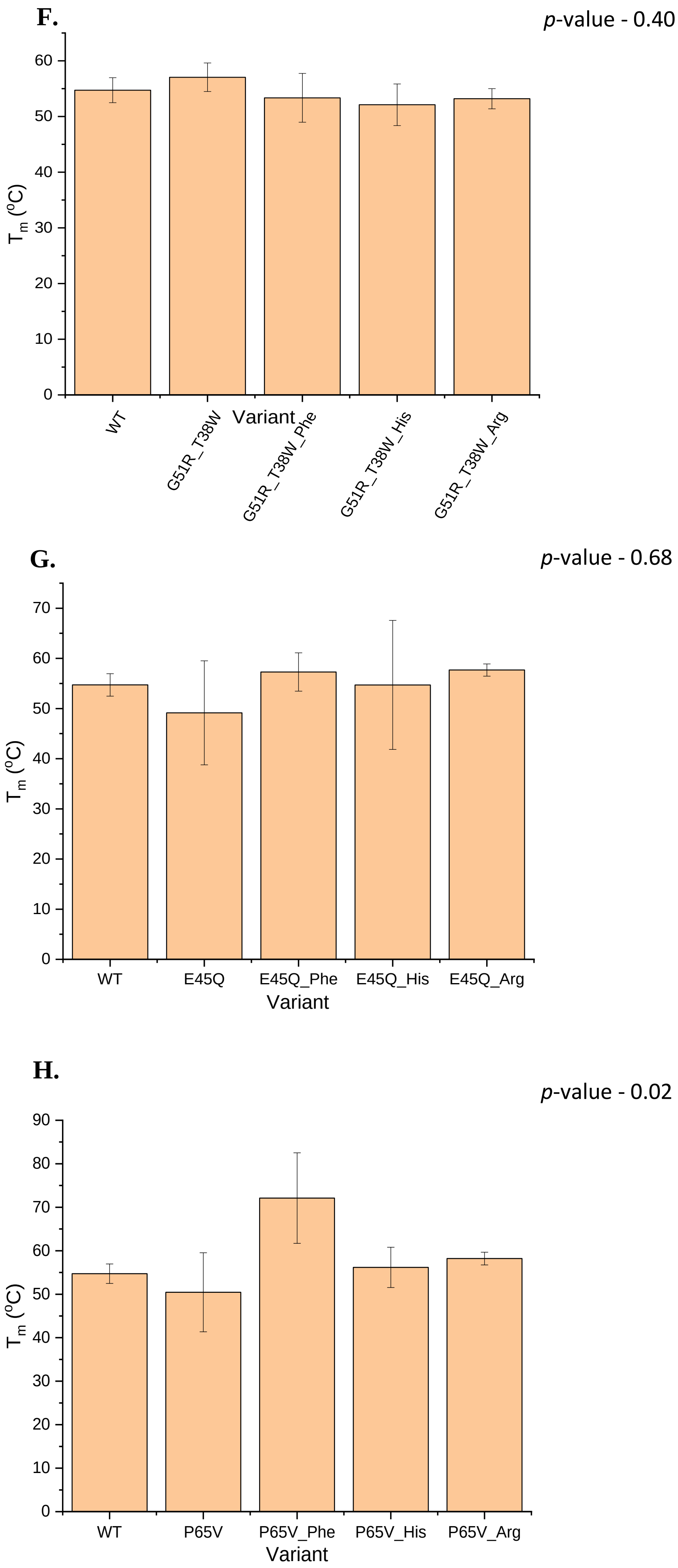

F.
p-value - 0.40
T_m (ºC)
0
10
20
30
40
50
60
Variant
WT
G51R_T38W
G51R_T38W_Phe
G51R_T38W_His
G51R_T38W_Arg
G.
p-value - 0.68
T_m (ºC)
0
10
20
30
40
50
60
70
WT
E45Q
E45Q_Phe
E45Q_His
E45Q_Arg
Variant
H.
p-value - 0.02
T_m (ºC)
0
10
20
30
40
50
60
70
80
90
WT
P65V
P65V_Phe
P65V_His
P65V_Arg
Variant

Figure 36. Examining Mutant Stability with T_m. The T_m was averaged from four runs in **A.** All variants were concentrated to 0.5mg/mL aside from P65V and a WT negative control, which were 0.3mg/mL. Variants in **B.-H.** were formulated to 0.15mg/mL with either 12.5mM phenylalanine (Phe), 25mM histidine (His), 50mM arginine (Arg) or more buffer (negative control) and T_m was averaged from three runs. P-values from one-way ANOVA analysis ($\alpha = 0.05$) are indicated in the top right corner of each graph. **A.** denotes the *p*-value for mutants compared with the control at 0.5mg/mL (top) and 0.3mg/mL (bottom).

5.1.4.1 Q134H and S12E_Q134H

A stabilising point mutation in regions T1-S12 and T133-A139 was anticipated to yield among the most impactful stabilising effects of all mutants because these regions appeared to be epistatic (chapter 4.1.8). Additionally, the two-state diffusion that occurs in both of these regions (suggested by linear peak trajectories in Figure 25) suggests that better packing of their backbone would slow this diffusion and improve structural resilience. Conversely, destabilising mutations in these regions are expected to have a more pronounced impact on stability than if they were in other regions. This has already been shown in previous mutation studies (Wood *et al*., 2021), where mutations proximal to T1-S12 and T133-A139 (F13A, Q131F and P132E) were among the most destabilising. The mutant Q134H (Figure 36a and c) was stabilising at 0.5mg/mL and 0.15mg/mL, possessing the highest stabilising effect amongst all mutants at 0.5mg/mL. Reasons for this could be the histidine at position 134 packed the backbone better, thus stabilising loop CD (and the N-terminal loop), substitution for a positively charged residue improves colloidal stability, or a mixture of these scenarios. This extra positive charge could explain why histidine's stabilising effect is hindered (compared to having no excipient) for this mutant but returns for the double mutant S12E_Q134H, where a negatively charged residue is substituted in. The large variation in T_m for mutants Q134H and S12E_Q134H with 25mM histidine could be because a more packed backbone around one of the excipient-induced cluster (red cluster in Figure 31b) could impact the excipient mechanism of action. However, the same outcome would be expected for arginine (Figure 31c and d), which showed much less variation.

Although expression of the S12E mutant failed, the mutant S12W was successfully tested (Wood *et al*., 2021) and showed a moderate decrease in stability. This likely explains the large drop in thermal stability from Q134H to S12E_Q134H because tryptophan also possesses a large sidechain like glutamic acid. The suggested epistatic relationship between T1-S12 and T133-A139 regions stipulates that the stabilising Q134H mutation would rescue the destabilising S12E mutation. While this is not the case at 0.15mg/mL protein concentration, it is true for 0.5mg/mL. The addition of 12.5mM phenylalanine improved the T_m of S12E_Q134H by 12°C, which comports with the notion that this variant is sensitive to increased surface tension of water, brought on by phenylalanine (section 4.1.8.1) and increased protein concentration (Figure 36a).

The stabilising Q134H mutation may be a dominant factor in this sensitivity because better packing of the loop CD backbone would be complemented by increased surface tension.

5.1.4.2 P65V and Q67V

Loop AB is pivotal to the structure-function relationship and as confirmed by NMR and MD (Figure 9d, 33c and d), this region undergoes significant structure remodelling, thus exposing the APR region on helix D (Figure S.10). Therefore, because G-CSF aggregation is limited by conformational stability (Raso *et al.*, 2005), improving the structural resilience of the loop AB C-terminus should improve conformational and colloidal stability. The mutant Q67V improves conformational stability at both protein concentrations (Figure 36a and d), whereas P65V only shows improvement at 0.3mg/mL. Aside from better packing, replacing the larger glutamine residue with the smaller valine in Q67V may have permitted more flexibility in this region, reducing strain on Q70-S8 (Figure S.9a) and W58-H156 (Figure 14) interactions. Nevertheless, substituting out a more rigid residue in the loop AB C-terminus proved to be detrimental to stability in the case of P65V. This mutant may serve as a caveat for software like Cartesian_ddg that try to relax/better pack the backbone (and proximal side chains) to improve protein stability. In this case, a more flexible loop AB backbone comes at the cost of conformational stability, perhaps because P65 was important for positioning of Q70 and W58 and local structural resilience.

The addition of 12.5mM phenylalanine to P65V reinforces the proposed mechanism of action of this excipient (Figure 31a). The focus of this excipient on stabilising the loop AB C-terminus compounds with the increased flexibility in this region for P65V, thus presenting as a benefit to conformational stability, possessing the highest average T_m of 72^{o}C. This underlines the importance of the VT-NMR approach presented in chapter 4 in its ability to identify potential regions for mutation-excipient relationships.

5.1.4.3 E45Q

Based on the “switch” mechanism, residue E45 plays an important role in capturing H43 when is makes its ring flip. MD shows this event to stabilise the unstructured loop preceding H43 (Figure 34e) given that the histidine side chain has less freedom to move. Therefore, the mutant E45Q was expected to take away electrostatic attraction to H43, without changing the local structure, thus confirming the importance of E45 to the “switch” mechanism. However, without this attraction, H43 could still be directed toward residue 45, as shown in Figure 34f, due to the positioning of P44 or perhaps the negative charge from E46. Thus, the drop in thermostability for this variant (Figure 36g) could be due to flailing of H43 as it tries to complete the “switch”. The structural remodelling of loop AB is likely sensitive to these N-terminal residues, meaning that H43 flailing could result in the compromised stability of E45Q.

NMR observations suggest that histidine as an excipient interacts with/significantly alters the conformation of E45 (Figure 31b). Hence, while the stabilising effect of this excipient was expected to be hindered when added to E45Q, it remained. This could be because the excipient instead interacts with E46, thus reducing H43 flailing as it cannot attempt to attract to E46. Alternatively the addition of cosolvent could just be generally stabilising to this variant given that all excipients improve T_m.

5.1.4.4 G51R_T38W

The point mutations G51R and T38W were mildly destabilising (Wood *et al*., 2021) and are located in dynamically correlated regions according to MD (Figure 32e). In both cases a smaller side chain is substituted for a larger one, which would likely rigidify the N-terminal loop AB region. This region relies on flexibility given its functional significance. Therefore, rigidifying with a point mutation may cause another proximal region of this loop to compensate by becoming more dynamic, compromising stability as a result. This proposed compensatory effect may explain the clusters of anti-correlation in this part of loop AB from MD. Improved stability resulting from combining the mutants G51R and T38W (Figure 36f) could be explained thusly: The compensatory effect of rigidifying one region is diminished by the addition of another rigidifying mutation. Hence, while this mutation enhances conformational stability, it should decrease bioactivity because of restricted loop AB movement. Nonetheless, this double mutant was only mildly stabilising at 0.15mg/mL and destabilising at 0.5mg/mL (by 0.4^{o}C). Furthermore, this is the only mutant for which all excipient conditions are destabilising. Both histidine and phenylalanine exhibit their stabilising effect via their influence on both termini of loop AB (Figure 31a and b). Therefore, their stabilising effect may be rendered useless given that G51R_T38W already stabilised this loop.

5.1.5 A Structure-function Relationship does not always lead to a Trade-off

Now that mutant thermostability had been assessed, the next step was to examine bioactivity. Taken together with thermostability studies, information yielded from bioactivity assays can help elucidate regions that are important to the structure-function trade-off. The mutant G51R_T38W concentrates two rigidifying mutations near to the “switch” region, which stabilises the protein. Therefore, functional assays, described in section 2.7, were conducted with this variant and WT at the various storage conditions and formulations outlined in Figure 37. Mutants were tested after FD and storage for a month at either -20^{o}C or 45^{o}C given the importance of this process in biotherapeutic storage. The excipients phenylalanine and histidine were selected given their stabilising effect (Figure 36) and consequently their ability probe the stability-functionality trade-off. The potential of histidine excipient to interaction with E45 also allows us to probe the importance of this residue to the "switch". Sorbitol and Mannitol were selected because of their cryoprotective abilities (Carpenter and Crowe, 1988; Storey and Storey, 1991).

Figure 37a-c illustrates GNFS-60 cell response curves for the G-CSF variant and excipient condition specified in the legends. All response curves decreased and converged into the baseline as the protein concentration was reduced, which signified successful titration. The strength of the response curves were generally spread out, showing some to achieve high CPM values while others barely achieved above the baseline. Moreover, all storage conditions showed WT to achieve the strongest response, whether formulated at 3.5% mannitol and 0.4% histidine or 3.5% sorbitol. The non-FD conditions gave the lower responses than the FD conditions, however this could just be more related to cell preparation than sample.

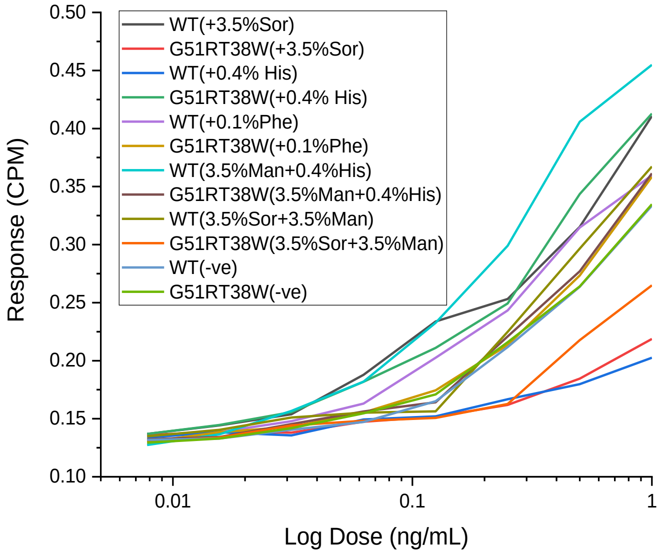
Table 10 denotes the change in response at respective storage conditions, from negative control to excipient addition, as Δresponse. Therefore, when the addition of an excipient increases the response, the Δresponse is 1. The opposite scenario is denoted with a value of -1 and 0 represents very little change when the curves appear to overlay. The final three table columns show the Δresponse between negative controls for WT and G51R_T38W, where a value of 1 signifies an improvement in response brought about by the double mutant. Overall, this table underlines consistencies in Δresponse for excipients, namely; the 3.5% sorbitol for WT, 0.4% histidine for WT and 0.1% phenylalanine for both variants. The response is improved for WT at 3.5% sorbitol and 0.1% phenylalanine and diminished at 0.4% histidine, all of which are excipients that improve thermostability of WT G-CSF. G51R_T38W maintained native-like bioactivity in the non-FD state but reduced in the in the post-FD state. Therefore, the slightly improved stability of this variant did not come at a cost to bioactivity until the FD process. This could be because the two mutations made loop AB more vulnerable to structural change during FD which hinders the functionally significant structural reconfiguration needed.

Phenylalanine is the only excipient to show identical Δresponse values for both G-CSF variants, where bioactivity is improved at non-FD conditions but reduced for both FD conditions. This does not follow the stability-bioactivity trade-off for non-FD WT, because phenylalanine improved stability, but it does for G51R_T38W. This could be because phenylalanine generally stabilises the protein-receptor complex, by reducing protein-protein interaction or stabilising the protein-receptor complex. Alternatively, this scenario could just happen for WT, and instead for G51R_T38W, phenylalanine offsets the rigidification of the loop AB N-terminus. Of note, phenylalanine was the only excipient that showed a potential change in the side chain rotamer of H43 (Figure 27h). Therefore, this increase in bioactivity reinforces the conclusion that the "switch" is not significantly altered by this excipient (Figure 29a). The diminished bioactivity for both FD variants could be attributed to loss in general structure because the excipient is not known to be cryoprotectant.

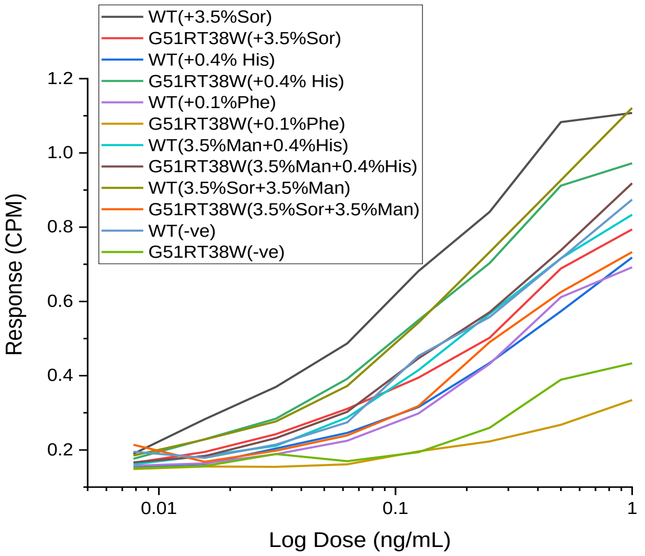
Sorbitol consistently improves the response for WT, pointing to a general stabilising effect similar to phenylalanine, with non-FD condition, but the added benefit of being cryoprotective as a sugar. The two main hypothesised stabilisation mechanisms that protect against lyophilisation are the "water substitution" (Carpenter, Arakawa and Crowe, 1992; Bjelošević *et al*., 2020) and "glass dynamics" (Chang *et al*., 2005; Franks, 1994). The water substitution mechanism hypothesises that excipients can replace the protein-water hydrogen bonds that are lost during the drying process, which thermodynamically favours the folded state. On the other hand, the glass dynamics mechanism stipulates that the cosolvent (particularly sugars) provides a rigid, inert matrix where the proteins are dispersed and mobility is very low (Pikal, 2004). Sorbitol reduces the response for G51R_T38W at non-FD and FD with 45ºC storage conditions. This perhaps indicates that the water substitution mechanism plays a more dominant role in sorbitol cryoprotection because mutations can render this excipient non-cryoprotective (Table 10 and Figure 38). If the glass dynamics mechanism was dominant, it would be less likely that mutations could affect the excipient's ability to form an inert matrix. The diminished response for G51R_T38W with sorbitol in the non-FD state may infer mutant-excipient interaction because these mutations alone do not decrease bioactivity.

Bioactivity diminishes at all conditions for WT when formulated with 0.4% histidine, reinforcing the proposed mechanism of action for histidine (Figure 31b). Excipient interaction or induced structural change at E45 has compromised bioactivity for WT, which also validates the importance of E45 to the "switch" mechanism. On the other hand, 0.4% histidine improved bioactivity for G51R_T38W at non-FD and post-FD -20ºC storage. Furthermore, bioactivity at these conditions for G51R_T38W with 0.4% histidine surpassed that of WT (negative control), as shown in Figure 37a and b. At post-FD -20ºC storage (Figure 37b), this revival of bioactivity was so profound that the response curve went from near baseline for G51R_T38W negative control to nearly to strongest response for this data set in the presence of histidine. Therefore, while the mutations for this variant may have disposed loop AB to functionally unfavourable structural change during the FD process, histidine may have acted as a cryoprotectant by interacting with the N-terminal loop AB region and preventing said structural change. Mannitol also acts as a cryoprotectant and appeared to prevent this mechanism of action for histidine given that when the two excipients are combined, bioactivity of WT either stayed the same or increased (Table 10).

A. Non-FD



B. Post FD -20



C. Post FD +45

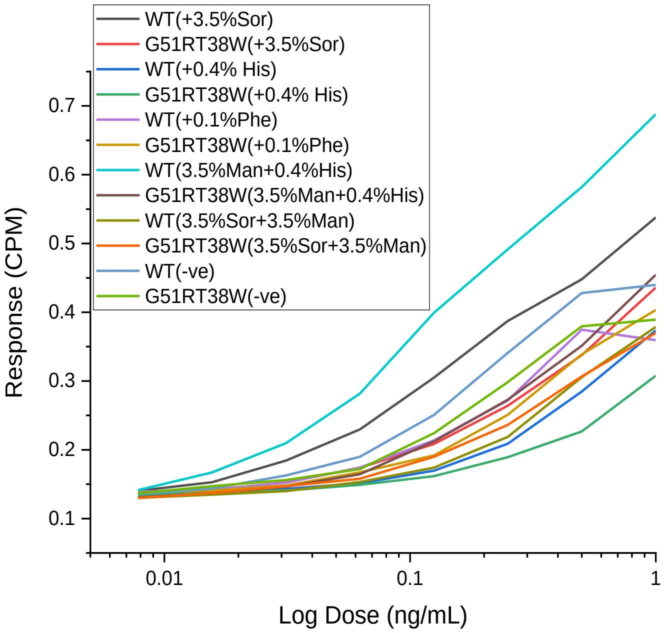


Table 10.

		Formulation																	
		Non FD (+3.5% Sor)	-20 (+3.5 %Sor)	+ 45 (+3.5 %Sor)	Non FD (+0.4% His)	- 20 (+0.4% His)	+ 45 (+0.4% His)	Non FD (+0.1% Phe)	- 20 (+0.1 %Phe)	+ 45 (+0.1% Phe)	Non FD (+3.5 %Man +0.4% His)	-20 (+3.5 %Man +0.4% His)	+45 (+3.5% Man+0.4%His)	Non FD (+3.5 %Sor +3.5% Man)	-20 (+3.5 %Sor +3.5 %Man)	+ 45 (+3.5 %Sor +3.5 %Man)	G51R_T38W	- 20 G51R_T38W	+45 G51R_T38W
Δ Response	WT	1	1	1	-1	-1	-1	1	-1	-1	1	0	1	1	1	-1	0	-1	-1
	G51R_T38W	-1	1	-1	1	1	-1	1	-1	-1	1	1	0	-1	1	-1			

Figure 37. Bioactivity of G51R_T38W vs WT. A.-C. GNFS-60 response curves for WT and G51R_T38W formulated with either sorbitol (Sor), histidine (His), phenylalanine (Phe), mannitol (Man) or more buffer (-ve). Mutants were tested without undergoing FD (non-FD) and after FD and storage for a month at either -20ºC or 45ºC. **Table 10**. Δresponse of 1, -1 and 0 means the addition of an excipient increases, decreases and has very little change to the response curve. The final three table columns show the Δresponse between negative controls for WT and G51R_T38W.

Following bioactivity assays with G51R_T38W, the mutants constructed based on NMR observations (E45Q, P65V, Q67V, Q134H, S12E_Q134H) were also assessed using the same functional assays (Figure 38). Given the greater number of mutants to test, excipient conditions were narrowed down to 0.4% histidine (2) and 3.5% (3) sorbitol due to their potential for mutation-excipient interaction (Figure 37). Storage conditions were also narrowed down to non-FD and FD (stored at -20ºC for a month). The dilution for these assays appeared less successful than previously with G51R_T38W because many of the response curves were less smooth. The large deviation of some data points from the titration curve meant that some were omitted. This must have resulted from a sample problem instead of a problem with cells or assay preparation because the **G-CSF standard** produced a smooth response curve. This sample problem could result from bad transportation of samples between UCL and NIBSC. Nevertheless, the strength of the response curve for these mutants was comparable to that of the G51R_T38W assay and the general trends that histidine and sorbitol exhibit on WT bioactivity remained. Moreover, the strength response curves was spread out and they all eventually reached a baseline (Figure 38a and g).

5.1.5.1 Q134H and S12E_Q134H

The mutant Q134H increased thermostability (Figure 36a and c) and bioactivity (Figure 38f), albeit mildly. This may be accredited to the mutation being positioned in a region, not directly important to bioactivity, eliciting structural and dynamic correlation with other regions. Resultantly, this mutation likely had a stabilising effect on the protein-receptor complex without affecting the structural changes in loop AB needed to exposed binding site-III.

Both WT and Q134H had shallow response curves post-FD (Figure 38l), revealing that the Q134H mutation did not offer cryoprotection. However, unlike WT, 0.4% histidine improves bioactivity for Q134H post-FD, hinting at a structural change in the excipient red cluster (Figure 27b) caused by the mutation, as earlier discussed. Although, this structural change appeared to reduce histidine's stabilising capability for Q134H in the non-FD state (Figure 36c), it may be what promotes water substitution in this region during the FD process. The double mutant S12E_Q134H saw a decrease in bioactivity (Figure 38b) but an improvement in cryoprotection

(Figure 38h). Moreover, the return of histidine's stabilising effect to this double mutant (Figure 36e) was accompanied by an increase in bioactivity with 0.4% histidine (and 3.5% sorbitol) at both storage conditions. What's more, S12E_Q134H with 3.5% sorbitol yielded the strongest response curve in the non-FD state (with 0.4% histidine not far behind), which suggests that this mutant improves excipient interact (Figure 38a). This improvement could result from an extra negatively charged residue (improving histidine's interaction) or structural disruption from the large side chain of glutamic acid, which could improve excipient interaction near the loop AB C-terminal (Figure 30a).

5.1.5.2 P65V and Q67V

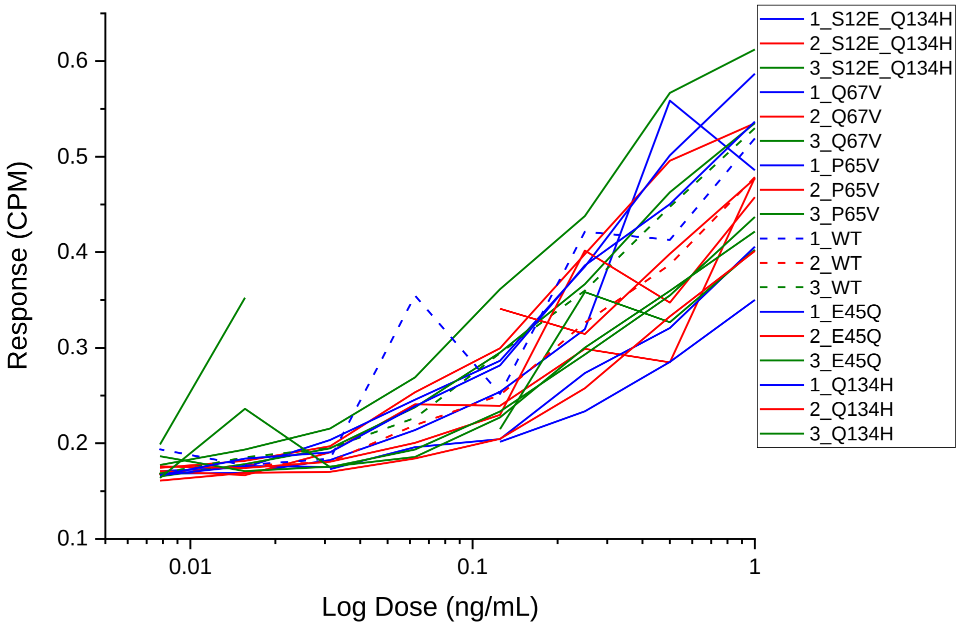
The mutant P65V displayed a diminished bioactivity at both storage conditions (Figure 38d and j). Therefore, taken together with its relatively low thermostability at 0.15mg/mL (Figure 36h), this mutant alone promoted undesirable qualities. The reasoning for this is likely the greater freedom of movement that the valine mutation gives the loop AB C-terminus, as previously proposed. Consequently, this loop could become more dynamic, thus compromising its conformational stability and binding site-III-receptor interaction. It is difficult to interpret the effect that 0.4% histidine and 3.5% sorbitol had on P65V in the non-FD state because of the noise of the data. However, these excipients rescued bioactivity for P65V post-FD. Given that all excipients rescued the thermostability of this mutant in Figure 36h, the general stabilising effect of co-solvent could have prevented mutation-inducted conformational changes in loop AB, protecting P65V from FD.

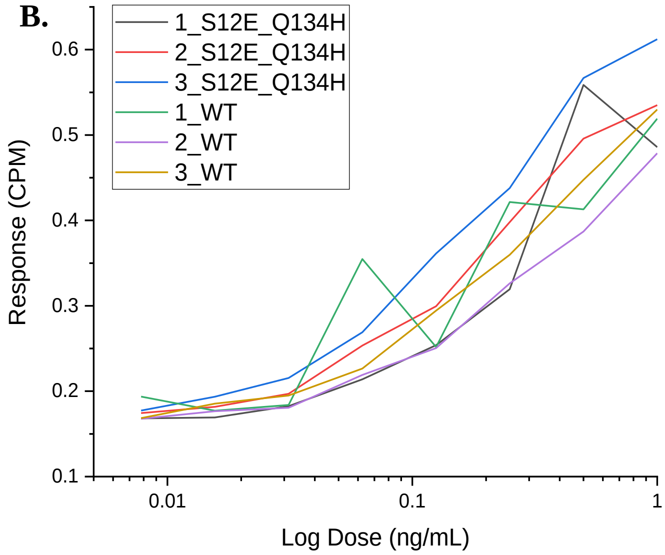
The packing of the loop AB C-terminus was likely better for Q67V than P65V because this mutant yielded improved thermostability and bioactivity compared to WT (Figure 36d, 38c and i). In addition, Q67V possessed the highest non-FD response curve out of all variants without excipients (Figure 38a). This supports the previous suggestion that the valine mutation at Q67 improves Q70-S8 hydrogen bonding, which would stabilise much of loop AB given the significant thermal-induced environmental change that Q70 experiences (Table 4a). Excipients either decreased the response curve for Q67V, for the non-FD condition, or had little effect post-FD (Figure 38c and i). The stabilisation of loop AB by Q67V is perhaps what ablates the stabilising effect of these excipients (Figure 36d), which translates to little change in/diminished bioactivity.

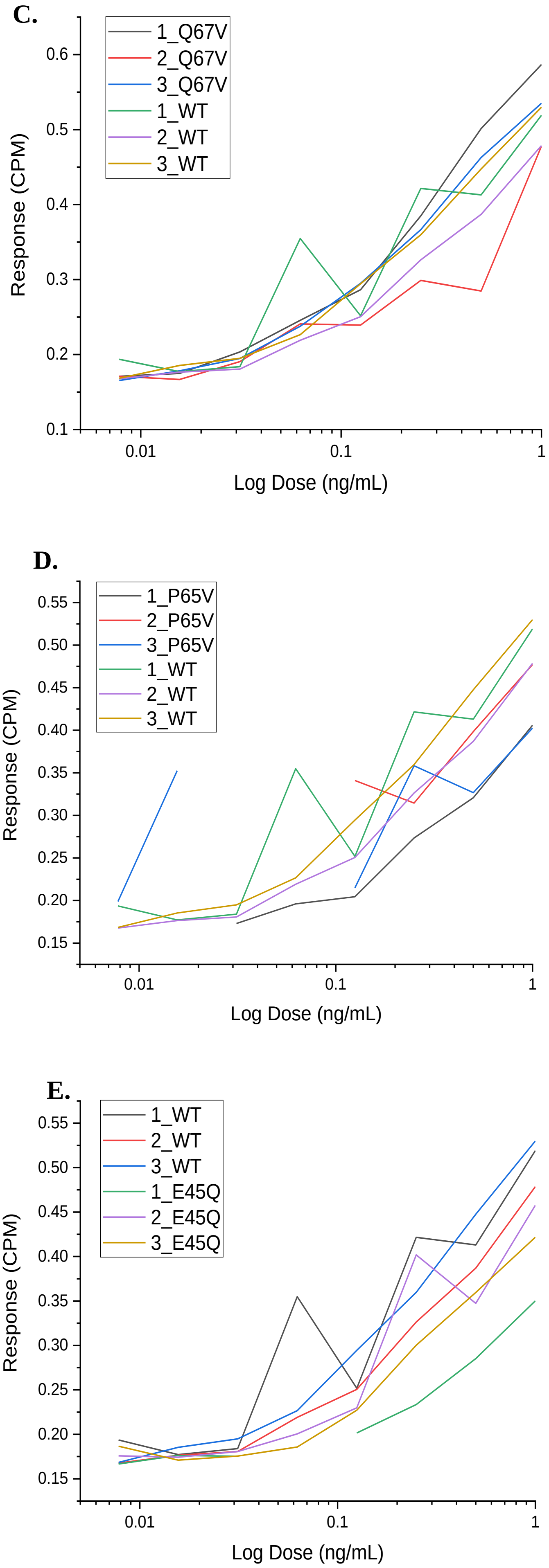
5.1.5.3 E45Q

The importance of residue E45 to the "switch" mechanism was emphasised by E45Q because this mutant yielded the lowest bioactivity amongst all variants at the non-FD condition (Figure 38e). Decreased E45Q thermostability caused by H43 flailing, as earlier proposed, appears to have translated into a destabilised binding site-III and resultantly diminished bioactivity. However, the rescue of E45Q thermostability by histidine (Figure 36g) also rescued bioactivity at both storage conditions (Figure 38e and k). This contradicts the notion of histidine excipient interaction with residue E46 instead of E45 for E46Q because E46 is part of binding site-III and excipient interaction would hinder receptor interaction. Nevertheless, if histidine gets displaced from residue E46 during the receptor binding process, i.e. binding site-II interaction before site-III, then this would explain the rescue qualities of the excipient.

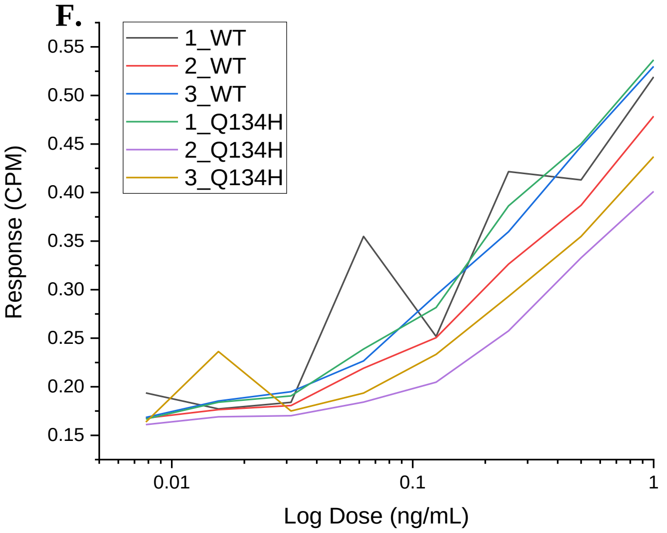
Non-FD Bioassays





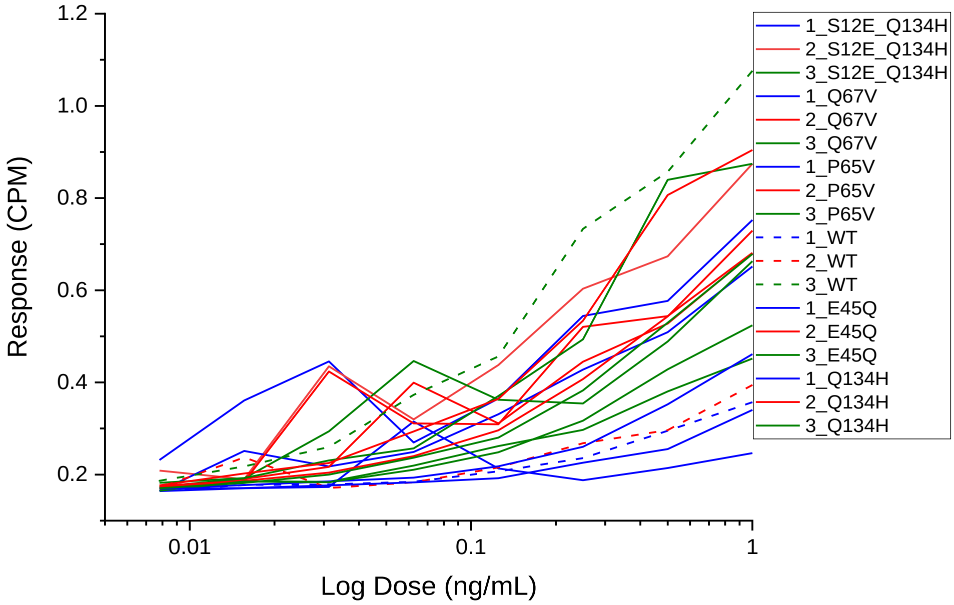

C.
1_Q67V
2_Q67V
3_Q67V
1_WT
2_WT
3_WT
Response (CPM)
0.1
0.2
0.3
0.4
0.5
0.6
0.01
0.1
1
Log Dose (ng/mL)
D.
1_P65V
2_P65V
3_P65V
1_WT
2_WT
3_WT
Response (CPM)
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.01
0.1
1
Log Dose (ng/mL)
E.
1_WT
2_WT
3_WT
1_E45Q
2_E45Q
3_E45Q
Response (CPM)
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.01
0.1
1
Log Dose (ng/mL)

F.

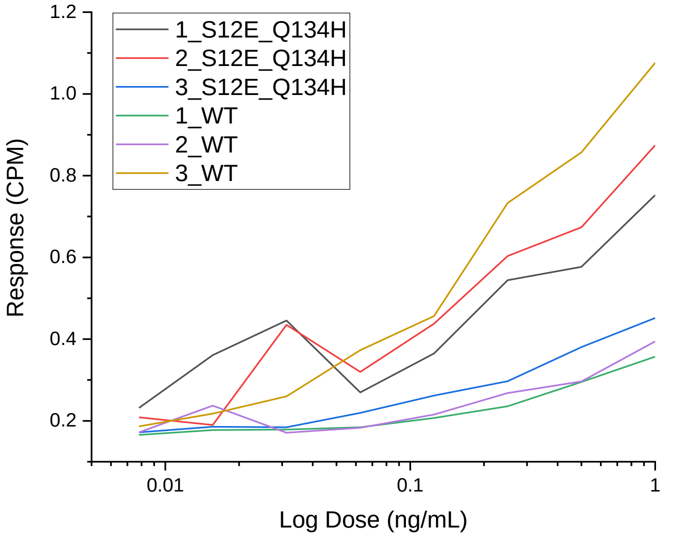


FD Bioassays

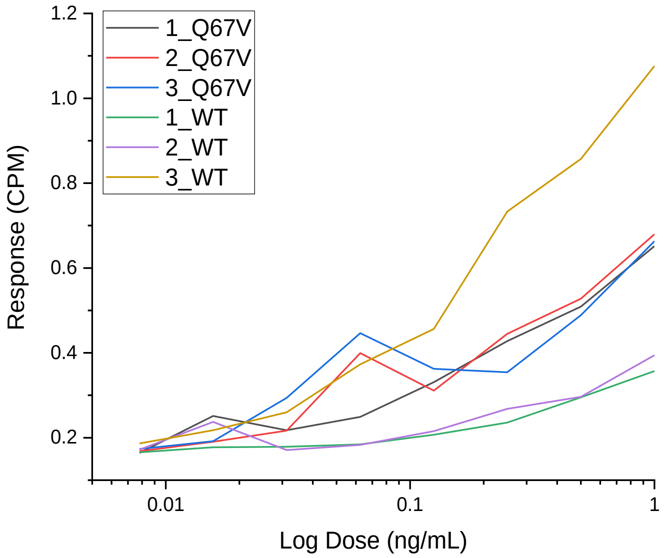
G.



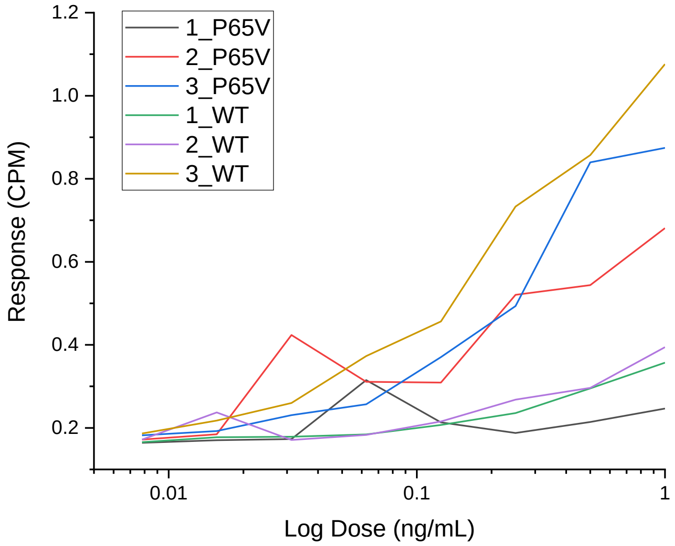
H.



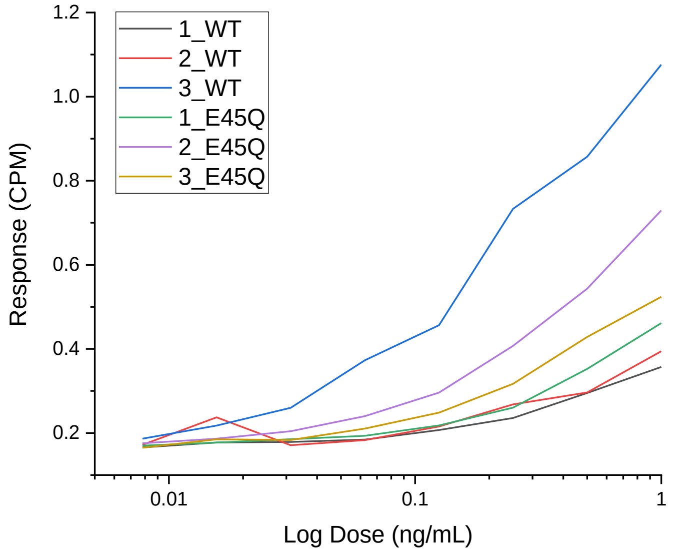
I.


1_Q67V
2_Q67V
3_Q67V
1_WT
2_WT
3_WT
Response (CPM)
1.2
1.0
0.8
0.6
0.4
0.2
0.01
0.1
1
Log Dose (ng/mL)

J.


1_P65V
2_P65V
3_P65V
1_WT
2_WT
3_WT
Response (CPM)
1.2
1.0
0.8
0.6
0.4
0.2
0.01
0.1
1
Log Dose (ng/mL)

K.


1_WT
2_WT
3_WT
1_E45Q
2_E45Q
3_E45Q
Response (CPM)
1.2
1.0
0.8
0.6
0.4
0.2
0.01
0.1
1
Log Dose (ng/mL)

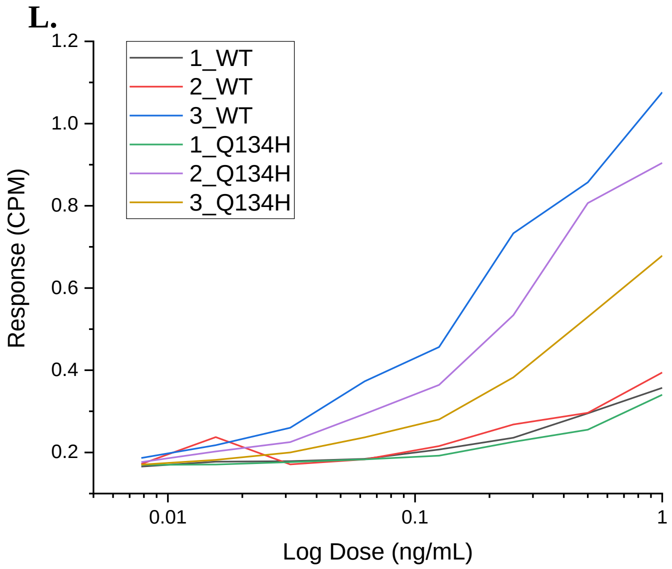


Figure 38. Bioactivity of Mutants Designed from NMR Observations. GNFS-60 response curves for variants without FD treatment (**A.-F.**) and after FD and storage at -20ºC for one month (**G.-L.**). Variants are formulated with more buffer as a negative control (1), 0.4% histidine (2) or 3.5% sorbitol (3).

5.2 Discussion

5.2.1 In Silico Modelling Guides Protein Engineering Approaches

MD correctly predicted the change in G-CSF dynamics between α-helices and unstructured regions as well as between pH 4.4 and 7.4 (Figure 32a and b). Moreover, MD correctly underlined the large conformational change in loop AB relative to the rest of the protein (Figure 33c and d) and cooperative manner in which these changes take place (Figure 32e). Therefore, the general trends observed from NMR were picked up with computation, making the case that MD is at least effective as a guiding tool in understanding native ensemble dynamics. The all-atom MD approach used here was not sufficient to make conclusions on G-CSF behaviour upon unfolding due to the use of relatively low temperatures and timescales. CG-MD offers a computationally cost-effect way of probing these longer timescale dynamics needed to observe unfolding. Although larger global structure perturbation was observed from the CG-MD run in Figure S.13, the analysis conducted could not conclude that unfolding occurred. However, CG-MD sacrifices resolution about the residues because they are simulated as beads. This resolution can make all-atom MD effective when engineering proteins because residue-level information can narrow down the target. Nevertheless, analysis on all-atom MD with G-CSF highlights regions instead of individual residues for their importance to structural and dynamic changes, at least until compared with NMR observations. The prime example of this being the "switch" mechanism (Figure 34d and e), which was only identified hindsight after the VT-NMR experiments. This highlights an issue with analysis and being able to extract these significant functional mechanisms. Of note, the "switch" mechanism occurred on a longer timescale than tested with MD and was still identified. In this case, events could have unfolded quicker because of the missing C36-C42 disulphide bond, which would have stabilised the region more. Hence, a potential question is: Are we keeping up with what MD is telling us? Residue C36 is the only residue that stood out as significant without context from NMR because it did not disulphide bond with C42 and was characteristic of an allosteric –LHStaple configuration.

Computational modelling with Rosetta's Cartesian_ddg application (Figure 35b), on the other hand, can reveal specific residues that compromise structural resilience as opposed to broad areas like with MD. These suggested residues generally elicited structural resilience (Figure 35c) from NMR and were either highly conserved residues themselves or near them (Figure 25a). Therefore, Cartesian_ddg appears to highlight residues that are important to conserved regions, and thus protein viability, due to the resilience they offer. Consequently, this application sometimes suggests stabilising mutations for residues that may prevent local structure packing (i.e. H79 and H156) but form significant interactions with neighbouring regions. Hence, utilising Rosetta without contextual knowledge of the protein does not advance a semi-rational mutagenesis approach.

5.2.2 Combining VT-NMR with In Silico Modelling adds Rationale to Protein Engineering

The semi-rational approach to engineering G-CSF presented here contextualises *in silico* modelling with NMR observations. The mutant P65V underlines the importance of this approach because it could have been regarded as an instable (at 0.15mg/mL) and inactive mutant (Figure 36h and 38d) if potential excipient-mutation interaction not been identified. A valine substitution for residue P65 was the most stabilising mutation predicted from Rosetta, which comported with NMR data because this loop AB region experienced the largest thermal-induced environmental changes. However, this mutation likely gave this region more freedom to move, which, taken together with phenylalanine's mechanism of action (Figure 31a), endowed P65V in 12.5mM phenylalanine with the highest thermostability amongst the variants.

The rationality behind the mutant H156F was more conflicting than other variants. This was because VT-NMR (Table 4) showed significant $\Sigma\Delta\delta$ and dynamics for this residue (and its spatial neighbour, W58 for dynamics) and Rosetta highlighted H156 as disruptive to the packing of local structure. Thus, these observations appear to portray H156 as a cause of disruption to its proximal loop AB region. On the other hand, a previous NMR study highlighting H156 as part of a pH sensitive region suggests that this residue forms a cation-π bond with W58, which is important to loop AB stability (Aubin *et al*., 2015). The latter observation is likely the reason behind the failure to refold H156F, perhaps suggesting that the high $\Sigma\Delta\delta$ and dynamics for H156 was owing to the attempt for this residue to keep W58 (and neighbouring loop AB) packed onto helix D.

Nevertheless, all mutants formulated at 0.5 or 0.3 mg/mL improved thermostability on average (except for G51R_T38W where it remained the same). Furthermore, the mutant Q67V improved both bioactivity and stability, and histidine excipient rescued bioactivity for mutants G51R_T38W and E45Q. Therefore, the presented engineering approach combining VT-NMR with *in silico* modelling provides (semi-) rationality behind constructing mutants and formulations to yield desirable protein characteristics. A larger library of mutants, both semi-rationally designed and negative controls, would be preferred because provide more statistical robustness behind this approach. In addition, more amino acid excipient conditions would ideally have been studied for mechanism of action and possible interactions with mutations pertaining to stability and bioactivity. Arginine was not used in bioactivity assays because it is typically destabilising (Wood *et al*., 2020), however, it improves Tm for some mutants (Figure 36) and could possibly rescue bioactivity post-FD (Stärtzel, 2018). Hence, going forward, this excipient should be included in bioassays along with phenylalanine to assess if they save bioactivity as well as stability.

6 Conclusion

The consistency of these findings with conclusions from previous and current biophysical studies not only shows that NMR can mechanistically detail the reasons for these studies, it validates mapping functionally and structurally significant regions to G-CSF. Achieving the first aim of this project provided insight to four structural clusters, concentrated around loop AB, that were pivotal to the structure-function relationship for G-CSF. The potential for significant structural remodelling within these clusters (promoted by residues H43, V48, Q70, H156 and the C36-C42 –LHStaple disulphide bond in particular) endows G-CSF with an adaptation to bone marrow. These observations were key to addressing the second project aim of determining mechanisms of (de-)stabilisation for excipients. Probing excipient mechanisms of action demonstrated that the four structural clusters, "switch" mechanism and concerted structural and dynamic changes largely remained the same with all co-solvents. This emphasises these importance of these features to the protein and the ability of NMR to identify them. The different mechanisms of action exhibited by each excipient were highlighted by another set of structure clusters, this time identified by changes in hydrogen bond strength, side chain rotamer, conformational diffusion and potential interaction sites. 12.5mM phenylalanine and 50mM arginine generally had the largest impact on conformational remodelling, exemplified by such changes for residues T38, H43, I56, Q70, A72, L92 and I95. What's more, the impact of 50mM arginine on L92 had the most pronounced knock-on effect, showing major disruption to the typical concerted structural changes in G-CSF. On the other hand, 25mM histidine and arginine had a milder impact in this respect, showing stronger interaction with the protein instead.

MD confirmed the major site of stabilisation for phenylalanine (in loop AB and helix D) and the potentially pH-sensitive cluster around residue H79 for arginine, thus supporting the use of these clusters for engineering G-CSF. *In silico* modelling proved useful in cross-referencing NMR observations and designing mutants to address the final project aim of validating the use of VT-NMR for semi-rational protein engineering. MD correctly predicted the significant conformational change in loop AB and the "switch" mechanism and Rosetta's Cartesian_ddg highlighted residues in highly conserved regions that play an important role in thermal resilience. Therefore, considering these observations alongside the allosteric –LHStaple for the C36-C42 disulphide bond, it could be argument that MD alone provided rationality behind mutating regions to significant to structure and function. However, the mutants P65V and H156 were destabilising despite being predicted as stabilising from Cartesian_ddg. This was due to a lack of context regarding structural significance, which when obtained from NMR allowed the rescue of stability and functionality for mutants E45W, P65V and G51R_T38W. The context provided by NMR outlined potential for excipient-mutation interaction, leading to the high T_m of 72°C for P65V in

12.5mM phenylalanine. Furthermore, the mutants Q67V and Q134H increased both stability and bioactivity, defying the typical structure-function trade-off. Therefore, while *in silico* modelling mirrors experimental observations on many occasions, it has proven to be more effective as a guiding tool in this study.

Overall, the semi-rational protein engineering approach presented here successfully outlines mutants (also in combination with excipients) that improve protein fitness. Moreover, this approach outlines key phenomena, such as structure-function trade-off and epistasis, which underpin protein survival. However, this approach can be labour intensive (especially for NMR data processing) and yields a low number of mutant and formulation variants to test. Consequently, this approach currently would not likely be appeal to mainstream protein engineering, as these issues would have a monetary cost not accompanied by a large explored mutational space. In addition, NMR can only currently examine proteins up to ~40 kDa, which is not sufficient for many therapeutic proteins. Nevertheless, examining separate chains of larger proteins, automating VT-NMR spectra assignment and identification of structural clusters and combining this with high-throughput mutation generation and analysis could offset some of these issues. This could guide the direction of these mutant libraries so that the optimal experimental space is explored.

7 Future Work

7.1 Validating Current VT-NMR Approach

One of the largest validation problems facing this semi-rational engineering approach is the mutant library size, thus, immediate next steps would be to expand it by including mutations in loop BC, helix D and negative control mutations outside of structure clusters. Another validation problem is the fact that this approach has only been tested on one protein, which could be addressed by applying it to other proteins with backbone assignments. In addition, the fastest way to test the ability of this approach to probe epistasis (like with residues T1-S12 and T133-A139 in G-CSF) would be to examine proteins that are already suspected to possess epistatic regions, such as cheZ phosphatase and E1 glycoprotein.

Using VT-NMR to analyse the library of mutants and formulation conditions in this study would reinforce the conclusions made. This is because detailed mechanisms of action have been proposed and, thus, NMR observables should reflect the changes made. The mutants (and formulations) that should show the most significant of these changes are P65V with 25mM phenylalanine, E45Q and G51R_T38W (also with 0.4% histidine). WT G-CSF was briefly compared before and after FD with NMR and, given the significance of FD in biotherapeutic storage, should be further explored. This would include assessing cryoprotective properties of the excipient sorbitol and buffer sodium acetate on the structure of G-CSF.

7.2 Improving Computational Methods

In silico modelling proved effective in cross-validation with NMR observables and should be further explored in its predictive capabilities. This would be achieved by expanding comparative computational data with more MD simulations, for example; higher temperature MD, longer run times, thermal replica-exchange MD and simulating multiple G-CSF molecules in one system. On a longer timescale, comparing more proteins analysed by the same VT-NMR approach with *in silico* methods could develop these computational methods by adding a competitiveness similar to the Critical Assessment of protein Structure Prediction (CASP) community. Here, instead of predicting structure, methods could be developed to better predict/identify structure remodelling significant to function and stability. Furthermore, a large enough library of proteins that are comparable with VT-NMR would make machine-learning approaches to predict VT-NMR data more feasible. Improving these tools that predict protein behaviour will facilitate biosimilars and *de novo* protein design, thus advancing the market of novel therapeutics.

7.3 Workflow for the Ideal VT-NMR-guided Protein Engineering Approach

The automation of as much of the VT-NMR analysis process as possible is crucial to its widespread use. Fortunately, automation of cross-peak propagation is made possible by the Shift-T web server (Trainor et al., 2020). Therefore, focus can be shifted to automation post-initial data processing where significant residues can be determined and mapped onto protein structure. This includes top percentiles of ΣΔδ, PI and percentage increase in PI, temperature coefficients, temperature-lines and peak trajectory linearity. Therefore, this would automatically map regions of interest on to G-CSF structure, guiding the engineer to look in detail at these areas in simulation or other computational methods. Highlighting regions of interest would also function like a design of experiments platform, guiding mutation libraries to focus on certain spots. This would also help with implementing epistatic mutations (in correlating regions from VT-NMR) to higher degrees of order because larger combinations of mutations can be constructed. Manufacturers of biosimilars would benefit from this approach because it removes the arduous steps from the current VT-NMR process and provides a semi-rational path to develop successful mutants that would show a significant improvement from the originator molecule.

8 Bibliography

Abdul-Fattah, Ahmad M., Vu Truong-Le, Luisa Yee, Lauren Nguyen, Devendra S. Kalonia, Marcus T. Cicerone, and Michael J. Pikal. "Drying-induced variations in physico-chemical properties of amorphous pharmaceuticals and their impact on stability (I): Stability of a monoclonal antibody." *Journal of Pharmaceutical Sciences* 96, no. 8 (2007): 1983-2008.

Aldeghaither, D., Smaglo, B. G. and Weiner, L. M. (2015) 'Beyond peptides and mAbs - Current status and future perspectives for biotherapeutics with novel constructs', *Journal of Clinical Pharmacology*. doi: 10.1002/jcph.407.

Andersen, N.H., Neidigh, J.W., Harris, S.M., Lee, G.M., Liu, Z. and Tong, H., 1997. Extracting information from the temperature gradients of polypeptide NH chemical shifts. 1. The importance of conformational averaging. *Journal of the American Chemical Society*, *119*(36), pp.8547-8561.

Arakawa, T. and Timasheff, S. N. (1985) 'The stabilization of proteins by osmolytes', *Biophysical Journal*. doi: 10.1016/S0006-3495(85)83932-1.

Arbely, E., Neuweiler, H., Sharpe, T.D., Johnson, C.M. and Fersht, A.R., 2010. The human peripheral subunit-binding domain folds rapidly while overcoming repulsive Coulomb forces. *Protein Science*, *19*(9), pp.1704-1713.

Aritomi, M., Kunishima, N., Okamoto, T., Kuroki, R., Ota, Y. and Morikawa, K., 1999. Atomic structure of the GCSF–receptor complex showing a new cytokine–receptor recognition scheme. *Nature*, *401*(6754), pp.713-717.

Armstrong, D.A., Kaas, Q. and Rosengren, K.J., 2018. Prediction of disulfide dihedral angles using chemical shifts. *Chemical science*, *9*(31), pp.6548-6556).

Arvedson, T. L. and Giffin, M. J. (2012) 'Structural biology of G-CSF and its receptor', in *Twenty Years of G-CSF: Clinical and Nonclinical Discoveries*. doi: 10.1007/978-3-0348-0218-5_5.

Aubin, Y., Hodgson, D.J., Thach, W.B., Gingras, G. and Sauvé, S., 2015. Monitoring effects of excipients, formulation parameters and mutations on the high order structure of filgrastim by NMR. *Pharmaceutical research*, *32*(10), pp.3365-3375.

Baker, F.N. and Porollo, A., 2016. CoeViz: a web-based tool for coevolution analysis of protein residues. *BMC bioinformatics*, *17*(1), pp.1-7.

Baldwin, R. L. (1986) 'Temperature dependence of the hydrophobic interaction in protein folding.', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.83.21.8069.

Barata, T.S., Zhang, C., Dalby, P.A., Brocchini, S. and Zloh, M., 2016. Identification of protein–excipient interaction hotspots using computational approaches. *International journal of molecular sciences*, *17*(6), p.853.

Bellissent-Funel, M. C. *et al.* (2016) 'Water Determines the Structure and Dynamics of Proteins', *Chemical Reviews*. doi: 10.1021/acs.chemrev.5b00664.

Berkowitz, S. A. and Houde, D. J. (2014) 'The Complexity of Protein Structure and the

Challenges it Poses in Developing Biopharmaceuticals', in *Biophysical Characterization of Proteins in Developing Biopharmaceuticals*. doi: 10.1016/B978-0-444-59573-7.00001-4.

Bjelošević, M., Pobirk, A.Z., Planinšek, O. and Grabnar, P.A., 2020. Excipients in freeze-dried biopharmaceuticals: Contributions toward formulation stability and lyophilisation cycle optimisation. *International Journal of Pharmaceutics*, *576*, p.119029.

Bouvignies, G., Vallurupalli, P., Cordes, M.H., Hansen, D.F. and Kay, L.E., 2011. Measuring 1HN temperature coefficients in invisible protein states by relaxation dispersion NMR spectroscopy. *Journal of biomolecular NMR*, *50*(1), pp.13-18.

Bristow, A.F., Bird, C., Bolgiano, B. and Thorpe, R., 2012. Regulatory requirements for therapeutic proteins: the relationship between the conformation and biological activity of filgrastim. *Pharmeuropa Bio & Scientific Notes*, *2012*, pp.103-117.

Carpenter, J.F. and Crowe, J.H., 1988. The mechanism of cryoprotection of proteins by solutes. *Cryobiology*, *25*(3), pp.244-255.

Carpenter, J.F., Arakawa, T. and Crowe, J.H., 1992. Interactions of stabilizing additives with proteins during freeze-thawing and freeze-drying. *Developments in biological standardization*, *74*, pp.225-38.

Caulkins, B.G., Cervantes, S.A., Isas, J.M. and Siemer, A.B., 2018. Dynamics of the proline-rich C-terminus of huntingtin exon-1 fibrils. *The Journal of Physical Chemistry B*, *122*(41), pp.9507-9515.

Chakroun, N., Hilton, D., Ahmad, S.S., Platt, G.W. and Dalby, P.A., 2016. Mapping the aggregation kinetics of a therapeutic antibody fragment. *Molecular pharmaceutics*, *13*(2), pp.307-319.

Chang, L.L., Shepherd, D., Sun, J., Ouellette, D., Grant, K.L., Tang, X.C. and Pikal, M.J., 2005. Mechanism of protein stabilization by sugars during freeze-drying and storage: native structure preservation, specific interaction, and/or immobilization in a glassy matrix?. *Journal of pharmaceutical sciences*, *94*(7), pp.1427-1444.

Chen, K. and Arnold, F. H. (1993) 'Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide.', *Proceedings of the National Academy of Sciences of the United States of America*.

Chen, X., Weber, I. and Harrison, R.W., 2008. Hydration water and bulk water in proteins have distinct properties in radial distributions calculated from 105 atomic resolution crystal structures. *The Journal of Physical Chemistry B*, *112*(38), pp.12073-12080.

Chi, E.Y., Krishnan, S., Kendrick, B.S., Chang, B.S., Carpenter, J.F. and Randolph, T.W., 2003. Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Science*, *12*(5), pp.903-913.

Cierpicki, T. and Otlewski, J., 2001. Amide proton temperature coefficients as hydrogen bond indicators in proteins. *Journal of biomolecular NMR*, *21*(3), pp.249-261.

Codina, N., Hilton, D., Zhang, C., Chakroun, N., Ahmad, S.S., Perkins, S.J. and Dalby, P.A., 2019. An expanded conformation of an antibody Fab region by X-ray scattering, molecular dynamics, and smFRET identifies an aggregation mechanism. *Journal of molecular biology*, *431*(7), pp.1409-1425.

Consalvi, V., Chiaraluce, R., Giangiacomo, L., Scandurra, R., Christova, P., Karshikoff, A., Knapp, S. and Ladenstein, R., 2000. Thermal unfolding and conformational stability of the recombinant domain II of glutamate dehydrogenase from the hyperthermophile Thermotoga maritima. *Protein engineering*, *13*(7), pp.501-507.

Cravens, A., Jamil, O.K., Kong, D., Sockolosky, J.T. and Smolke, C.D., 2021. Polymerase-guided base editing enables in vivo mutagenesis and rapid protein engineering. *Nature communications*, *12*(1), pp.1-12.

Cui, J.Y., Zhang, F., Nierzwicki, L., Palermo, G., Linhardt, R.J. and Lisi, G.P., 2020. Mapping the structural and dynamic determinants of pH-sensitive heparin binding to granulocyte macrophage Colony stimulating factor. *Biochemistry, 59*(38), pp.3541-3553.

Dalby, P. A. (2011) 'Strategy and success for the directed evolution of enzymes', *Current Opinion in Structural Biology*. doi: 10.1016/j.sbi.2011.05.003.

De Dios, A.C., Pearson, J.G. and Oldfield, E., 1993. Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. *Science*, *260*(5113), pp.1491-1496.

Declerck, P. J. (2007) 'Biotherapeutics in the era of biosimilars: What really matters is patient safety', *Drug Safety*. doi: 10.2165/00002018-200730120-00002.

Deechongkit, S., Wen, J., Narhi, L.O., Jiang, Y., Park, S.S., Kim, J. and Kerwin, B.A., 2009. Physical and biophysical effects of polysorbate 20 and 80 on darbepoetin alfa. *Journal of pharmaceutical sciences*, *98*(9), pp.3200-3217.

Ding, N. S., Hart, A. and De Cruz, P. (2016) 'Systematic review: Predicting and optimising response to anti-TNF therapy in Crohn's disease - Algorithm for practical management', *Alimentary Pharmacology and Therapeutics*. doi: 10.1111/apt.13445.

Dinwoodie, N. (2011) 'Biobetters and the Future Biologics Market', *BioPharm International*.

Dobson, J., Kumar, A., Willis, L.F., Tuma, R., Higazi, D.R., Turner, R., Lowe, D.C., Ashcroft, A.E., Radford, S.E., Kapur, N. and Brockwell, D.J., 2017. Inducing protein aggregation by extensional flow. *Proceedings of the National Academy of Sciences*, *114*(18), pp.4673-4678.

Doherty, C.P., Young, L.M., Karamanos, T.K., Smith, H.I., Jackson, M.P., Radford, S.E. and Brockwell, D.J., 2018. A peptide-display protein scaffold to facilitate single molecule force studies of aggregation-prone peptides. *Protein Science*, *27*(7), pp.1205-1217.

Dolinsky, T.J., Czodrowski, P., Li, H., Nielsen, J.E., Jensen, J.H., Klebe, G. and Baker, N.A., 2007. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic acids research*, *35*(suppl_2), pp.W522-W525.

Dombkowski, A. A. (2003) ‘Disulfide by Design^{TM}: A computational method for the rational design of disulfide bonds in proteins’, *Bioinformatics*. doi: 10.1093/bioinformatics/btg231.

Dong, X., Gong, Z., Lu, Y.B., Liu, K., Qin, L.Y., Ran, M.L., Zhang, C.L., Liu, Z., Zhang, W.P. and Tang, C., 2017. Ubiquitin S65 phosphorylation engenders a pH-sensitive conformational switch. *Proceedings of the National Academy of Sciences*, *114*(26), pp.6770-6775.

Du, X., Li, Y., Xia, Y.L., Ai, S.M., Liang, J., Sang, P., Ji, X.L. and Liu, S.Q., 2016. Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, *17*(2), p.144.

Edelhoch, H., 1967. Spectroscopic determination of tryptophan and tyrosine in proteins. *Biochemistry*, *6*(7), pp.1948-1954.

Estrada, J., Bernadó, P., Blackledge, M. and Sancho, J., 2009. ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble. *BMC bioinformatics*, *10*(1), pp.1-8.

Franks, F., 1994. Long–term stabilization of biologicals. *Bio/technology*, *12*(3), pp.253-256.

Frenz, B., Lewis, S.M., King, I., DiMaio, F., Park, H. and Song, Y., 2020. Prediction of protein mutational free energy: benchmark and sampling improvements increase classification accuracy. *Frontiers in bioengineering and biotechnology*, p.1175.

Ghasriani, H., Hodgson, D.J., Brinson, R.G., McEwen, I., Buhse, L.F., Kozlowski, S., Marino, J.P., Aubin, Y. and Keire, D.A., 2016. Precision and robustness of 2D-NMR for structure assessment of filgrastim biosimilars. *Nature biotechnology*, *34*(2), pp.139-141.

Ghasriani, H., Frahm, G.E., Johnston, M.J. and Aubin, Y., 2020. Effects of excipients on the structure and dynamics of filgrastim monitored by thermal unfolding studies by CD and NMR spectroscopy. *ACS omega*, *5*(49), pp.31845-31857.

Gill, S.C. and Von Hippel, P.H., 1989. Calculation of protein extinction coefficients from amino acid sequence data. *Analytical biochemistry*, *182*(2), pp.319-326.

Gokarn, Y.R., Kras, E., Nodgaard, C., Dharmavaram, V., Fesinmeyer, R.M., Hultgen, H., Brych, S., Remmele Jr, R.L., Brems, D.N. and Hershenson, S., 2008. Self-buffering antibody formulations. *Journal of pharmaceutical sciences*, *97*(8), pp.3051-3066.

Goldenzweig, A., Goldsmith, M., Hill, S.E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J. and Lieberman, R.L., 2016. Automated structure-and sequence-based design of proteins for high bacterial expression and stability. *Molecular cell*, *63*(2), pp.337-346.

Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A. and Caves, L.S., 2006. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, *22*(21), pp.2695-2696.

Grant, Y., Matejtschuk, P., Bird, C., Wadhwa, M. and Dalby, P.A., 2012. Freeze drying formulation using microscale and design of experiment approaches: a case study using granulocyte colony-stimulating factor. *Biotechnology letters*, *34*(4), pp.641-648.

Grant, B.J., Skjærven, L. and Yao, X.Q., 2021. The Bio3D packages for structural bioinformatics. *Protein Science*, *30*(1), pp.20-30.

Gribenko, A.V., Patel, M.M., Liu, J., McCallum, S.A., Wang, C. and Makhatadze, G.I., 2009. Rational stabilization of enzymes by computational redesign of surface charge–charge interactions. *Proceedings of the National Academy of Sciences*, *106*(8), pp.2601-2606.

Heisel, K.A. and Krishnan, V.V., 2014. NMR based solvent exchange experiments to understand the conformational preference of intrinsically disordered proteins using FG-nucleoporin peptide as a model. *Peptide Science*, *102*(1), pp.69-77.

Herman, A.C., Boone, T.C. and Lu, H.S., 2002. Characterization, formulation, and stability of Neupogen®(Filgrastim), a recombinant human granulocyte-colony stimulating factor. *Formulation, characterization, and stability of protein drugs: case histories*, pp.303-328.

Holmgren, A., 1989. Thioredoxin and glutaredoxin systems. *Journal of Biological Chemistry*, *264*(24), pp.13963-13966.

Houde, D. J. and Berkowitz, S. A. (2014) 'Biopharmaceutical Industry's Biophysical Toolbox', in *Biophysical Characterization of Proteins in Developing Biopharmaceuticals*. doi: 10.1016/B978-0-444-59573-7.00003-8.

Hsu, K.C., Chen, Y.F., Lin, S.R. and Yang, J.M., 2011. iGEMDOCK: a graphical environment of enhancing GEMDOCK using pharmacological interactions and post-screening analysis. *BMC bioinformatics*, *12*(1), pp.1-11.

Ichiye, T. and Karplus, M. (1991) 'Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations', *Proteins: Structure, Function, and Bioinformatics*. doi: 10.1002/prot.340110305.

Igumenova, T.I., Frederick, K.K. and Wand, A.J., 2006. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chemical reviews*, *106*(5), pp.1672-1699.

Irudayam, S.J. and Henchman, R.H., 2009. Entropic cost of protein− ligand binding and its dependence on the entropy in solution. *The Journal of Physical Chemistry B*, *113*(17), pp.5871-5884.

Israelachvili, J. (2011) *Intermolecular and Surface Forces*, *Intermolecular and Surface Forces*. doi: 10.1016/C2009-0-21560-1.

Jain, T., Sun, T., Durand, S., Hall, A., Houston, N.R., Nett, J.H., Sharkey, B., Bobrowicz, B., Caffry, I., Yu, Y. and Cao, Y., 2017. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, *114*(5), pp.944-949.

Jemal, A. and Bray, F., 2011. Center MM, Ferlay J, Ward E, et al (2011) Global cancer statistics. *Ca Cancer J Clin*, *61*(2), pp.69-90.

Jiang, Y., Jiang, W., Qiu, Y. and Dai, W., 2011. Effect of a structurally modified human granulocyte colony stimulating factor, G-CSFa, on leukopenia in mice and monkeys. *Journal of hematology & oncology*, *4*(1), pp.1-8.

Jorgensen, L., Hostrup, S., Moeller, E.H. and Grohganz, H., 2009. Recent trends in stabilising peptides and proteins in pharmaceutical formulation–considerations in the choice of excipients. *Expert opinion on drug delivery*, *6*(11), pp.1219-1230.

Kamerzell, T.J., Esfandiary, R., Joshi, S.B., Middaugh, C.R. and Volkin, D.B., 2011. Protein–excipient interactions: Mechanisms and biophysical characterization applied to protein formulation development. *Advanced drug delivery reviews*, *63*(13), pp.1118-1159.

Kellogg, E. H., Leaver-Fay, A. and Baker, D. (2011) 'Role of conformational sampling in computing mutation-induced changes in protein structure and stability', *Proteins: Structure, Function and Bioinformatics*. doi: 10.1002/prot.22921.

Keskin, O., Gursoy, A., Ma, B. and Nussinov, R., 2008. Principles of protein− protein interactions: what are the preferred ways for proteins to interact?. *Chemical reviews*, *108*(4), pp.1225-1244.

Kheddo, P., Tracka, M., Armer, J., Dearman, R.J., Uddin, S., Van Der Walle, C.F. and Golovanov, A.P., 2014. The effect of arginine glutamate on the stability of monoclonal antibodies in solution. *International journal of pharmaceutics*, *473*(1-2), pp.126-133.

Kheddo, P., Cliff, M.J., Uddin, S., van der Walle, C.F. and Golovanov, A.P., 2016, October. Characterizing monoclonal antibody formulations in arginine glutamate solutions using 1H NMR spectroscopy. In *Mabs* (Vol. 8, No. 7, pp. 1245-1258). Taylor & Francis.

Kim, S.J., Lee, J.A., Joo, J.C., Yoo, Y.J., Kim, Y.H. and Song, B.K., 2010. The development of a thermostable CiP (Coprinus cinereus peroxidase) through in silico design. *Biotechnology Progress*, *26*(4), pp.1038-1046.

Kleckner, I.R. and Foster, M.P., 2011. An introduction to NMR-based approaches for measuring protein dynamics. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1814*(8), pp.942-968.

Ko, S.K., Berner, C., Kulakova, A., Schneider, M., Antes, I., Winter, G., Harris, P. and Peters, G.H., 2022. Investigation of the pH-dependent aggregation mechanisms of GCSF using low resolution protein characterization techniques and advanced molecular dynamics simulations. *Computational and Structural Biotechnology Journal*.

Konski, A. F. (2011) 'Generic biologics: A comparative analysis of regulatory review', *BioProcess International*.

Krishnan, S., Chi, E.Y., Webb, J.N., Chang, B.S., Shan, D., Goldenberg, M., Manning, M.C., Randolph, T.W. and Carpenter, J.F., 2002. Aggregation of granulocyte colony stimulating factor under physiological conditions: characterization and thermodynamic inhibition. *Biochemistry*, *41*(20), pp.6422-6431.

Kumar, R. (2009) 'Role of naturally occurring osmolytes in protein folding and stability', *Archives of Biochemistry and Biophysics*. doi: 10.1016/j.abb.2009.09.007.

Lakomek, N.A., Lange, O.F., Walter, K.F., Farès, C., Egger, D., Lunkenheimer, P., Meiler, J., Grubmüller, H., Becker, S., de Groot, B.L. and Griesinger, C., 2008. Residual dipolar couplings as a tool to study molecular recognition of ubiquitin. *Biochemical Society Transactions*, *36*(6), pp.1433-1437.

Lehmann, M. and Wyss, M. (2001) 'Engineering proteins for thermostability: The use of sequence alignments versus rational design and directed evolution', *Current Opinion in Biotechnology*. doi: 10.1016/S0958-1669(00)00229-9.

Levitt, M.H., 2013. *Spin dynamics: basics of nuclear magnetic resonance*. John Wiley & Sons.

Levy, Y. and Onuchic, J.N., 2004. Water and proteins: A love–hate relationship. *Proceedings of the National Academy of Sciences*, *101*(10), pp.3325-3326.

Liao, H., Yeh, W., Chiang, D., Jernigan, R.L. and Lustig, B., 2005. Protein sequence entropy is closely related to packing density and hydrophobicity. *Protein Engineering Design and Selection*, *18*(2), pp.59-64.

Lichtarge, O., Bourne, H.R. and Cohen, F.E., 1996. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, *257*(2), pp.342-358.

Lipiäinen, T., Peltoniemi, M., Sarkhel, S., Yrjönen, T., Vuorela, H., Urtti, A. and Juppo, A., 2015. Formulation and stability of cytokine therapeutics. *Journal of pharmaceutical sciences*, *104*(2), pp.307-326.

Liu, F., Poursine-Laurent, J. and Link, D. C. D. (2000) 'Expression of the G-CSF receptor on hematopoietic progenitor cells is not required for their mobilization by G-CSF', *Blood*.

Liu, Y. and Kuhlman, B. (2006) 'RosettaDesign server for protein design', *Nucleic Acids Research*. doi: 10.1093/nar/gkl163.

Liu, Y., Zhao, Y. and Feng, X. (2008) 'Exergy analysis for a freeze-drying process', *Applied Thermal Engineering*. doi: 10.1016/j.applthermaleng.2007.06.004.

Liu, B. and Zhou, X. (2015) 'Freeze-drying of proteins', *Methods in Molecular Biology*. doi: 10.1007/978-1-4939-2193-5_23.

Llinas, M., Llinas, E.S.D. and Klein, M.P., 1974. Charge relay at the peptide bond: A proton magnetic resonance study of solvation effects on the amide electron density distribution.

Luo, P., Hayes, R.J., Chan, C., Stark, D.M., Hwang, M.Y., Jacinto, J.M., Juvvadi, P., Chung, H.S., Kundu, A., Ary, M.L. and Dahiyat, B.I., 2002. Development of a cytokine analog with enhanced stability using computational ultrahigh throughput screening. *Protein Science*, *11*(5), pp.1218-1226.

Ma, B., Elkayam, T., Wolfson, H. and Nussinov, R., 2003. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences*, *100*(10), pp.5772-5777.

Machado, M.R., Barrera, E.E., Klein, F., Sóñora, M., Silva, S. and Pantano, S., 2019. The SIRAH 2.0 force field: altius, fortius, citius. *Journal of Chemical Theory and Computation*, *15*(4), pp.2719-2733.

Mahler, H.C., Friess, W., Grauschopf, U. and Kiese, S., 2009. Protein aggregation: pathways, induction factors and analysis. *Journal of pharmaceutical sciences*, *98*(9), pp.2909-2934.

Majumdar, A.B., Kim, I.J. and Na, H., 2020. Effect of solvent on protein structure and dynamics. *Physical Biology*, *17*(3), p.036006.

Manning, M.C., Chou, D.K., Murphy, B.M., Payne, R.W. and Katayama, D.S., 2010. Stability of protein pharmaceuticals: an update. *Pharmaceutical research*, *27*(4), pp.544-575.

Massa, A., Perut, F., Chano, T., Woloszyk, A., Mitsiadis, T.A., Avnet, S. and Baldini, N., 2017. The effect of extracellular acidosis on the behaviour of mesenchymal stem cells in vitro. *European Cells and Materials (ECM)*, *33*, pp.252-267.

Mitragotri, S., Burke, P. A. and Langer, R. (2014) 'Overcoming the challenges in administering biopharmaceuticals: Formulation and delivery strategies', *Nature Reviews Drug Discovery*. doi: 10.1038/nrd4363.

Miyazaki, K. and Arnold, F. H. (1999) 'Exploring nonnatural evolutionary pathways by saturation mutagenesis: Rapid improvement of protein function', *Journal of Molecular Evolution*. doi: 10.1007/PL00006593.

Molineux, G. (2004) 'The Design and Development of Pegfilgrastim (PEG-rmetHuG-CSF, Neulasta&#174;)', *Current Pharmaceutical Design*. doi: 10.2174/1381612043452613.

Motojima, H., Kobayashi, T., Shimane, M., Kamachi, S.I. and Fukushima, M., 1989. Quantitative enzyme immunoassay for human granulocyte colony stimulating factor (G-CSF). *Journal of immunological methods*, *118*(2), pp.187-192.

Newberry, R.W. and Raines, R.T., 2019. Secondary forces in protein folding. *ACS chemical biology*, *14*(8), pp.1677-1686.

Nikolaeva, L.P., 2018. Features of acid–base balance of bone marrow. *Acta Medica International*, *5*(2), p.55.

Nilsson, M. R., Driscoll, M. and Raleigh, D. P. (2009) 'Low levels of asparagine deamidation can have a dramatic effect on aggregation of amyloidogenic peptides: Implications for the study of amyloid formation', *Protein Science*. doi: 10.1110/ps.48702.

Olivera, P., Danese, S. and Peyrin-Biroulet, L. (2017) 'Next generation of small molecules in inflammatory bowel disease', *Gut*. doi: 10.1136/gutjnl-2016-312912.

Park, H., Bradley, P., Greisen Jr, P., Liu, Y., Mulligan, V.K., Kim, D.E., Baker, D. and DiMaio, F., 2016. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of chemical theory and computation, 12*(12), pp.6201-6212.

Palmer III, A.G., 1997. Probing molecular motion by NMR. *Current opinion in structural biology, 7*(5), pp.732-737.

Panchenko, A. R., Luthey-Schulten, Z. and Wolynes, P. G. (1996) 'Foldons, protein structural modules, and exons.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 93(5), pp. 2008–13. doi: 10.1073/PNAS.93.5.2008.

Panick, G., Malessa, R., Winter, R., Rapp, G., Frye, K.J. and Royer, C.A., 1998. Structural characterization of the pressure-denatured state and unfolding/refolding kinetics of staphylococcal nuclease by synchrotron small-angle X-ray scattering and Fourier-transform infrared spectroscopy. *Journal of molecular biology, 275*(2), pp.389-402.

Panopoulos, A. D. and Watowich, S. S. (2008) 'Granulocyte colony-stimulating factor: Molecular mechanisms of action during steady state and "emergency" hematopoiesis', *Cytokine*. doi: 10.1016/j.cyto.2008.03.002.

Papaleo, E., Mereghetti, P., Fantucci, P., Grandori, R. and De Gioia, L., 2009. Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: the myoglobin case. *Journal of molecular graphics and modelling, 27*(8), pp.889-899.

Parera, M. and Martinez, M. A. (2014) 'Strong epistatic interactions within a single protein', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msu113.

Parthasarathy, S. and Murthy, M. R. N. (2002) 'Protein thermal stability: insights from atomic displacement parameters (B values)', *Protein Engineering, Design and Selection*. doi: 10.1093/protein/13.1.9.

Parui, S. and Jana, B., 2019. Factors promoting the formation of clathrate-like ordering of water in biomolecular structure at ambient temperature and pressure. *The Journal of Physical Chemistry B*, *123*(4), pp.811-824.

Philo, J. and Arakawa, T. (2009) 'Mechanisms of Protein Aggregation', *Current Pharmaceutical Biotechnology*. doi: 10.2174/138920109788488932.

Pikal-Cleland, K.A., Cleland, J.L., Anchordoquy, T.J. and Carpenter, J.F., 2002. Effect of glycine on pH changes and protein stability during freeze–thawing in phosphate buffer systems. *Journal of pharmaceutical sciences*, *91*(9), pp.1969-1979.

Pikal, M.J., 2004. Mechanisms of protein stabilization during freeze-drying and storage: The relative importance of thermodynamic stabilization and glassy state relaxation dynamic. *Freeze-drying/lyophilization of pharmaceutical and biological products*.

Poelwijk, F.J., Socolich, M. and Ranganathan, R., 2019. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications*, *10*(1), pp.1-11.

Porasuphatana, S., Weaver, J., Budzichowski, T.A., Tsai, P. and Rosen, G.M., 2001. Differential effect of buffer on the spin trapping of nitric oxide by iron chelates. *Analytical Biochemistry*, *298*(1), pp.50-56.

Posfai, A., Zhou, J., Plotkin, J.B., Kinney, J.B. and McCandlish, D.M., 2018. Selection for protein stability enriches for epistatic interactions. *Genes*, *9*(9), p.423.

Ptitsyn, O.B. and Ting, K.L.H., 1999. Non-functional conserved residues in globins and their possible role as a folding nucleus. *Journal of molecular biology*, *291*(3), pp.671-682.

Ramírez-Sarmiento, C.A., Baez, M., Wilson, C.A., Babul, J., Komives, E.A. and Guixé, V., 2013. Observation of solvent penetration during cold denaturation of E. coli phosphofructokinase-2. *Biophysical journal*, *104*(10), pp.2254-2263.

Raso, S.W., Abel, J., Barnes, J.M., Maloney, K.M., Pipes, G., Treuheit, M.J., King, J. and Brems, D.N., 2005. Aggregation of granulocyte-colony stimulating factor in vitro involves a conformationally altered monomeric state. *Protein science*, *14*(9), pp.2246-2257.

Reetz, M. T. (2013) 'The importance of additive and non-additive mutational effects in protein engineering', *Angewandte Chemie - International Edition*. doi: 10.1002/anie.201207842.

Reetz, M. T. and Carballeira, J. D. (2007) 'Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes.', *Nature protocols*. doi: 10.1038/nprot.2007.72.

Richardson, J.S., 1989. DC Richardson in: Prediction of Protein Structure and the Principles of Protein Conformation. *GD Fasman, Ed*, pp.1-99.

Ringe, D. and Petsko, G. A. (2003) 'The "glass transition" in protein dynamics: What it is, why it occurs, and how to exploit it', *Biophysical Chemistry*. doi: 10.1016/S0301-4622(03)00096-6.

Roberts, A.W., Foote, S., Alexander, W.S., Scott, C., Robb, L. and Metcalf, D., 1997. Genetic influences determining progenitor cell mobilization and leukocytosis induced by granulocyte colony-stimulating factor. *Blood, The Journal of the American Society of Hematology*, *89*(8), pp.2736-2744.

Roberts, B.C. and Mancera, R.L., 2008. Ligand− protein docking with water molecules. *Journal of chemical information and modeling*, *48*(2), pp.397-408.

Roberts, C.J., Das, T.K. and Sahin, E., 2011. Predicting solution aggregation rates for therapeutic proteins: approaches and challenges. *International journal of pharmaceutics*, *418*(2), pp.318-333.

Robinson, M.J., Matejtschuk, P., Bristow, A.F. and Dalby, P.A., 2018. T m-values and unfolded fraction can predict aggregation rates for granulocyte colony stimulating factor variant formulations but not under predominantly native conditions. *Molecular pharmaceutics*, *15*(1), pp.256-267.

Saito, S., Hasegawa, J., Kobayashi, N., Kishi, N., Uchiyama, S. and Fukui, K., 2012. Behavior of monoclonal antibodies: relation between the second virial coefficient (B 2) at low concentrations and aggregation propensity and viscosity at high concentrations. *Pharmaceutical research*, *29*(2), pp.397-410.

Salvi, G., De Los Rios, P. and Vendruscolo, M., 2005. Effective interactions between chaotropic agents and proteins. *Proteins: Structure, Function, and Bioinformatics*, *61*(3), pp.492-499.

Sandeep, V., Parveen, J. and Chauhan, P. (2016) 'Biobetters: the better biologics and their regulatory overview', *International Journal of Drug Regulatory Affairs*.

Schmidt, B., Ho, L. and Hogg, P.J., 2006. Allosteric disulfide bonds. *Biochemistry*, *45*(24), pp.7429-7433.

Schmidt, B. and Hogg, P.J., 2007. Search for allosteric disulfide bonds in NMR structures. *BMC Structural Biology*, *7*(1), pp.1-12.

Semerad, C.L., Liu, F., Gregory, A.D., Stumpf, K. and Link, D.C., 2002. G-CSF is an essential regulator of neutrophil trafficking from the bone marrow to the blood. *Immunity*, *17*(4), pp.413-423.

Sethi, A., Eargle, J., Black, A.A. and Luthey-Schulten, Z., 2009. Dynamical networks in tRNA: protein complexes. *Proceedings of the National Academy of Sciences*, *106*(16), pp.6620-6625.

Shaoxiong Wu, 2011. *1D and 2D NMR Experiment Methods*, Chemistry Department Emory University 1515 Pierce Drive Atlanta, GA 30322: .

Shoemaker, B.A., Portman, J.J. and Wolynes, P.G., 2000. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proceedings of the National Academy of Sciences*, *97*(16), pp.8868-8873.

Sinha, S. and Wang, S.M., 2020. Classification of VUS and unclassified variants in BRCA1 BRCT repeats by molecular dynamics simulation. *Computational and structural biotechnology journal*, *18*, pp.723-736.

Sohrabi, C., Foster, A. and Tavassoli, A., 2020. Methods for generating and screening libraries of genetically encoded cyclic peptides in drug discovery. *Nature Reviews Chemistry*, *4*(2), pp.90-101.

Souza, L.M., Boone, T.C., Gabrilove, J., Lai, P.H., Zsebo, K.M., Murdock, D.C., Chazin, V.R., Bruszewski, J., Lu, H., Chen, K.K. and Barendt, J., 1986. Recombinant human granulocyte colony-stimulating factor: effects on normal and leukemic myeloid cells. *Science*, *232*(4746), pp.61-65.

Spassov, V. Z., Karshikoff, A. D. and Ladenstein, R. (1994) ‘Optimization of the electrostatic interactions in proteins of different functional and folding type’, *Protein Science*. doi: 10.1002/pro.5560030921.

Spencer, J.A., Ferraro, F., Roussakis, E., Klein, A., Wu, J., Runnels, J.M., Zaher, W., Mortensen, L.J., Alt, C., Turcotte, R. and Yusuf, R., 2014. Direct measurement of local oxygen concentration in the bone marrow of live animals. *Nature*, *508*(7495), pp.269-273.

Starr, T. N. and Thornton, J. W. (2016) ‘Epistasis in protein evolution’, *Protein Science*. doi: 10.1002/pro.2897.

Stärtzel, P., 2018. Arginine as an excipient for protein freeze-drying: a mini review. *Journal of pharmaceutical sciences*, *107*(4), pp.960-967.

Stemmer, W. P. C. (1994) ‘Rapid evolution of a protein in vitro by DNA shuffling’, *Nature*. doi: 10.1038/370389a0.

Storey, K.B. and Storey, J.M., 1991. Biochemistry of cryoprotectants. In *Insects at low temperature* (pp. 64-93). Springer, Boston, MA.

Strambini, G. B. and Gabellieri, E. (1996) 'Proteins in frozen solutions: Evidence of ice-induced partial unfolding', *Biophysical Journal*. doi: 10.1016/S0006-3495(96)79640-6.

Strohl, W. R. (2015) 'Fusion Proteins for Half-Life Extension of Biologics as a Strategy to Make Biobetters', *BioDrugs*. doi: 10.1007/s40259-015-0133-6.

Tamada, T., Honjo, E., Maeda, Y., Okamoto, T., Ishibashi, M., Tokunaga, M. and Kuroki, R., 2006. Homodimeric cross-over structure of the human granulocyte colony-stimulating factor (GCSF) receptor signaling complex. *Proceedings of the national academy of sciences*, *103*(9), pp.3135-3140.

Tang, X. and Pikal, M. J. (2004) 'Design of Freeze-Drying Processes for Pharmaceuticals: Practical Advice', *Pharmaceutical Research*. doi: 10.1023/B:PHAM.0000016234.73023.75.

Thiagarajan, G., Semple, A., James, J.K., Cheung, J.K. and Shameem, M., 2016, August. A comparison of biophysical characterization techniques in predicting monoclonal antibody stability. In *MAbs* (Vol. 8, No. 6, pp. 1088-1097). Taylor & Francis.

Timasheff, S.N., 2002. Protein-solvent preferential interactions, protein hydration, and the modulation of biochemical reactions by solvent components. *Proceedings of the National Academy of Sciences*, *99*(15), pp.9721-9726.

Tomlinson, J.H. and Williamson, M.P., 2012. Amide temperature coefficients in the protein G B1 domain. *Journal of biomolecular NMR*, *52*(1), pp.57-64.

Topp, E.M., Zhang, L., Zhao, H., Payne, R.W., Evans, G.J. and Manning, M.C., 2010. Chemical instability in peptide and protein pharmaceuticals. *Formulation and process development strategies for manufacturing biopharmaceuticals*, *2*, pp.41-67.

Tosstorff, A., Svilenov, H., Peters, G.H., Harris, P. and Winter, G., 2019. Structure-based discovery of a new protein-aggregation breaking excipient. *European Journal of Pharmaceutics and Biopharmaceutics*, *144*, pp.207-216.

Trainor, K., Palumbo, J.A., MacKenzie, D.W. and Meiering, E.M., 2020. Temperature dependence of NMR chemical shifts: Tracking and statistical analysis. *Protein Science*, *29*(1), pp.306-314.

Tsolis, A.C., Papandreou, N.C., Iconomidou, V.A., Hamodrakas, S.J., 2013. A Consensus Method for the Prediction of "Aggregation-Prone" Peptides in Globular Proteins. PLoS ONE, 8(1): e54175.

Tsuruta, T., Ishimoto, Y. and Masuoka, T. (1998) 'Effects of glycerol on intracellular ice formation and dehydration of onion epidermis', in *Annals of the New York Academy of Sciences*. doi: 10.1111/j.1749-6632.1998.tb10155.x.

Uversky, V.N., Yamin, G., Souillac, P.O., Goers, J., Glaser, C.B. and Fink, A.L., 2002. Methionine oxidation inhibits fibrillation of human α-synuclein in vitro. *FEBS letters*, *517*(1-3), pp.239-244.

Viles, J.H., Duggan, B.M., Zaborowski, E., Schwarzinger, S., Huntley, J.J., Kroon, G.J., Dyson, H.J. and Wright, P.E., 2001. Potential bias in NMR relaxation data introduced by peak intensity analysis and curve fitting methods. *Journal of biomolecular NMR*, *21*(1), pp.1-9.

Volkin, D. B. and Klibanov, A. M. (1987) 'Thermal destruction processes in proteins involving cystine residues.', *Journal of Biological Chemistry*. doi: 10.1111/j.1748-121X.2003.tb00230.x.

Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J. and Laue, E.D., 2005. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins: Structure, Function, and Bioinformatics*, *59*(4), pp.687-696.

Wadhwa, M., Bird, C., Hamill, M., Heath, A.B., Matejtschuk, P. and Thorpe, R., 2011. The 2nd International Standard for human granulocyte colony stimulating factor. Journal of immunological methods, 367(1-2), pp.63-69.

Walsh, I., Seno, F., Tosatto, S.C. and Trovato, A., 2014. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic acids research*, *42*(W1), pp.W301-W307.

Wang, P. L., Udeani, G. O. and Johnston, T. P. (1995) 'Inhibition of granulocyte colony stimulating factor (G-CSF) adsorption to polyvinyl chloride using a nonionic surfactant', *International Journal of Pharmaceutics*. doi: 10.1016/0378-5173(94)00236-X.

Wang, L. and Markley, J.L., 2009. Empirical correlation between protein backbone 15N and 13C secondary chemical shifts and its application to nitrogen chemical shift re-referencing. *Journal of biomolecular NMR*, *44*(2), pp.95-99.

Weinstein, Y., Ihle, J.N., Lavu, S. and Reddy, E.P., 1986. Truncation of the c-myb gene by a retroviral integration in an interleukin 3-dependent myeloid leukemia cell line. *Proceedings of the National Academy of Sciences*, *83*(14), pp.5010-5014.

Whitley, M. and Lee, A. (2009) 'Frameworks for Understanding Long-Range Intra-Protein Communication', *Current Protein & Peptide Science*. doi: 10.2174/138920309787847563.

Willis, L.F., Kumar, A., Dobson, J., Bond, N.J., Lowe, D., Turner, R., Radford, S.E., Kapur, N. and Brockwell, D.J., 2018. Using extensional flow to reveal diverse aggregation landscapes for three IgG1 molecules. *Biotechnology and bioengineering*, *115*(5), pp.1216-1225.

Won, C.M., Molnar, T.E., McKean, R.E. and Spenlehauer, G.A., 1998. Stabilizers against heat-induced aggregation of RPR 114849, an acidic fibroblast growth factor (aFGF). *International journal of pharmaceutics*, *167*(1-2), pp.25-36.

Wood, V.E., Groves, K., Cryar, A., Quaglia, M., Matejtschuk, P. and Dalby, P.A., 2020. HDX and In Silico Docking Reveal that Excipients Stabilize G-CSF via a Combination of Preferential Exclusion and Specific Hotspot Interactions. *Molecular Pharmaceutics*, *17*(12), pp.4637-4651.

Wood, V.E., Groves, K., Wong, L.M., Kong, L., Bird, C., Wadhwa, M., Quaglia, M., Matejtschuk, P. and Dalby, P.A., 2022. Protein Engineering and HDX Identify Structural Regions of G-CSF Critical to Its Stability and Aggregation. *Molecular Pharmaceutics*, *19*(2), pp.616-629.

Woycechowsky, K.J. and Raines, R.T., 2000. Native disulfide bond formation in proteins. *Current opinion in chemical biology*, *4*(5), pp.533-539.

Wu, Z., Kan, S.J., Lewis, R.D., Wittmann, B.J. and Arnold, F.H., 2019. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences*, *116*(18), pp.8852-8858.

Xu, A.Y., Castellanos, M.M., Mattison, K., Krueger, S. and Curtis, J.E., 2019. Studying excipient modulated physical stability and viscosity of monoclonal antibody formulations using small-angle scattering. *Molecular pharmaceutics*, *16*(10), pp.4319-4338.

Yang, J.M. and Shen, T.W., 2005. A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins: Structure, Function, and Bioinformatics*, *59*(2), pp.205-220.

Yu, H. and Dalby, P.A., 2017. Engineer flexible loops for improved enzyme thermostability.

Yu, H. and Dalby, P. A. (2018a) 'Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1810324115.

Yu, H. and Dalby, P. A. (2018b) 'Exploiting correlated molecular-dynamics networks to counteract enzyme activity–stability trade-off', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1812204115.

Yu, H. and Huang, H. (2014a) 'Engineering proteins for thermostability through rigidifying flexible sites', *Biotechnology Advances*. doi: 10.1016/j.biotechadv.2013.10.012.

Yu, L. (2001) 'Amorphous pharmaceutical solids: Preparation, characterization and stabilization', *Advanced Drug Delivery Reviews*. doi: 10.1016/S0169-409X(01)00098-9.

Zalar, M., Svilenov, H.L. and Golovanov, A.P., 2020. Binding of excipients is a poor predictor for aggregation kinetics of biopharmaceutical proteins. *European Journal of Pharmaceutics and Biopharmaceutics*, *151*, pp.127-136.

Zeeb, M. and Balbach, J., 2004. Protein folding studied by real-time NMR spectroscopy. *Methods*, *34*(1), pp.65-74.

Zhang, M.Z., Wen, J., Arakawa, T. and Prestrelski, S.J., 1995. A new strategy for enhancing the stability of lyophilized protein: the effect of the reconstitution medium on keratinocyte growth factor. *Pharmaceutical research*, *12*(10), pp.1447-1452.

Zhang, J., Frey, V., Corcoran, M., Zhang-van Enk, J. and Subramony, J.A., 2016. Influence of arginine salts on the thermal stability and aggregation kinetics of monoclonal antibody: dominant role of anions. *Molecular pharmaceutics*, *13*(10), pp.3362-3369.

Zhang, B., Powers, R. and O′Day, E.M., 2020. Evaluation of Non-Uniform Sampling 2D 1H–13C HSQC Spectra for Semi-Quantitative Metabolomics. *Metabolites*, *10*(5), p.203.

Zhang, H., Yang, Y., Zhang, C., Farid, S.S. and Dalby, P.A., 2021. Machine learning reveals hidden stability code in protein native fluorescence. *Computational and structural biotechnology* *journal*, *19*, pp.2750-2760.

Zhao, H. and Arnold, F. H. (2002) 'Directed evolution converts subtilisin E into a functional equivalent of thermitase', *Protein Engineering, Design and Selection*. doi: 10.1093/protein/12.1.47.

Zink, T., Ross, A., Lueers, K., Cieslar, C., Rudolph, R. and Holak, T.A., 1994. Structure and dynamics of the human granulocyte colony-stimulating factor determined by NMR spectroscopy. Loop mobility in a four-helix-bundle protein. *Biochemistry*, *33*(28), pp.8453-8463.

9 Supplementary Information

Using δ to Calculate ∑Δδ, Peak Linearity, Slope and the Normal Distribution of ∑Δδ

S.1 illustrates the concept of calculating ∑Δδ where, for example, the ∑Δδ at 297 K = Δδ from 295 K to 297 K, and the ∑Δδ at 301 K = (Δδ from 295 K to 297 K) + (Δδ from 297 K to 299 K) + (Δδ from 299 K to 301 K). The example in this figure demonstrates how the ∑Δδ for residue Q90 at 301 K is equal to the sum of the Δδ distances travelled between the peaks for residue Q90 at previous temperatures.

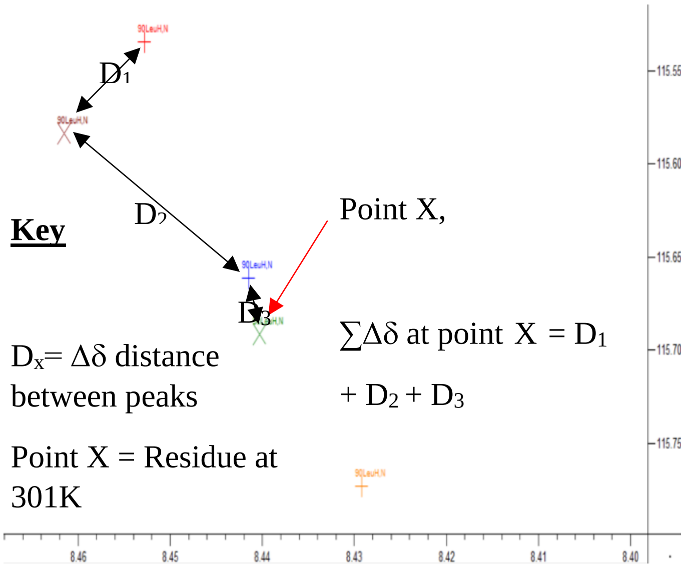


Figure S.1. Deriving ∑Δδ.

The movement of the peak maxima for residue L89 is given here by the different coloured crosses, with red representing its position at 295 K and orange at 305 K. D_x gives the Δδ distance between peaks at consecutive temperatures. ∑Δδ at point X, i.e. the total Δδ distance travelled by residue L89 at 301 K, is given as D_1+ D_2 + D_3.

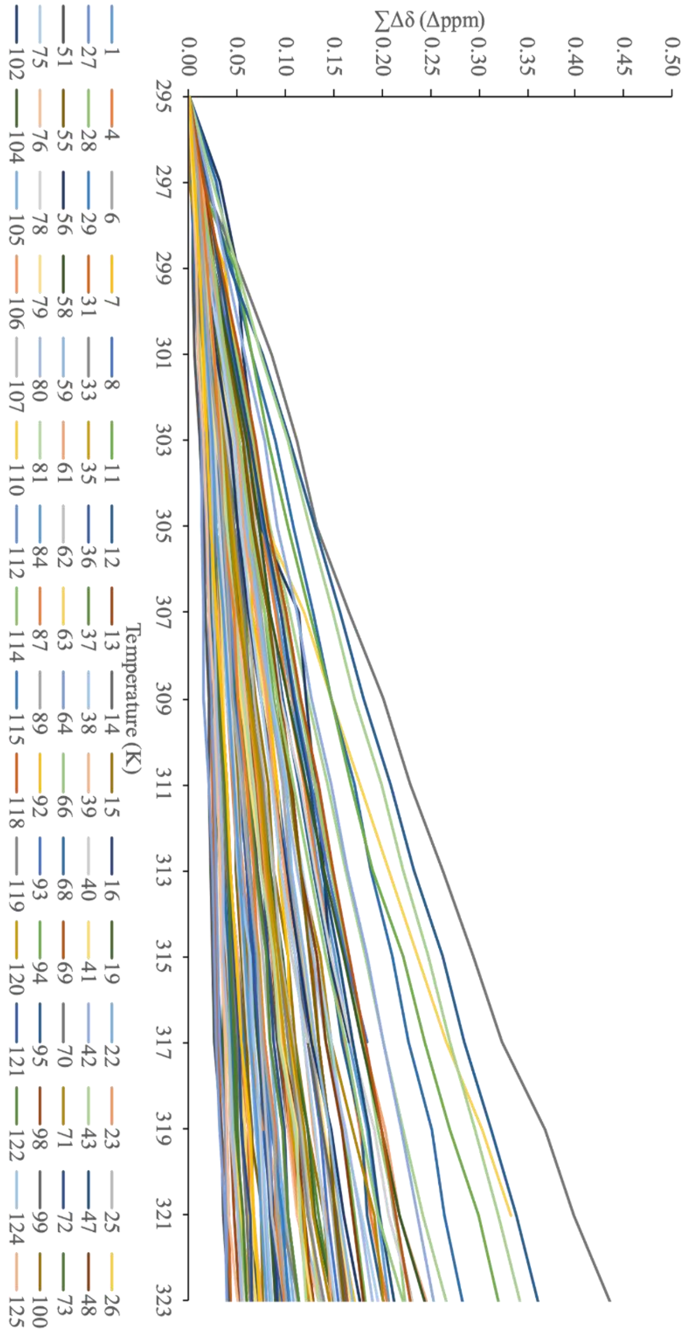
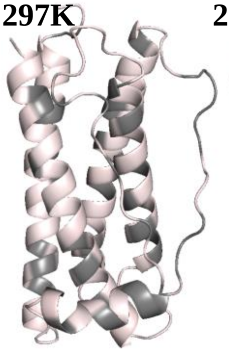
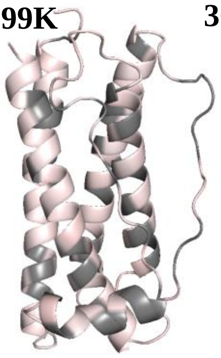


Figure S.2. Temperature dependence of ∑Δδ for all assigned residues. Trends for residues are coloured as shown in the legend. The starting temperature is 295 K and therefore has no cumulative Δδ distance.

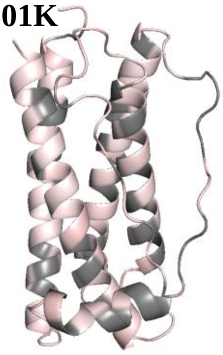
297K



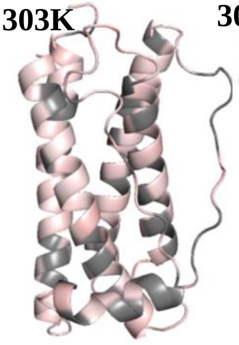
299K



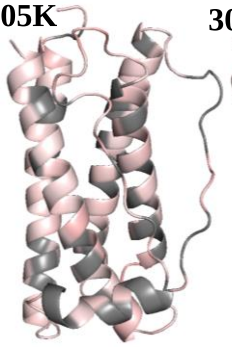
301K



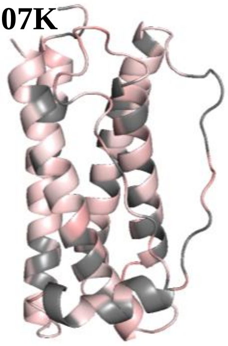
303K



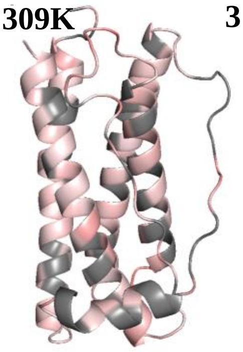
305K



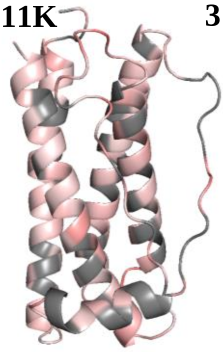
307K



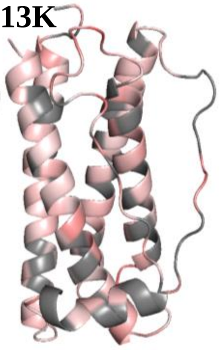
309K



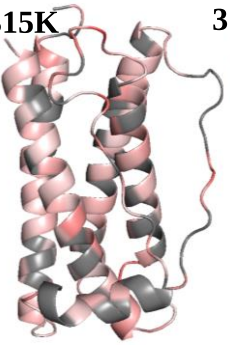
311K



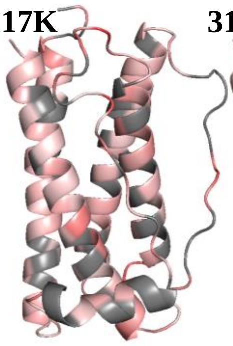
313K



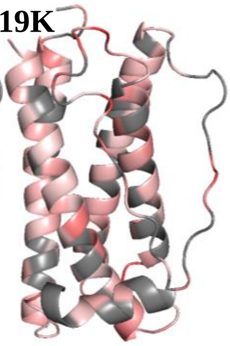
315K



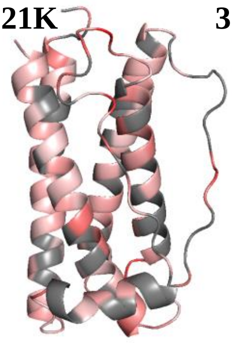
317K



319K



321K



323K

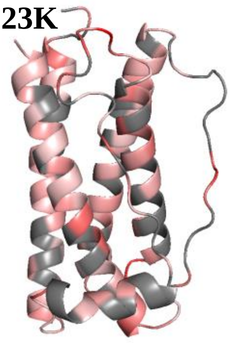


Figure S.3. Mapping ∑Δδ on G-CSF Structure. G-CSF (PDB:2D9Q) coloured according to ∑Δδ at their respective temperature. Red signifies the highest ∑Δδ value (0.435 Δppm), and white signifies the lowest ∑Δδ value (0.000425 Δppm) observed for the whole data set across the thermal melt, while grey represents unassigned residues. Unassigned resides are coloured grey. Residues T1 to A6 are missing in this structure.

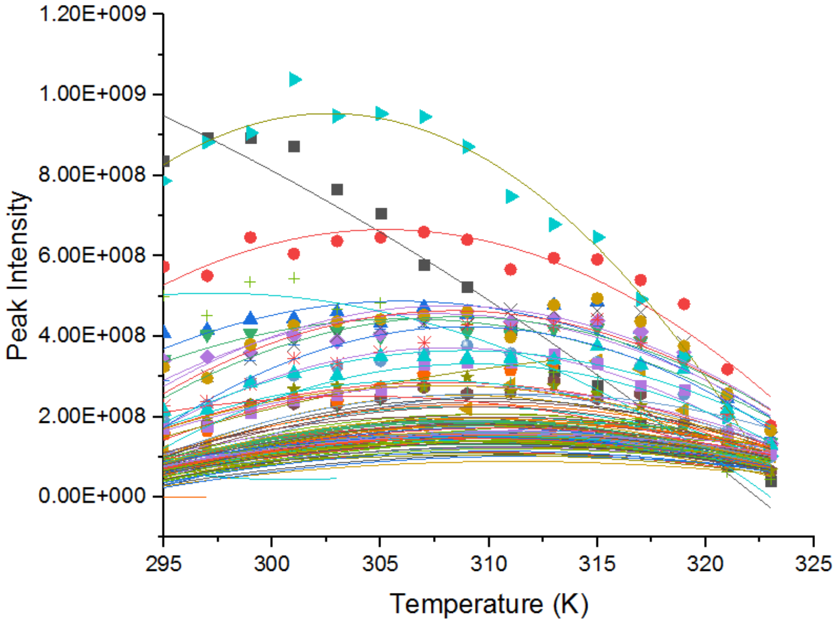
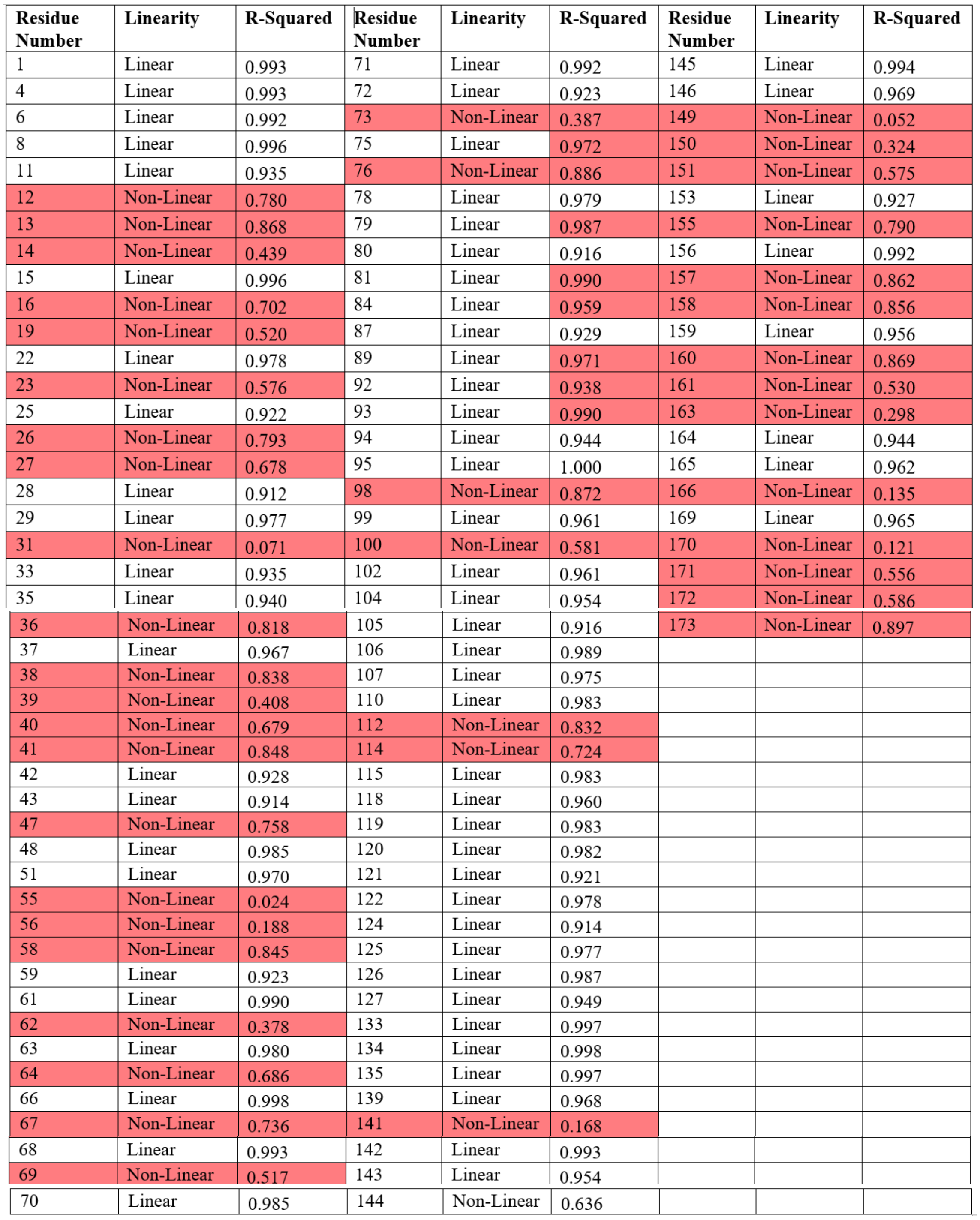


Figure S.4. Raw PI Data with Temperature. Each different coloured line represents a different assigned residue.

Linearity of WT peak trajectories Δδ are given in Table S.1, where 68 trajectories are considered as linear when they have an R^2 value over 0.9 (based on their 1-H and 15-N δ values). The remaining 44 residues for which linearity was calculated are non-linear. Accompanying this table is the structure of G-CSF, highlighting non-linear (red scale), linear (white) and unassigned (grey) residues (Figure S.5A).



Residue Number	Linearity	R-Squared	Residue Number	Linearity	R-Squared	Residue Number	Linearity	R-Squared
1	Linear	0.993	71	Linear	0.992	145	Linear	0.994
4	Linear	0.993	72	Linear	0.923	146	Linear	0.969
6	Linear	0.992	73	Non-Linear	0.387	149	Non-Linear	0.052
8	Linear	0.996	75	Linear	0.972	150	Non-Linear	0.324
11	Linear	0.935	76	Non-Linear	0.886	151	Non-Linear	0.575
12	Non-Linear	0.780	78	Linear	0.979	153	Linear	0.927
13	Non-Linear	0.868	79	Linear	0.987	155	Non-Linear	0.790
14	Non-Linear	0.439	80	Linear	0.916	156	Linear	0.992
15	Linear	0.996	81	Linear	0.990	157	Non-Linear	0.862
16	Non-Linear	0.702	84	Linear	0.959	158	Non-Linear	0.856
19	Non-Linear	0.520	87	Linear	0.929	159	Linear	0.956
22	Linear	0.978	89	Linear	0.971	160	Non-Linear	0.869
23	Non-Linear	0.576	92	Linear	0.938	161	Non-Linear	0.530
25	Linear	0.922	93	Linear	0.990	163	Non-Linear	0.298
26	Non-Linear	0.793	94	Linear	0.944	164	Linear	0.944
27	Non-Linear	0.678	95	Linear	1.000	165	Linear	0.962
28	Linear	0.912	98	Non-Linear	0.872	166	Non-Linear	0.135
29	Linear	0.977	99	Linear	0.961	169	Linear	0.965
31	Non-Linear	0.071	100	Non-Linear	0.581	170	Non-Linear	0.121
33	Linear	0.935	102	Linear	0.961	171	Non-Linear	0.556
35	Linear	0.940	104	Linear	0.954	172	Non-Linear	0.586
36	Non-Linear	0.818	105	Linear	0.916	173	Non-Linear	0.897
37	Linear	0.967	106	Linear	0.989			
38	Non-Linear	0.838	107	Linear	0.975			
39	Non-Linear	0.408	110	Linear	0.983			
40	Non-Linear	0.679	112	Non-Linear	0.832			
41	Non-Linear	0.848	114	Non-Linear	0.724			
42	Linear	0.928	115	Linear	0.983			
43	Linear	0.914	118	Linear	0.960			
47	Non-Linear	0.758	119	Linear	0.983			
48	Linear	0.985	120	Linear	0.982			
51	Linear	0.970	121	Linear	0.921			
55	Non-Linear	0.024	122	Linear	0.978			
56	Non-Linear	0.188	124	Linear	0.914			
58	Non-Linear	0.845	125	Linear	0.977			
59	Linear	0.923	126	Linear	0.987			
61	Linear	0.990	127	Linear	0.949			
62	Non-Linear	0.378	133	Linear	0.997			
63	Linear	0.980	134	Linear	0.998			
64	Non-Linear	0.686	135	Linear	0.997			
66	Linear	0.998	139	Linear	0.968			
67	Non-Linear	0.736	141	Non-Linear	0.168			
68	Linear	0.993	142	Linear	0.993			
69	Non-Linear	0.517	143	Linear	0.954			
70	Linear	0.985	144	Non-Linear	0.636			

Table S.1. Linear trajectories are defined as those with an $R^2 > 0.9$ and are coloured white. Non-linear trajectories are coloured red.

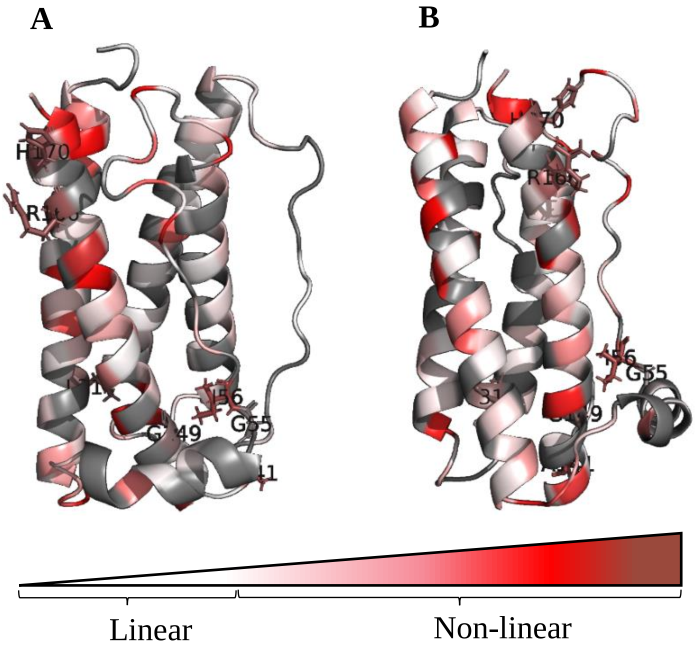
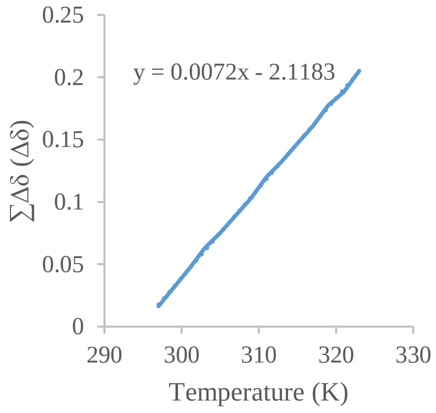


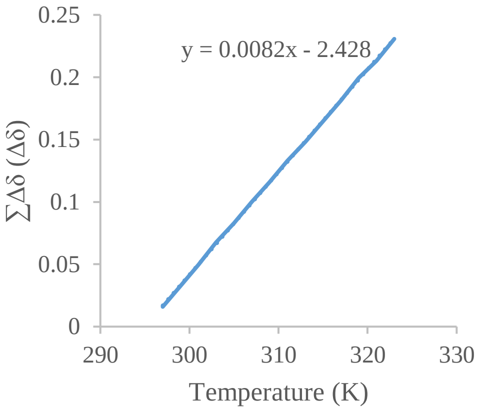
Figure S.5. Determining Peak Trajectory Linearity. Trajectory linearity for residues is coloured on a white to red scale, with darker red representing a lower R^2 value and so more non-linearity, white representing higher linearity. **A)** front (left-hand-side) and **B)** side (right-hand-side) view of PDB:2D9Q. Extremely non-linear residues (R^2 <0.3) are labelled and shown as sticks. Unassigned residues are light grey.

4



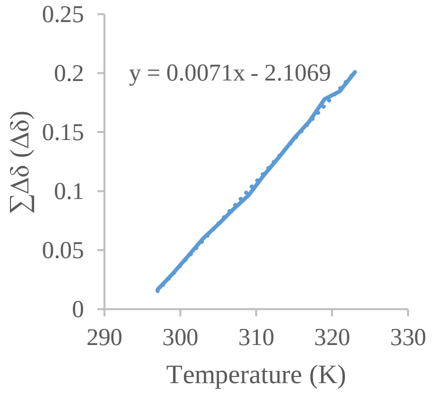
0.25
0.2
0.15
0.1
0.05
0
∑Δδ (Δδ)
y = 0.0072x - 2.1183
290
300
310
320
330
Temperature (K)

6



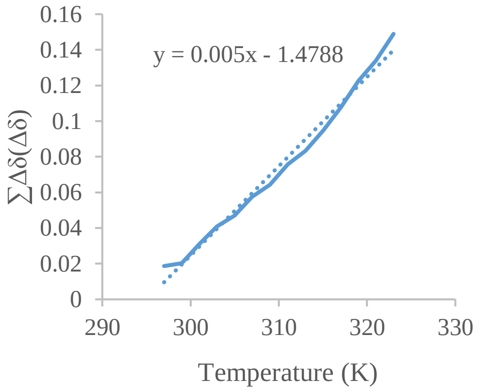
0.25
0.2
0.15
0.1
0.05
0
∑Δδ (Δδ)
y = 0.0082x - 2.428
290
300
310
320
330
Temperature (K)

12



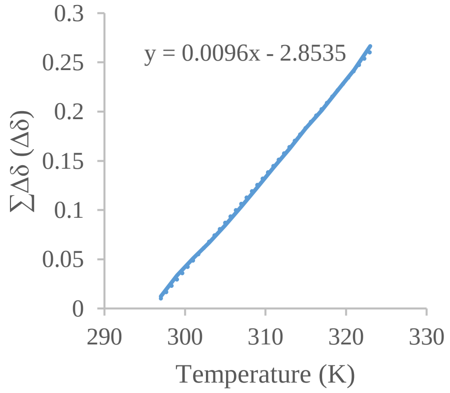
0.25
0.2
0.15
0.1
0.05
0
∑Δδ (Δδ)
y = 0.0071x - 2.1069
290
300
310
320
330
Temperature (K)

37



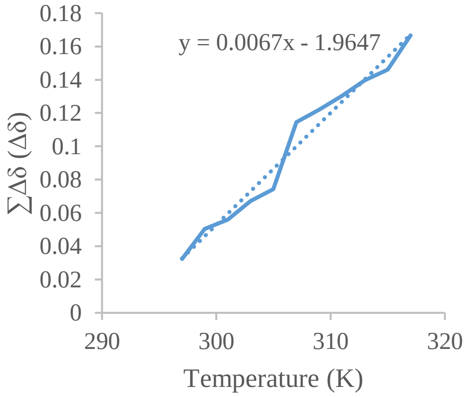
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
∑Δδ(Δδ)
y = 0.005x - 1.4788
290
300
310
320
330
Temperature (K)

43



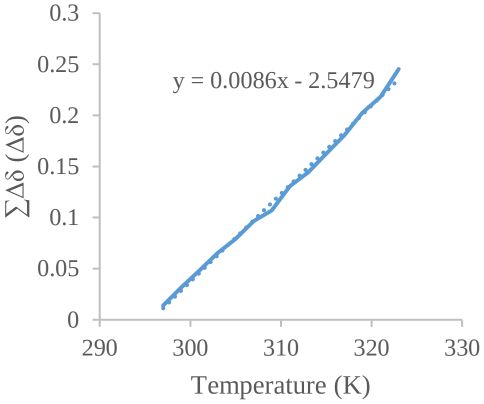
0.3
0.25
0.2
0.15
0.1
0.05
0
∑Δδ (Δδ)
y = 0.0096x - 2.8535
290
300
310
320
330
Temperature (K)

56

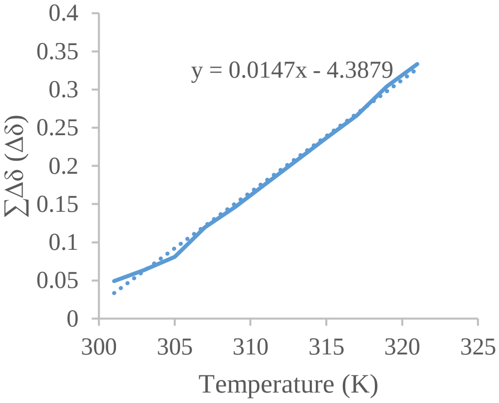


0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
∑Δδ (Δδ)
y = 0.0067x - 1.9647
290
300
310
320
Temperature (K)

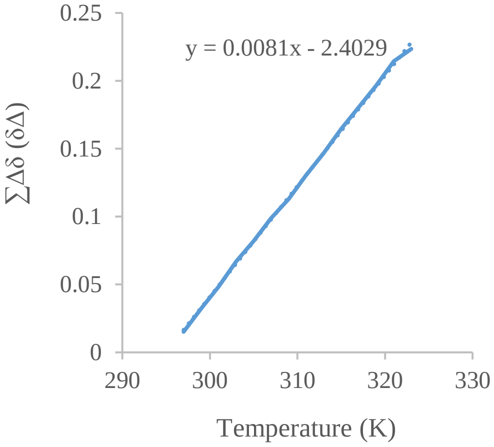
61



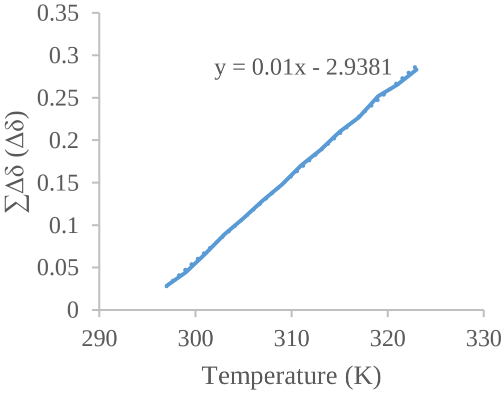
63



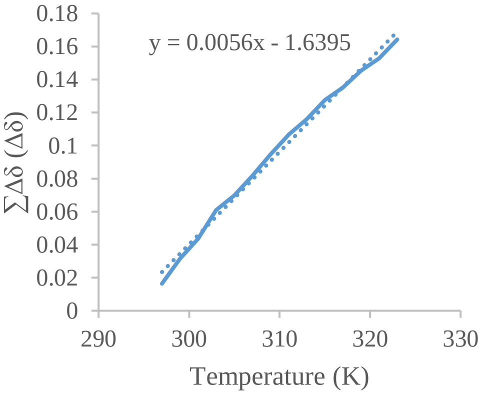
66



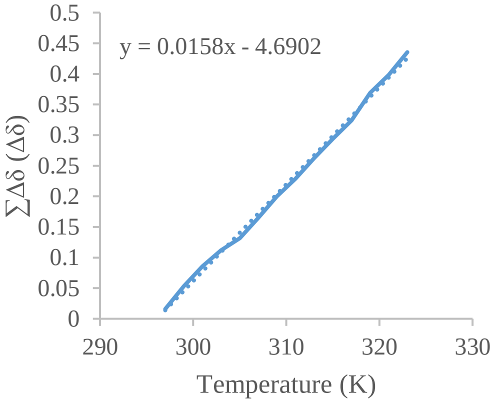
68

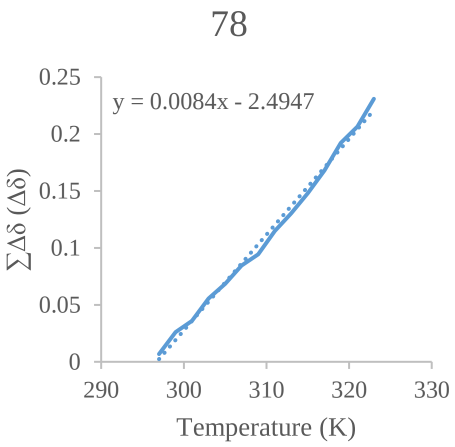


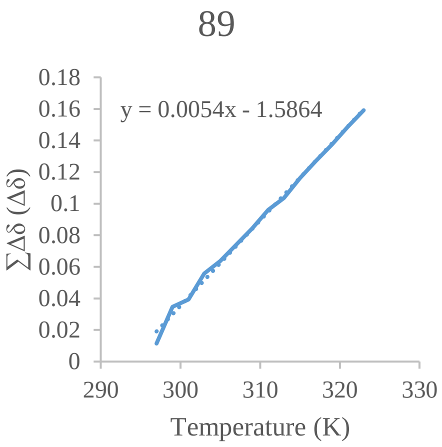
69

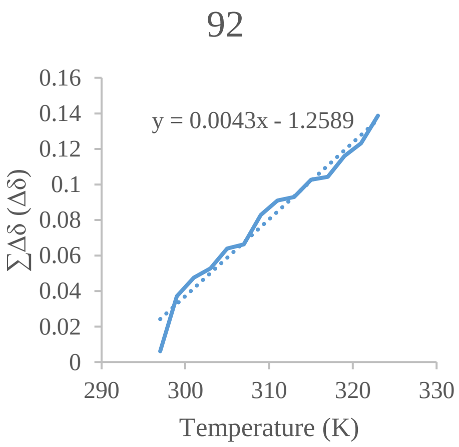


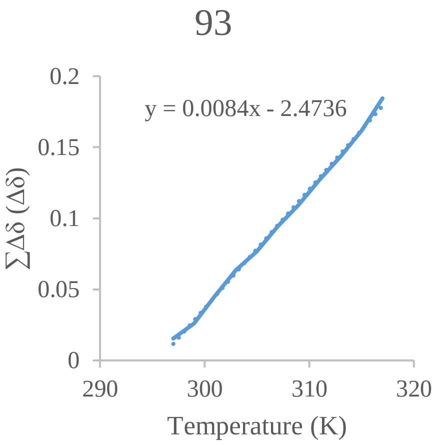
70

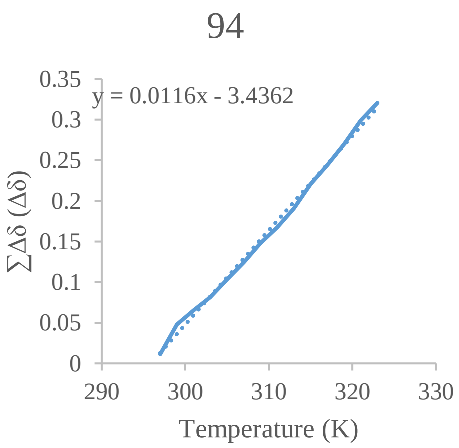


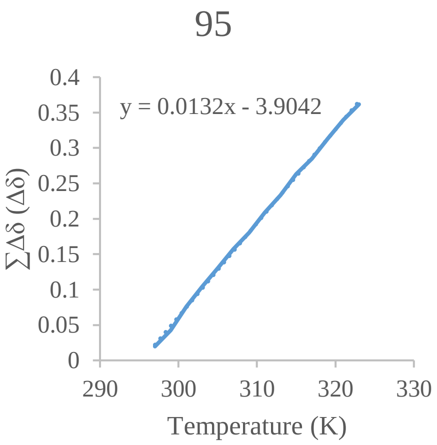

78
y = 0.0084x - 2.4947
∑Δδ (Δδ)
0.25
0.2
0.15
0.1
0.05
0
290
300
310
320
330
Temperature (K)


89
y = 0.0054x - 1.5864
∑Δδ (Δδ)
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
290
300
310
320
330
Temperature (K)


92
y = 0.0043x - 1.2589
∑Δδ (Δδ)
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
290
300
310
320
330
Temperature (K)


93
y = 0.0084x - 2.4736
∑Δδ (Δδ)
0.2
0.15
0.1
0.05
0
290
300
310
320
Temperature (K)


94
y = 0.0116x - 3.4362
∑Δδ (Δδ)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
290
300
310
320
330
Temperature (K)


95
y = 0.0132x - 3.9042
∑Δδ (Δδ)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
290
300
310
320
330
Temperature (K)

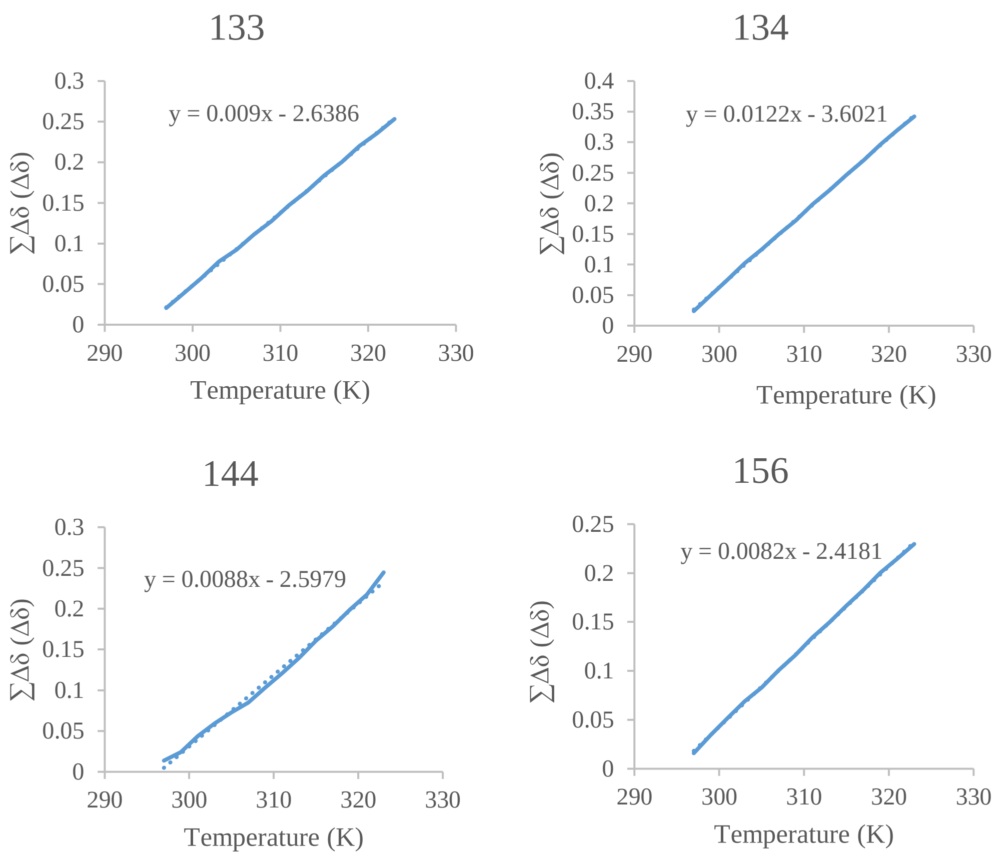


Figure S.6. Monitoring $\sum\Delta\delta$ for 90th Percentile Residues. Individual $\sum\Delta\delta$ vs temperature plots for residues in the top 90th percentile of $\sum\Delta\delta$ are shown. Each $\sum\Delta\delta$ vs temperature plot is accompanied by the respective $y= mx + c$ equation.

Although all $\sum\Delta\delta$ vs temperature plots are linear, deviations from the $y = mx + c$ line differs for all residues. Large deviations from this line may indicate significant transitions at certain temperatures. Therefore, to probe these residues that possibly experience significant transitions, R^2 values are calculated for all residues. Those with a value in the lower quartile (under 0.99) are considered to have a non-linear "$\sum\Delta\delta$-temperature relationship" and are tabulated in Table S.2A. The temperature point at which these residues (red) and those in the $\sum\Delta\delta$ 90th percentile (green) experience their largest deviation from their $y = mx + c$ line is shown in Table S.2B. Both of these residue categories are combined to examine how residues experiencing significant environmental changes may influence conformational transitions for other residues. Residues I56 and A37 are in the $\sum\Delta\delta$ 90th percentile and have R^2 values in the lower quartile. A "temperature line" illustrating residues in Table S.2B on the structure of G-CSF is given in Figure S.7A, showing that when $\sum\Delta\delta$ 90th percentile residues (in green) experience their largest deviation, as do proximal residues in the lower quartile for R^2 (in red).

Residue	R^2
19	0.930
150	0.950
100	0.961
92	0.967
56	0.972
141	0.973
151	0.975
87	0.978
47	0.980
41	0.981
55	0.981
16	0.984
33	0.985
59	0.985
31	0.985
26	0.986
163	0.986
37	0.986
64	0.986
166	0.987
127	0.987
73	0.987
104	0.989
14	0.989
149	0.989

Table S.2A. R^2 values for linearity of $\sum\Delta\delta$ vs temperature for each residue.

	Temperature (K)													
	297	**299**	**301**	**303**	**305**	**307**	**309**	**311**	**313**	**315**	**317**	**319**	**321**	**323**
Residues	69	94		134	63	56	92		33	95	31	6	4	43
	126	89		156	55	144	93			16		12	66	61
	19			149		59	127			41		14	73	68
	64						151			47		87	163	70
	100									141				133
										166				78
														37
														26
														104
														150

Table S.2B. Temperatures at which residues become non-linear in their trajectories for $\sum\Delta\delta$ vs temperature for all residues with $R^2 < 0.99$. Those in the $\sum\Delta\delta$ 90th percentile for lowest R^2 are highlighted in green. Non-linear residues in the lower quartile for R^2 are highlighted in red to match the shading in Figure S8.

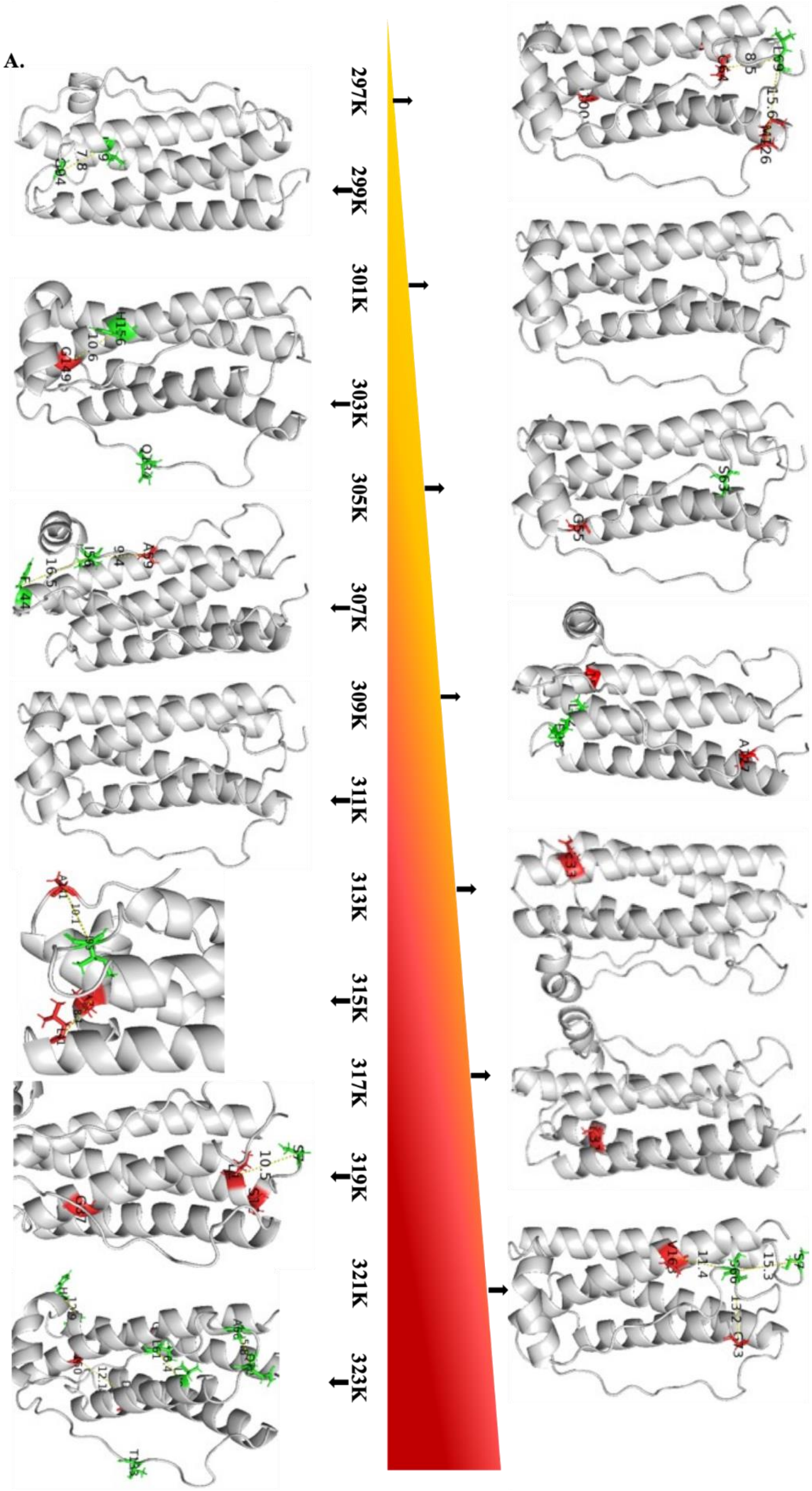


Figure S.7. Schematic temperature line. Residues are highlighted at the temperature at which non-linearity occurred for ∑Δδ 90th percentile residues (in green) and residues in the lower quartile for R^2 (in red). Residue S7 was used in place of G4 and A6.

Percentage Change in PI

Percentage change was calculated using the PI value at the start of the thermal melt and at the maximum point, which is typically around 309 K.

Percentage Increase in ? PI							
Residue Number	%Increase	Residue Number	%Increase	Residue Number	%Increase	Residue Number	%Increase
51	297	115	124	16	95	135	54
43	241	14	123	76	95	156	53
153	229	124	122	134	95	72	51
70	227	99	121	15	94	64	50
94	210	144	121	25	91	163	50
55	203	62	121	119	91	81	47
36	202	107	120	69	90	27	43
59	202	35	119	29	89	133	41
38	192	155	119	157	89	169	36
56	190	31	112	170	89	8	36
126	184	47	112	164	88	48	32
89	183	39	110	61	86	7	23
33	181	66	110	73	85	6	19
41	166	171	110	42	84	50	18
106	157	142	109	58	84	4	15
71	156	114	109	166	83	120	9
93	151	100	108	102	83	1	7
78	150	95	107	172	80		
63	149	87	106	68	78		
37	144	84	106	26	78		
28	143	80	105	141	77		
145	137	150	104	19	77		
92	137	127	102	12	75		
161	134	159	100	173	75		
151	134	105	100	160	72		
40	134	79	100	112	71		
146	134	110	99	75	70		
158	133	118	98	22	70		
165	133	149	98	104	56		
98	132	121	97	139	56		
13	130	125	96	143	55		
11	129	122	96	23	55		

Table S.3. Percentage increases in PI (in white columns) are given for all assigned residues (in grey columns), which is the only residue to show a decrease in PI at the start of the melt. Top 15 percentage increases are highlighted in yellow.

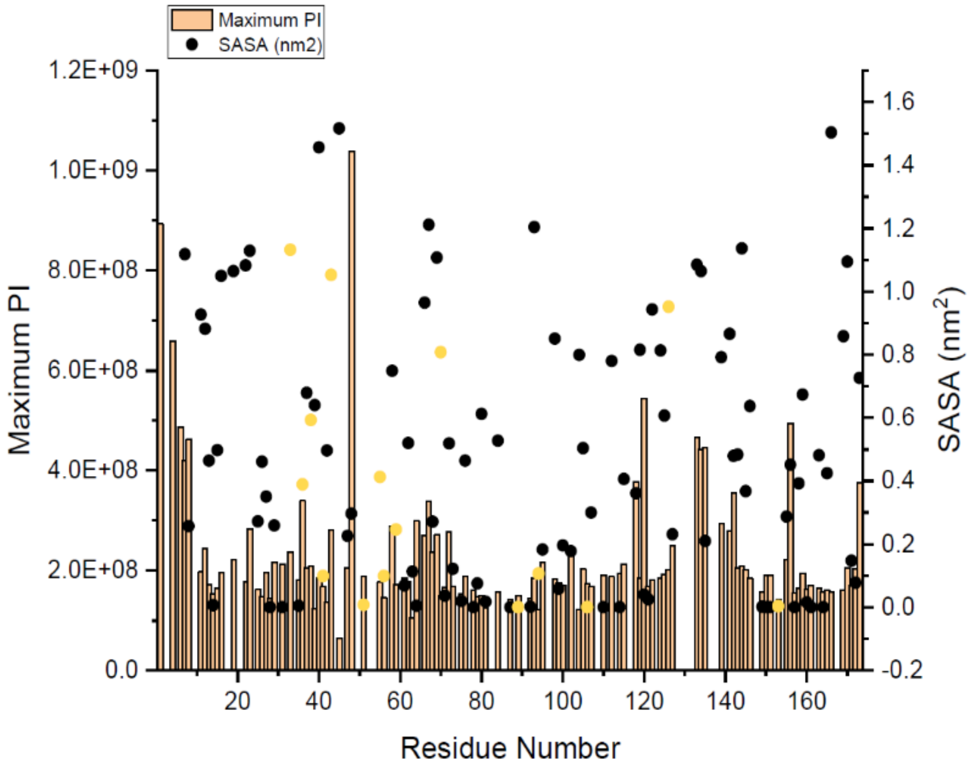


Figure S.8. Top 15 Residues with Highest Percentage Increase in ΔPI. Maximum PI vs. SASA. Residues in the sub-clusters have their SASA highlighted yellow.

Mapping Significant Structural Changes on to G-CSF

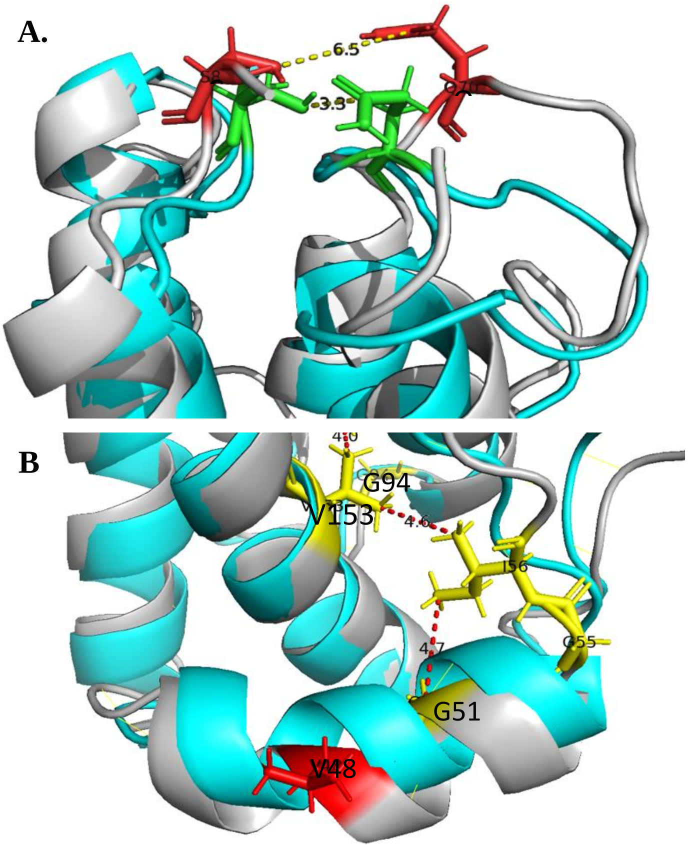
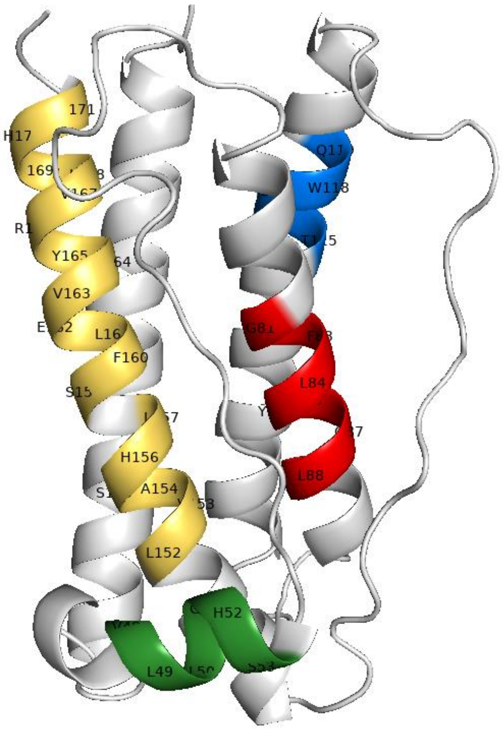


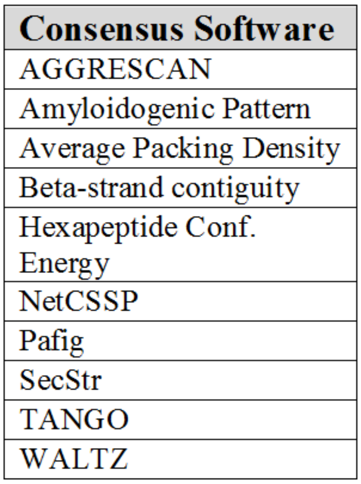
Figure S.9. Proximity of residue S8 to Q70 and exposure of Sub-cluster 2 by V48.

Relaxed (cyan) and unrelaxed (grey) G-CSF structures are overlaid. **A.** S8 and Q70 are highlighted green in the relaxed structure and red in the unrelaxed structure. Distance between both of these residues is 3.3 Ǻ in the relaxed structure and 6.5 Ǻ in the unrelaxed structure. **B.** shows residues in sub-cluster 2 (yellow). Distance between these residues are indicated along red dotted lines. Residue V48 is highlighted red.

APRs

AmylPred 2 (Tsolis *et al*., 2013) employs a consensus method to identify aggregation prone regions (APRs) APRs, combing results from several different software mentioned in Figure S.10. The 5 different APR consensus regions predicted by AmylPred 2 based on successful hits from at least 5 out of the 10 software are V48-S53, G81-L89, T115-Q119, L152-L157 and S159-L171. These regions are respectively coloured as green, red, blue and yellow (for L152-L157 and S159-L171 since they are so close in sequence) in. Helix D has the largest hotspot of APRs (coloured yellow).

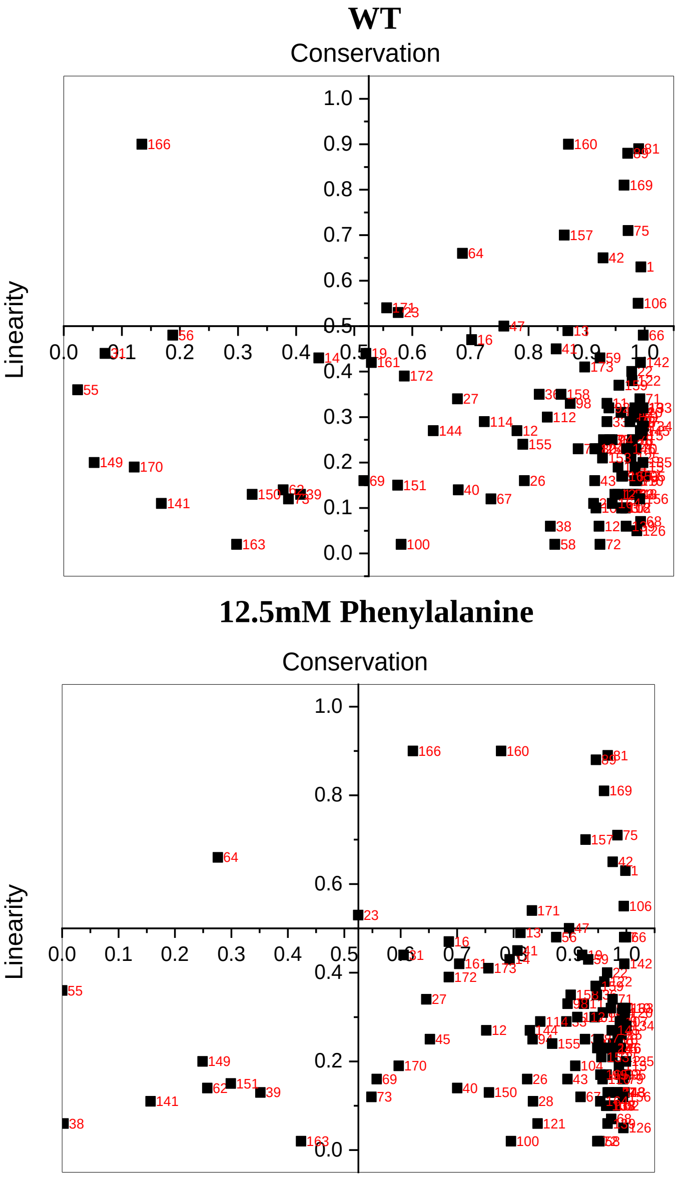




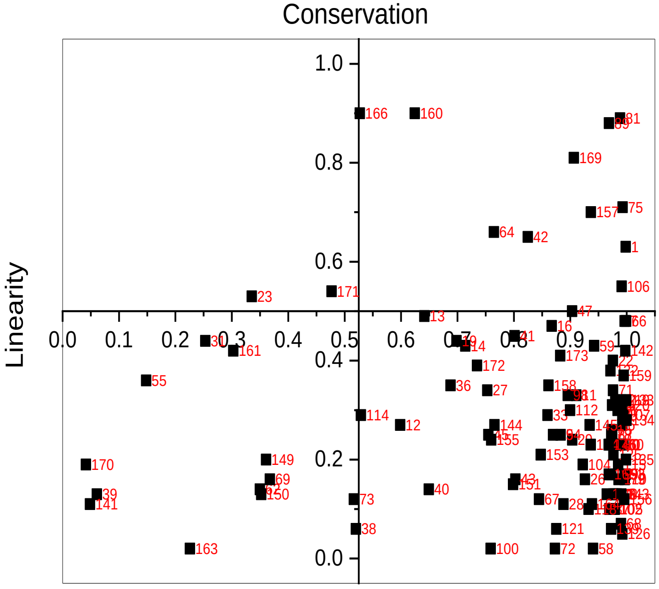
Consensus Software
AGGRESCAN
Amyloidogenic Pattern
Average Packing Density
Beta-strand contiguity
Hexapeptide Conf. Energy
NetCSSP
Pafig
SecStr
TANGO
WALTZ

Figure S.10. Identifying APRs. Five different Consensus APRs determined from AmylPred 2 are coloured yellow, red, green and blue. Two regions are coloured yellow because they are only one residue apart. The ten APR scanning software used for the consensus are shown in the accompanying table.

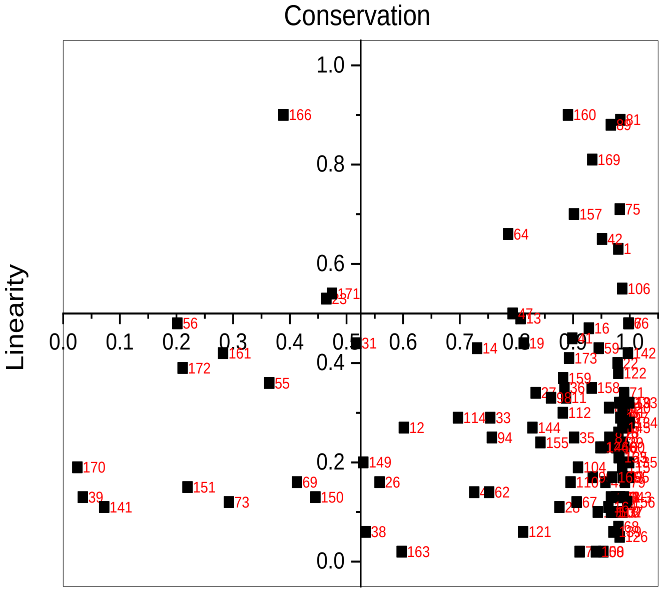
Residue Conservation and Peak Trajectory Linearity



25mM Histidine



25mM Arginine



50mM Arginine

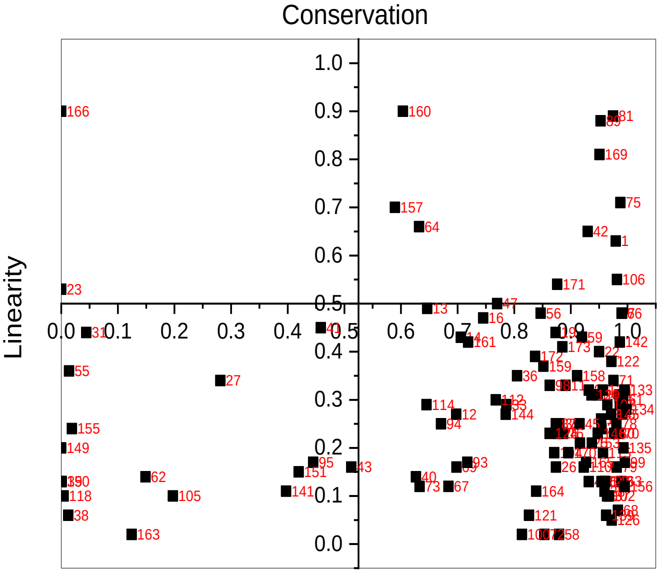
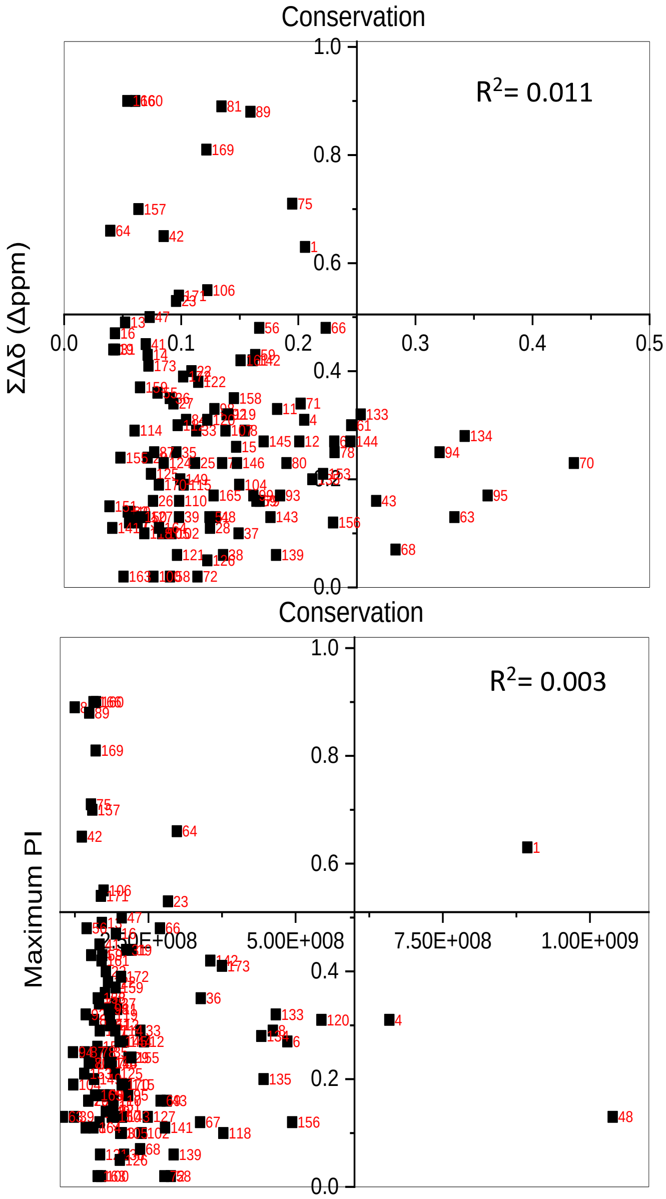
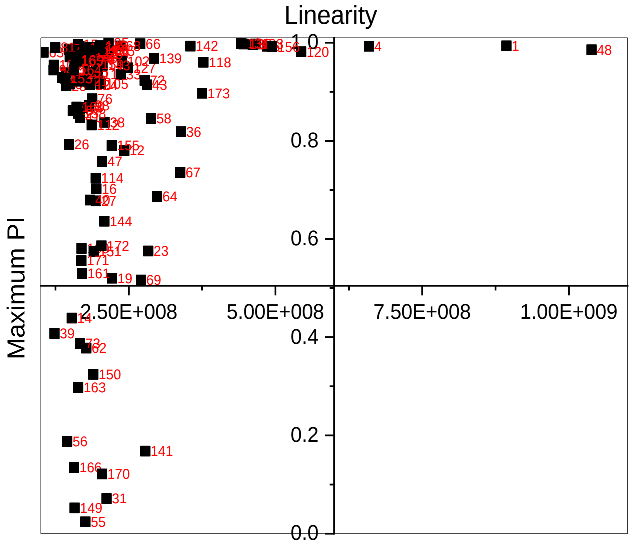
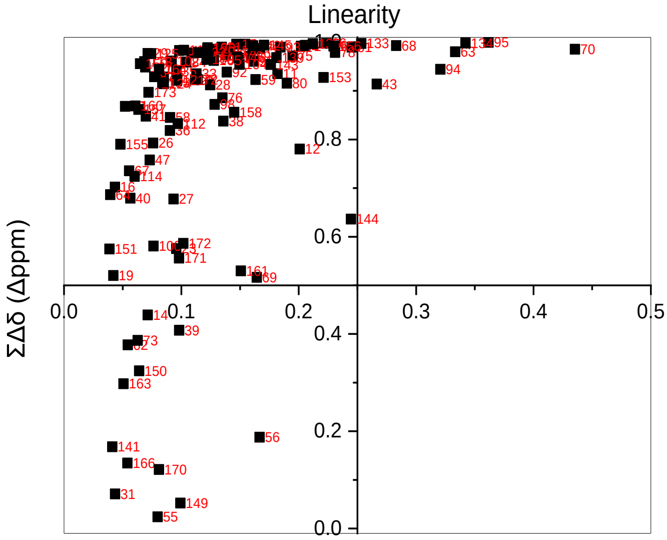


Figure S.11. Linearity vs Conservation. Residues are labelled in red.

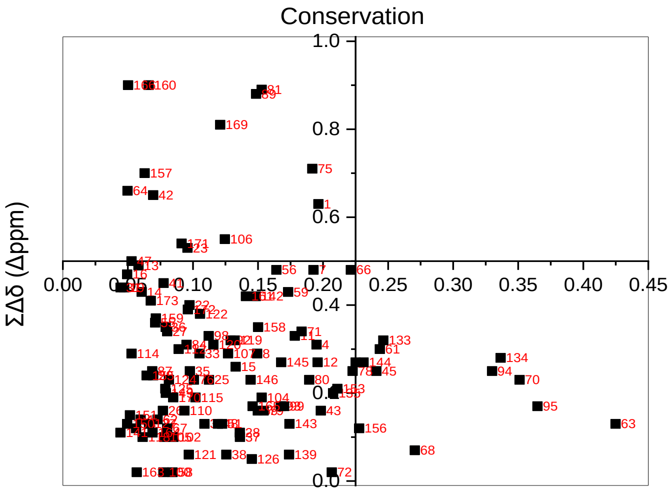
A. WT

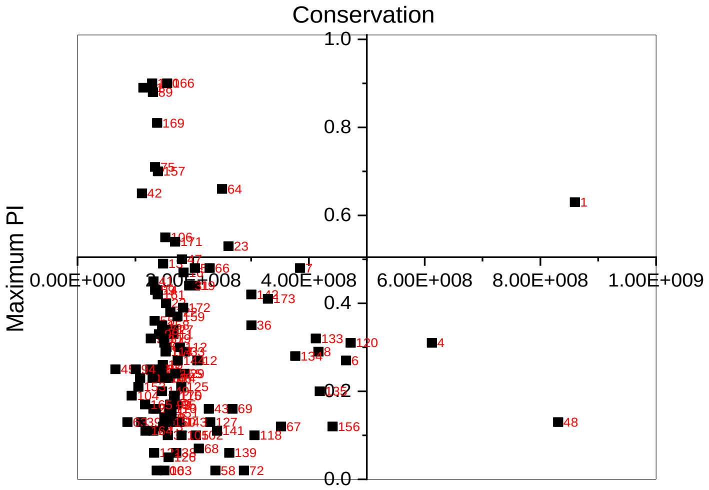


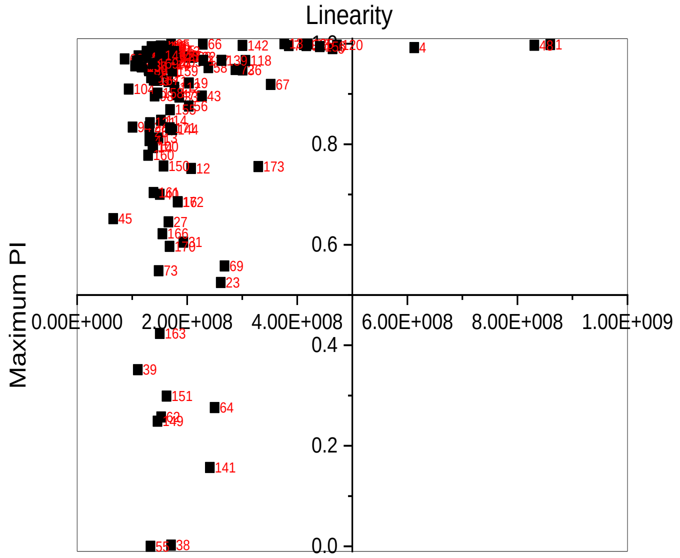


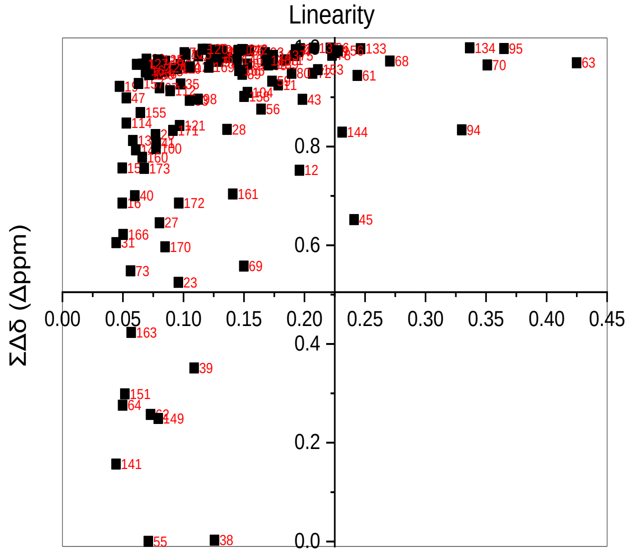


B. 12.5mM Phenylalanine

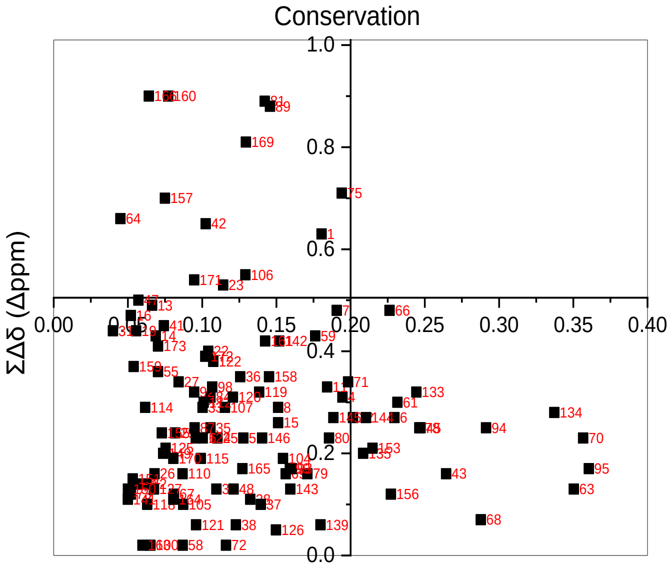


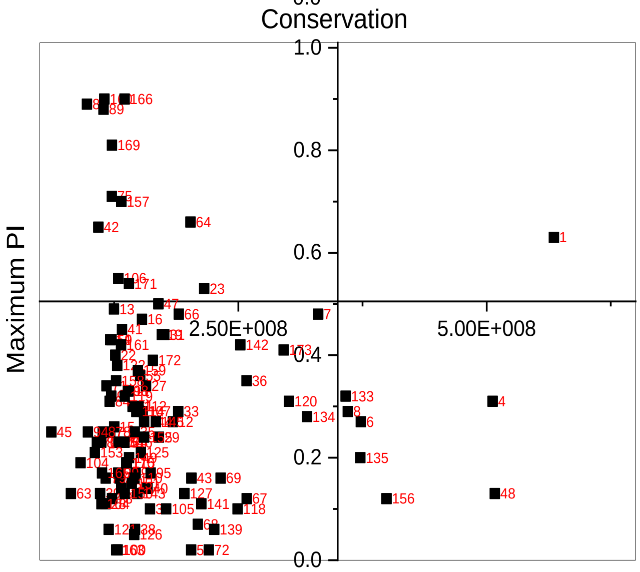

Conservation
Maximum PI
1.0
0.8
0.6
0.4
0.2
0.0
0.00E+000
2.00E+008
4.00E+008
6.00E+008
8.00E+008
1.00E+009

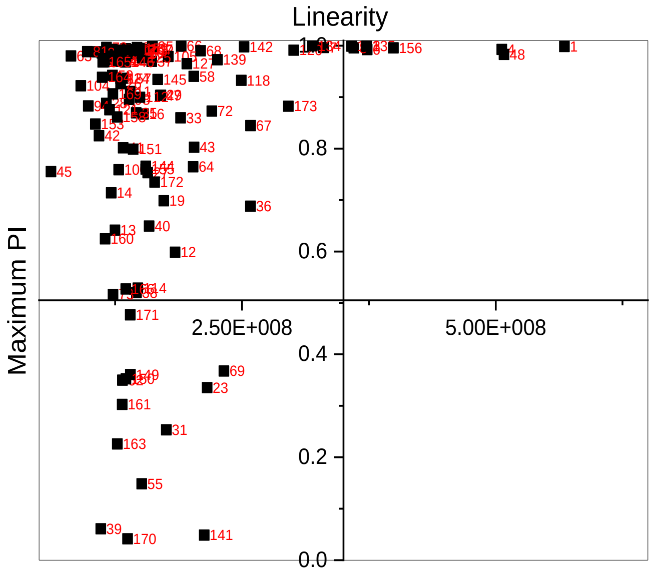

Linearity
Maximum PI
0.8
0.6
0.4
0.2
0.0
0.00E+000
2.00E+008
4.00E+008
6.00E+008
8.00E+008
1.00E+009

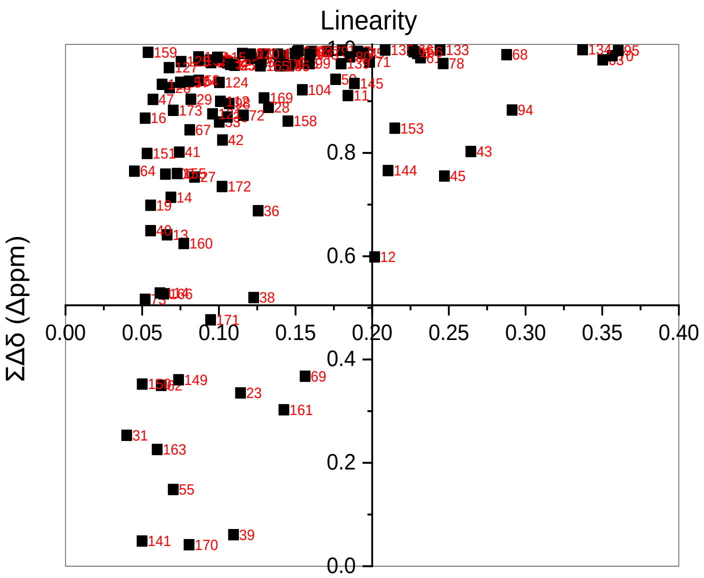

Linearity
ΣΔδ (Δppm)
0.8
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

C. 25mM Histidine

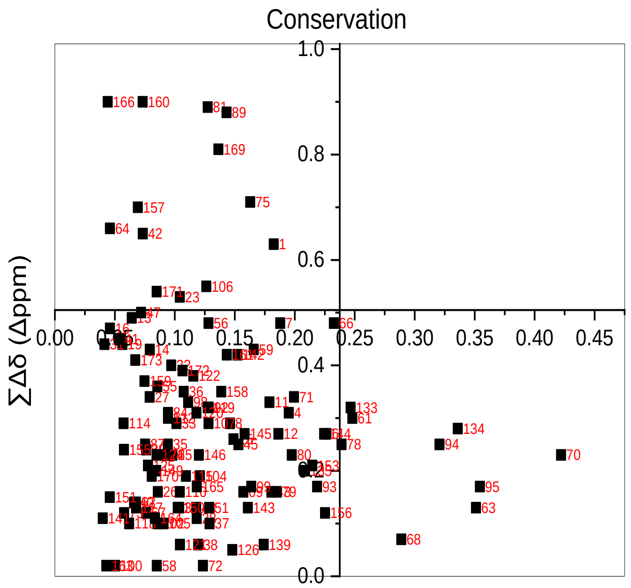


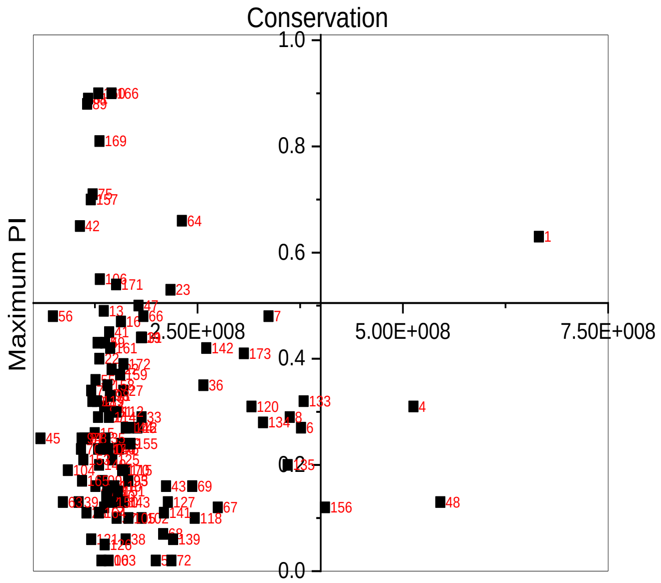


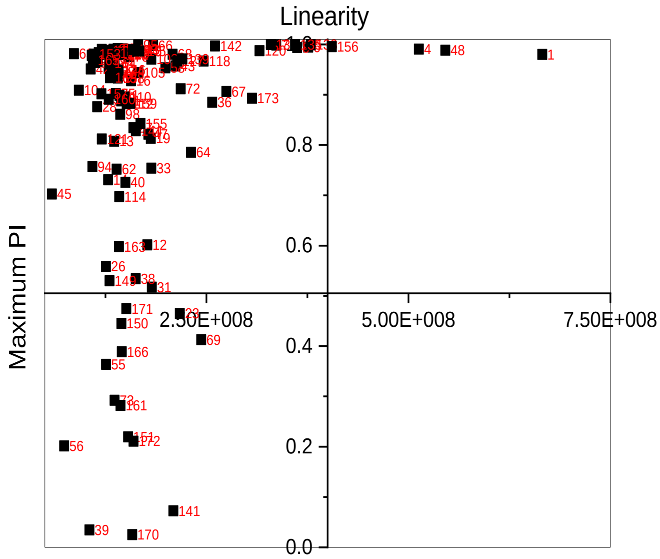


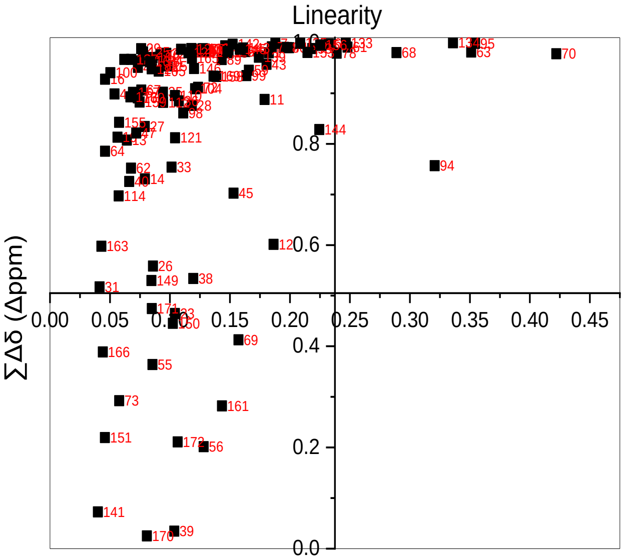


D. 25mM Arginine

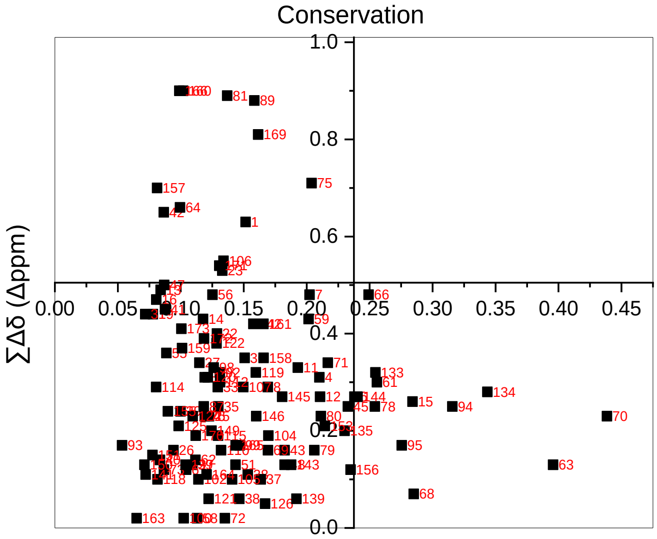


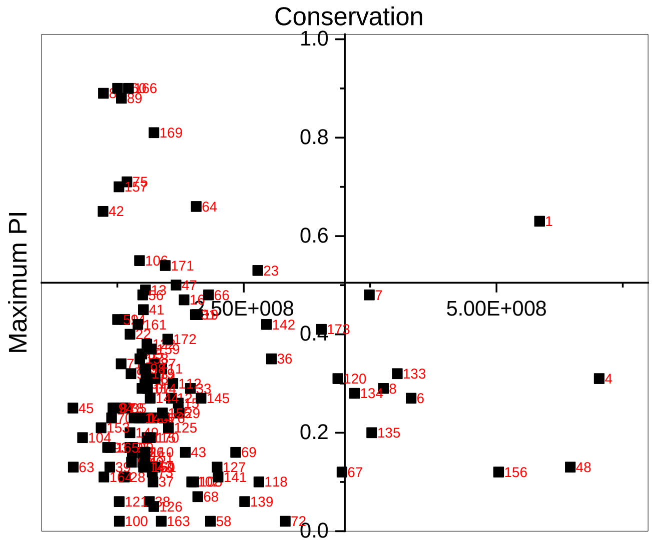


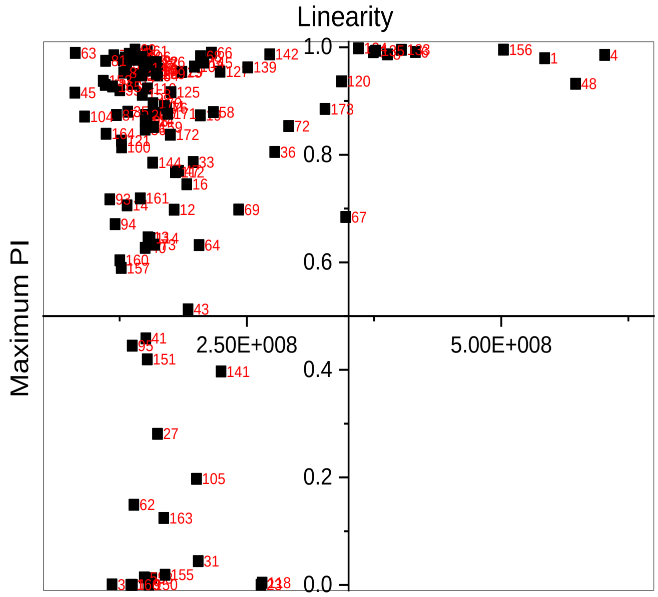




E. 50mM Arginine




Conservation
Maximum PI
1.0
0.8
0.6
0.4
0.2
0.0
2.50E+008
5.00E+008


Linearity
Maximum PI
1.0
0.8
0.6
0.4
0.2
0.0
2.50E+008
5.00E+008

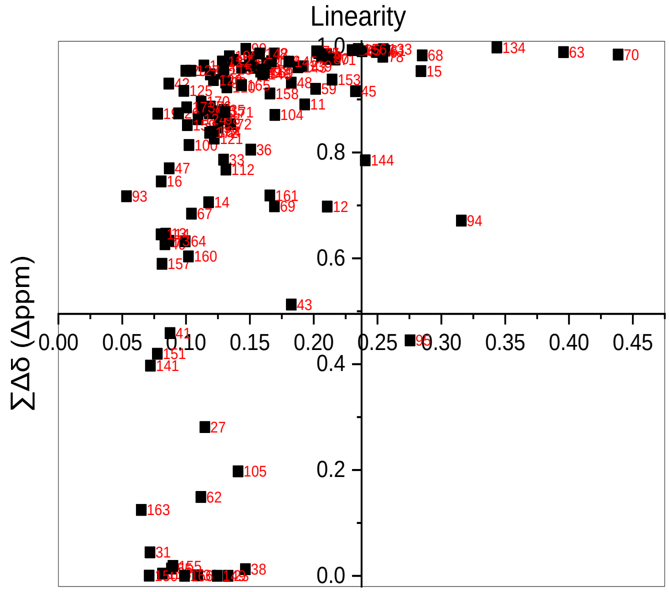

Linearity
∑Δδ (Δppm)
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

Figure S.12. ΣΔδ and Maximum PI vs Conservation and Linearity. A.-D. ΣΔδ and Maximum PI are compared with conservation on the top graphs and linearity on the bottom graphs. Residues are labelled red.

CG-MD

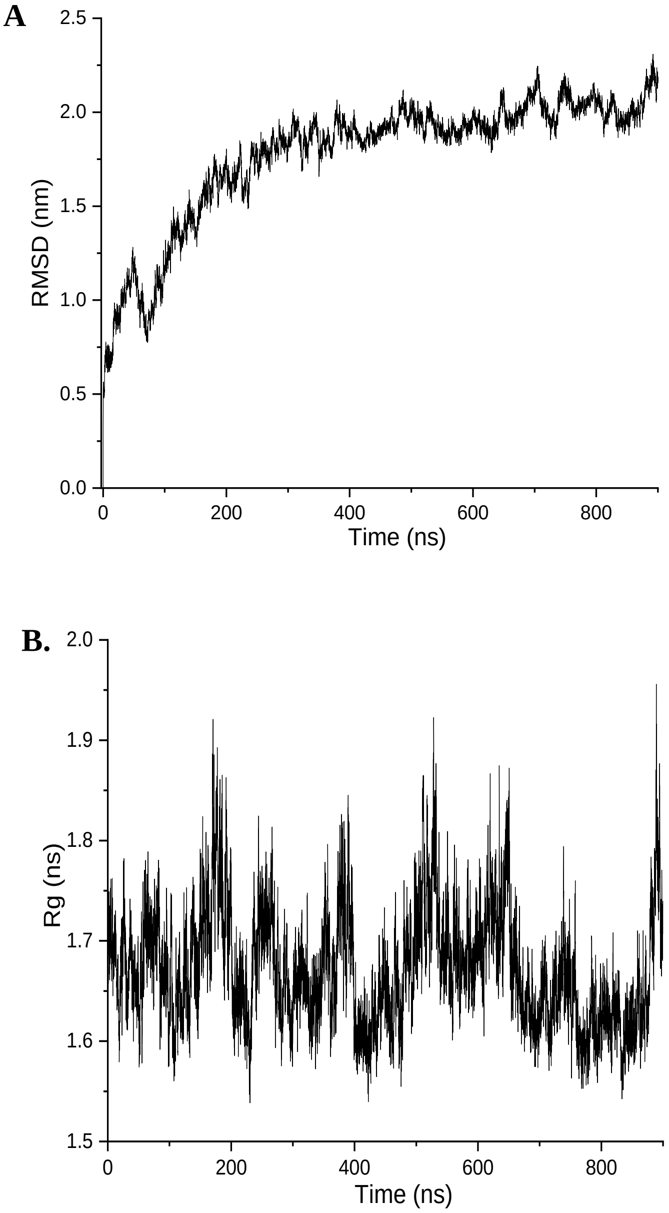
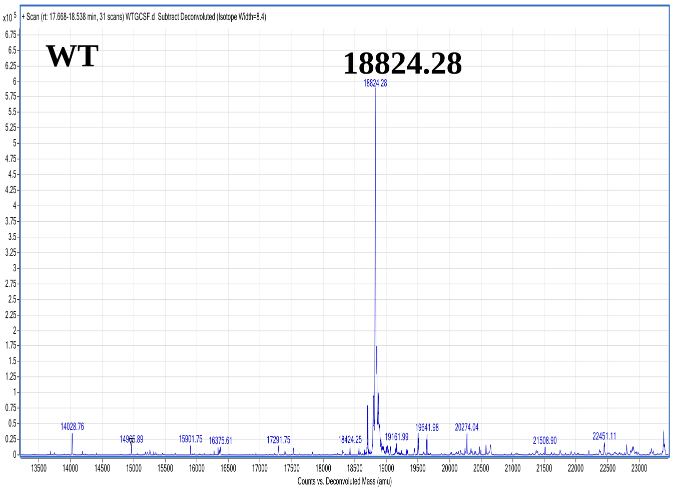
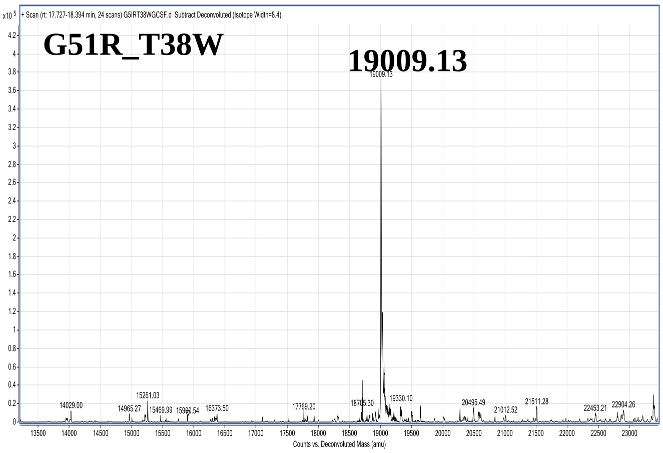
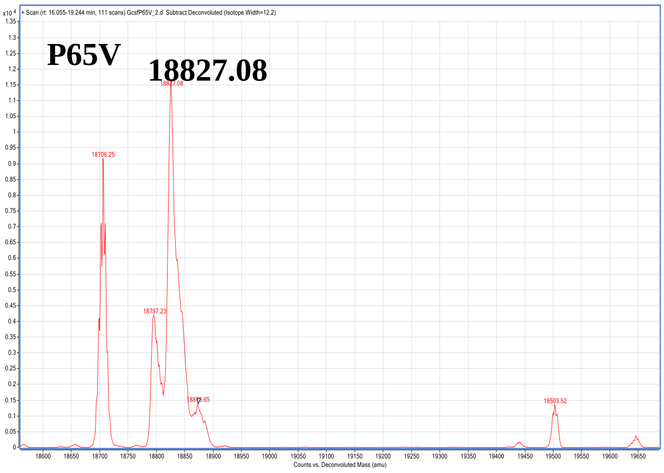


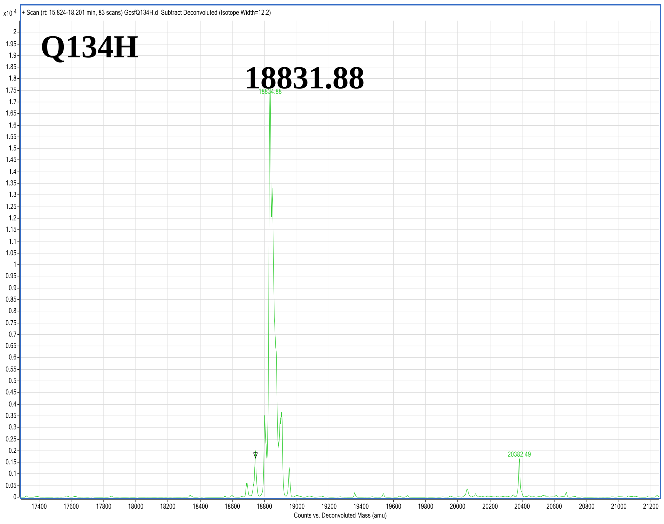
Figure S.13. CG-MD. RMSD (**A.**) and Rg (**B.**) calculated from a single CG-MD run.

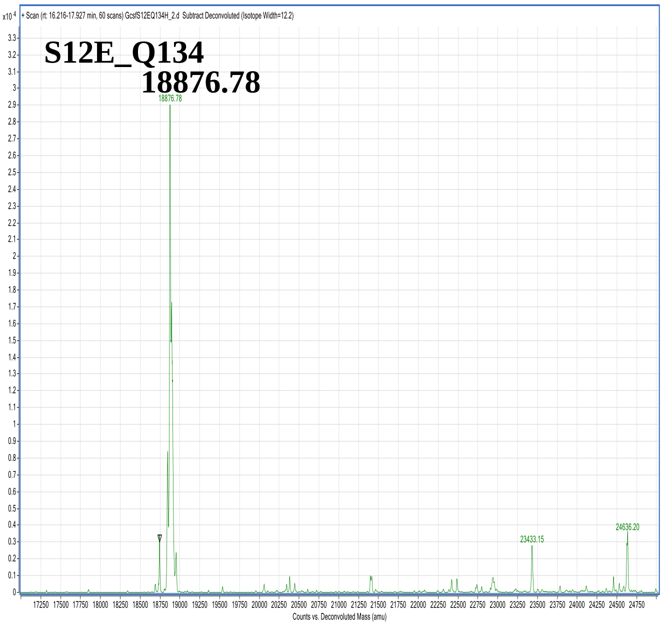
Mass Spectrometry

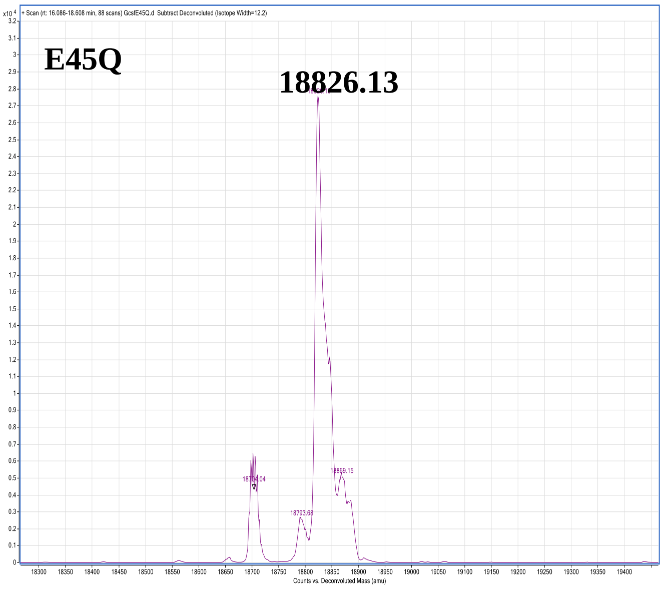


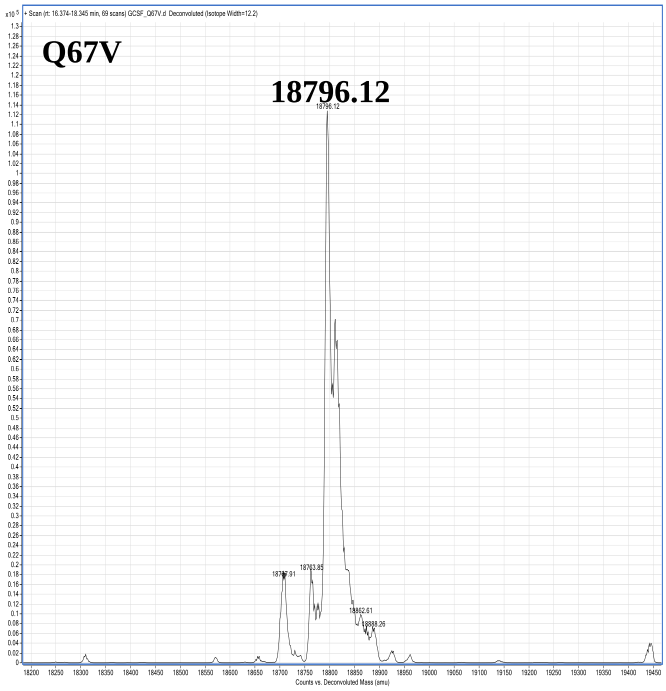





x10 4
+ Scan (rt: 15.824-18.201 min, 83 scans) GcsfQ134H.d Subtract Deconvoluted (Isotope Width=12.2)
Q134H
18831.88
18834.88
20382.49
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
1.05
1.1
1.15
1.2
1.25
1.3
1.35
1.4
1.45
1.5
1.55
1.6
1.65
1.7
1.75
1.8
1.85
1.9
1.95
2
17400
17600
17800
18000
18200
18400
18600
18800
19000
19200
19400
19600
19800
20000
20200
20400
20600
20800
21000
21200
Counts vs. Deconvoluted Mass (amu)


x10 4
+ Scan (rt: 16.216-17.927 min, 60 scans) GcsfS12EQ134H_2.d Subtract Deconvoluted (Isotope Width=12.2)
S12E_Q134
18876.78
18876.78
23433.15
24636.20
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3
3.1
3.2
3.3
17250
17500
17750
18000
18250
18500
18750
19000
19250
19500
19750
20000
20250
20500
20750
21000
21250
21500
21750
22000
22250
22500
22750
23000
23250
23500
23750
24000
24250
24500
24750
Counts vs. Deconvoluted Mass (amu)

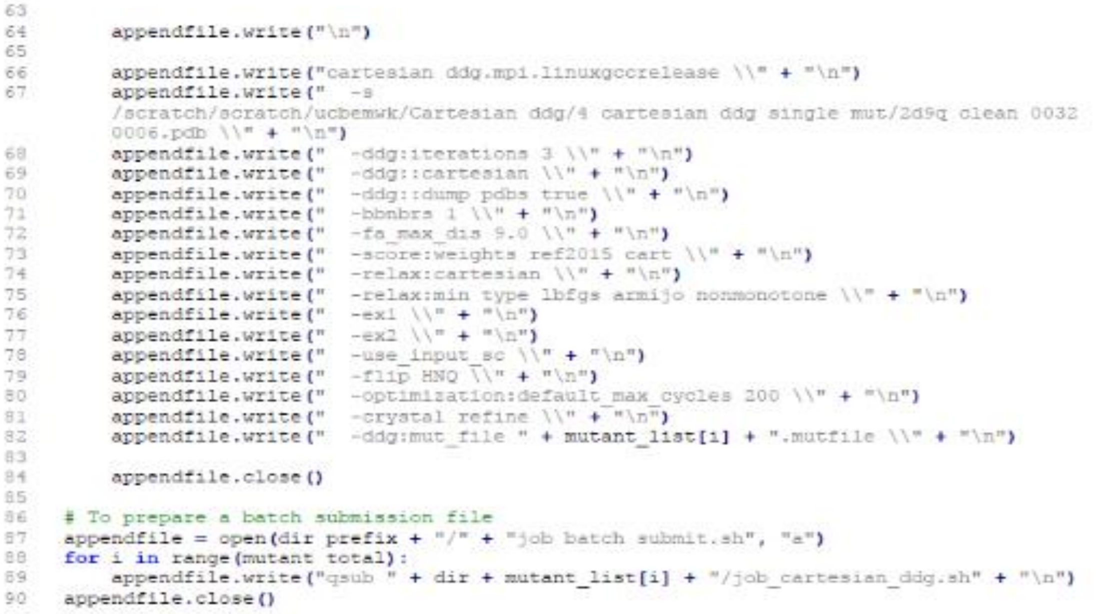

x10 4
+ Scan (rt: 16.086-18.608 min, 88 scans) GcsfE45Q.d Subtract Deconvoluted (Isotope Width=12.2)
E45Q
18826.13
18704.04
18793.68
18869.15
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3
3.1
3.2
18300
18350
18400
18450
18500
18550
18600
18650
18700
18750
18800
18850
18900
18950
19000
19050
19100
19150
19200
19250
19300
19350
19400
Counts vs. Deconvoluted Mass (amu)



S.14. Mutant Molecular Weights. The main peak corresponds to the molecular weight (labelled above) of the G-CSF variant.

Cartesian_ddg

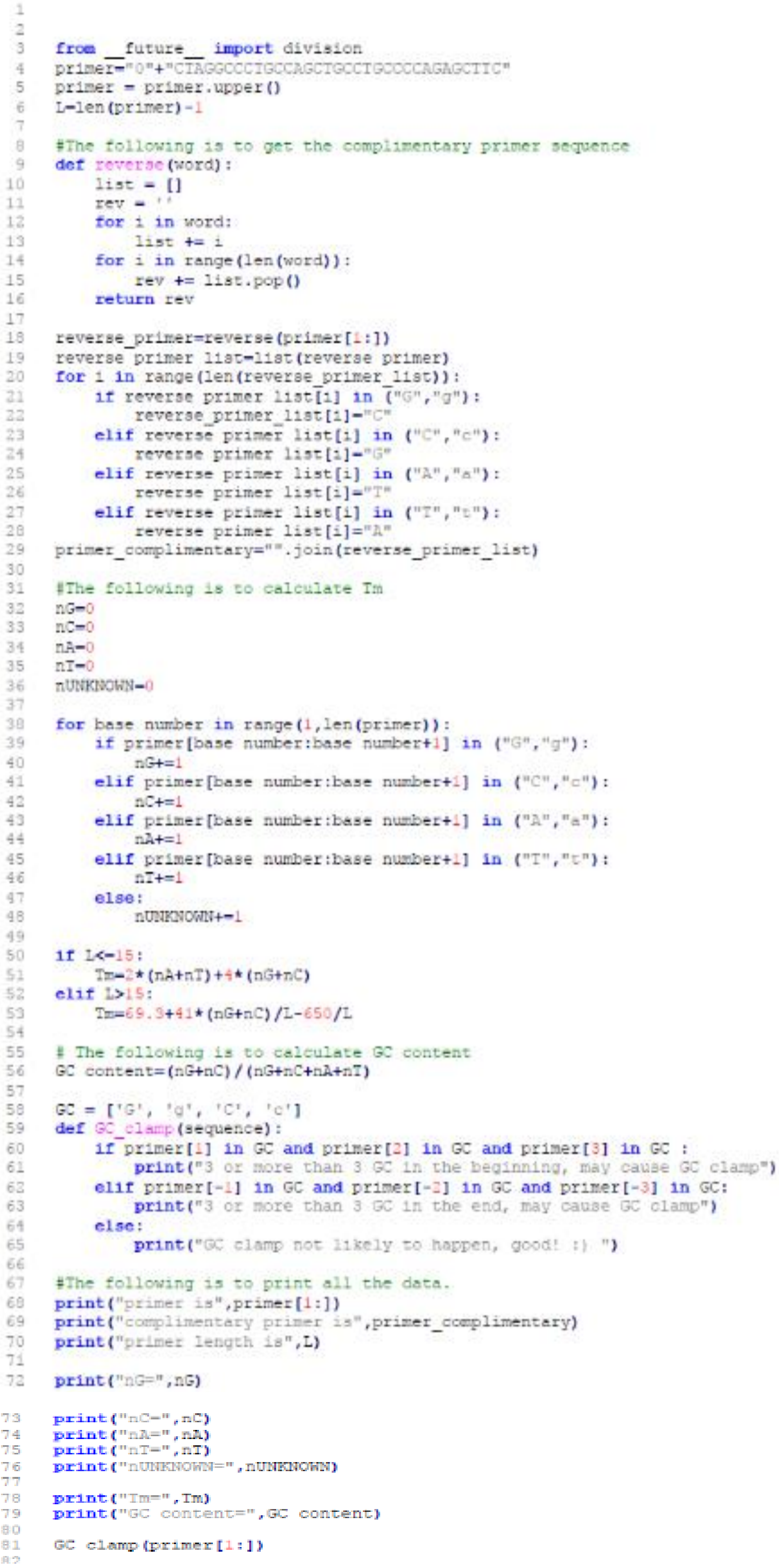
```python
# This file will generate single mutant folders (e.g. S350G).

# Each folder consists of job_cartesian_ddg.sh and mutfile (e.g. S350G.mutfile)

mutant_total = 3192 # 19*168 = 3192
residue_total = 168 # the total number of residues in the protein
AA_total = 20 # there are 20 amino acid
dir = "/scratch/scratch/ucbemwk/Cartesian_ddg/4_cartesian_ddg_single_mut/result/"
dir_prefix = "/scratch/scratch/ucbemwk/Cartesian_ddg/4_cartesian_ddg_single_mut/"

# The first thing is to build the names for 3192 mutants (e.g. D1A, D1C, ..., I1A...)
sequence=["S","S","L","P","Q","S","F","L","L","K","C","L","E","Q","V","R","K","I","Q","G","D","G","A","A","L","Q","E","K","L","C","A","T","Y","K","L","C","H","P","E","E","L","V","L","L","G","H","S","L","G","I","P","W","A","P","L","S","S","C","P","S","Q","A","L","Q","L","A","G","C","L","S","Q","L","H","S","G","L","F","L","Y","Q","G","L","L","Q","A","L","E","G","I","S","P","E","L","G","P","T","L","D","T","L","Q","L","D","V","A","D","F","A","T","T","I","W","Q","Q","M","E","E","L","G","M","A","P","A","L","Q","P","T","Q","G","A","M","P","A","F","A","S","A","F","Q","R","R","A","G","G","V","L","V","A","S","H","L","Q","S","F","L","E","V","S","Y","R","V","L","R","H","L","A","Q","P"]
AA=["A","C","D","E","F","G","H","I","K","L","M","N","P","Q","R","S","T","V","W","Y"]

mutant_list=[]
for i in range(mutant_total):
    mutant_list.append("")

i = 0
for residue_number in range(residue_total):
    for AA_number in range(AA_total):
        if sequence[residue_number]==AA[AA_number]:
            AA_number=AA_number+1
        else:

            mutant_list[i]=sequence[residue_number]+str(residue_number+1)+AA[AA_number]
            i+=1
#print(mutant_list)
# At this point, the mutant_list contains 3192 point mutations.

# To prepare the empty mutant folders
import os
for i in range(mutant_total):
    if not os.path.exists(dir+mutant_list[i]):
        os.makedirs(dir+mutant_list[i])

# To prepare .mutfile in each of the 3192 mutant folders
for i in range(mutant_total):
    appendfile = open(dir + mutant_list[i] + "/" + mutant_list[i]+".mutfile", "a")
    appendfile.write("total 1" + "\n")
    appendfile.write("1" + "\n")
    appendfile.write(mutant_list[i][0] + " " + mutant_list[i][1:-1] + " " + mutant_list[i][-1])
    appendfile.close()

# To prepare a job_cartesian_ddg.sh file in each of the 3192 mutant folders
for i in range(mutant_total):
    appendfile = open(dir + mutant_list[i] +"/" + "job_cartesian_ddg.sh", "a")
    appendfile.write("#!/bin/bash -l" + "\n")
    appendfile.write("#$ -S /bin/bash" + "\n")
    appendfile.write("#$ -l h_rt=1:0:0" + "\n")
    appendfile.write("#$ -l mem=2G" + "\n")
    appendfile.write("#$ -l tmpfs=15G" + "\n")
    appendfile.write("#$ -N " + mutant_list[i] + "\n")
    appendfile.write("#$ -pe mpi 1" + "\n")
    appendfile.write("#$ -wd " + dir + mutant_list[i] + "\n")

    appendfile.write("\n")

    appendfile.write("module unload compilers mpi" + "\n")
    appendfile.write("module load compilers/gnu/4.9.2" + "\n")
    appendfile.write("module load python2/recommended" + "\n")
    appendfile.write("module load mpi/openmpi/3.1.1/gnu-4.9.2" + "\n")
    appendfile.write("module load rosetta/2018.48.60516-mpi" + "\n")
```



```

        appendfile.write("\n")

        appendfile.write("cartesian ddg.mpi.linuxgccrelease \\" + "\n")
        appendfile.write("  -s
        /scratch/scratch/ucbemwk/Cartesian ddg/4 cartesian ddg single mut/2d9q clean 0032
        0006.pdb \\" + "\n")
        appendfile.write("  -ddg::iterations 3 \\" + "\n")
        appendfile.write("  -ddg::cartesian \\" + "\n")
        appendfile.write("  -ddg::dump pdbs true \\" + "\n")
        appendfile.write("  -bbnbrs 1 \\" + "\n")
        appendfile.write("  -fa_max_dis 9.0 \\" + "\n")
        appendfile.write("  -score:weights ref2015 cart \\" + "\n")
        appendfile.write("  -relax:cartesian \\" + "\n")
        appendfile.write("  -relax:min type lbfgs armijo nonmonotone \\" + "\n")
        appendfile.write("  -ex1 \\" + "\n")
        appendfile.write("  -ex2 \\" + "\n")
        appendfile.write("  -use_input_sc \\" + "\n")
        appendfile.write("  -flip HNQ \\" + "\n")
        appendfile.write("  -optimization:default_max_cycles 200 \\" + "\n")
        appendfile.write("  -crystal refine \\" + "\n")
        appendfile.write("  -ddg:mut_file " + mutant_list[i] + ".mutfile \\" + "\n")

        appendfile.close()

    # To prepare a batch submission file
    appendfile = open(dir prefix + "/" + "job batch submit.sh", "a")
    for i in range(mutant total):
        appendfile.write("qsub " + dir + mutant_list[i] + "/job_cartesian_ddg.sh" + "\n")
    appendfile.close()
```

S.15. Single Mutant Cartesian_ddg Script. Python script to calculate the folding energy for all possible single mutations (3192) for GCSF.

Primer Design



```
from __future__ import division
primer="0"+"CTAGGCCCTGCCAGCTGCCTGCCCCAGAGCTTC"
primer = primer.upper()
L=len(primer)-1

#The following is to get the complimentary primer sequence
def reverse(word):
    list = []
    rev = ''
    for i in word:
        list += i
    for i in range(len(word)):
        rev += list.pop()
    return rev

reverse_primer=reverse(primer[1:])
reverse primer list=list(reverse primer)
for i in range(len(reverse_primer_list)):
    if reverse primer list[i] in ("G","g"):
        reverse_primer_list[i]="C"
    elif reverse primer list[i] in ("C","c"):
        reverse primer list[i]="G"
    elif reverse primer list[i] in ("A","a"):
        reverse primer list[i]="T"
    elif reverse primer list[i] in ("T","t"):
        reverse primer list[i]="A"
primer_complimentary="".join(reverse_primer_list)

#The following is to calculate Tm
nG=0
nC=0
nA=0
nT=0
nUNKNOWN=0

for base number in range(1,len(primer)):
    if primer[base number:base number+1] in ("G","g"):
        nG+=1
    elif primer[base number:base number+1] in ("C","c"):
        nC+=1
    elif primer[base number:base number+1] in ("A","a"):
        nA+=1
    elif primer[base number:base number+1] in ("T","t"):
        nT+=1
    else:
        nUNKNOWN+=1

if L<=15:
    Tm=2*(nA+nT)+4*(nG+nC)
elif L>15:
    Tm=69.3+41*(nG+nC)/L-650/L

# The following is to calculate GC content
GC content=(nG+nC)/(nG+nC+nA+nT)

GC = ['G', 'g', 'C', 'c']
def GC_clamp(sequence):
    if primer[1] in GC and primer[2] in GC and primer[3] in GC :
        print("3 or more than 3 GC in the beginning, may cause GC clamp")
    elif primer[-1] in GC and primer[-2] in GC and primer[-3] in GC:
        print("3 or more than 3 GC in the end, may cause GC clamp")
    else:
        print("GC clamp not likely to happen, good! :) ")

#The following is to print all the data.
print("primer is",primer[1:])
print("complimentary primer is",primer_complimentary)
print("primer length is",L)

print("nG=",nG)

print("nC=",nC)
print("nA=",nA)
print("nT=",nT)
print("nUNKNOWN=",nUNKNOWN)

print("Tm=",Tm)
print("GC content=",GC content)

GC clamp(primer[1:])
```

S.16. Primer Design Script. Python script to obtain complementary primer sequence and calculate Tm and potential for GC clamping.

```
Pairwise Alignment
Sequence 1: WT
Sequence 2: Q134H
Sequence ends allowed to slide over each other
Alignment score: 1041

Identities:   0.9980843


             ....|....| ....|....| ....|....| ....|....| ....|....|
                      10         20         30         40         50
WT           ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG
Q134H        ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG


             ....|....| ....|....| ....|....| ....|....| ....|....|
                      60         70         80         90        100
WT           CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
Q134H        CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA

             ....|....| ....|....| ....|....| ....|....| ....|....|
                     110        120        130        140        150
WT           AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC
Q134H        AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC

             ....|....| ....|....| ....|....| ....|....| ....|....|
                     160        170        180        190        200
WT           GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA
Q134H        GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA

             ....|....| ....|....| ....|....| ....|....| ....|....|
                     210        220        230        240        250
WT           GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC
Q134H        GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC

             ....|....| ....|....| ....|....| ....|....| ....|....|
                     260        270        280        290        300
WT           TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT
Q134H        TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT

             ....|....| ....|....| ....|....| ....|....| ....|....|
                     310        320        330        340        350
WT           CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
Q134H        CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT

             ....|....| ....|....| ....|....| ....|....| ....|....|
                     360        370        380        390        400
WT           CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC
Q134H        CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC

             ....|....| ....|....| ....|....| ....|....| ....|....|
                     410        420        430        440        450
WT           AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
Q134H        ATGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
```

```
            ....|....| ....|....| ....|....| ....|....| ....|....|
                    460        470        480        490        500
WT          GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT
Q134H       GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT

            ....|....| ....|....| ..
                    510        520
WT          TCTACGCCAC CTTGCCCAGC CC
Q134H       TCTACGCCAC CTTGCCCAGC CC


Pairwise Alignment
Sequence 1: WT
Sequence 2: S8C
Sequence ends allowed to slide over each other
Alignment score: 1041

Identities:   0.9980843


            ....|....| ....|....| ....|....| ....|....| ....|....|
                     10         20         30         40         50
WT          ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG
S8C         ACACCCCTAG GCCCTGCCAG CTGCCTGCCC CAGAGCTTCC TGCTCAAGTG

            ....|....| ....|....| ....|....| ....|....| ....|....|
                     60         70         80         90        100
WT          CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
S8C         CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA

            ....|....| ....|....| ....|....| ....|....| ....|....|
                    110        120        130        140        150
WT          AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC
S8C         AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC

            ....|....| ....|....| ....|....| ....|....| ....|....|
                    160        170        180        190        200
WT          GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA
S8C         GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA

            ....|....| ....|....| ....|....| ....|....| ....|....|
                    210        220        230        240        250
WT          GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC
S8C         GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC

            ....|....| ....|....| ....|....| ....|....| ....|....|
                    260        270        280        290        300
WT          TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT
S8C         TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT

            ....|....| ....|....| ....|....| ....|....| ....|....|
                    310        320        330        340        350
WT          CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
S8C         CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
```

```
             ....|....| ....|....| ....|....| ....|....| ....|....|
                    360        370        380        390        400
WT           CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC
S8C          CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    410        420        430        440        450
WT           AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
S8C          AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    460        470        480        490        500
WT           GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT
S8C          GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT

             ....|....| ....|....| ..
                    510        520
WT           TCTACGCCAC CTTGCCCAGC CC
S8C          TCTACGCCAC CTTGCCCAGC CC
```

```
Pairwise Alignment
Sequence 1: WT
Sequence 2: E45Q
Sequence ends allowed to slide over each other
Alignment score: 1041

Identities:   0.9980843
```

```
             ....|....| ....|....| ....|....| ....|....| ....|....|
                     10         20         30         40         50
WT           ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG
E45Q         ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG

             ....|....| ....|....| ....|....| ....|....| ....|....|
                     60         70         80         90        100
WT           CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
E45Q         CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    110        120        130        140        150
WT           AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC
E45Q         AGCTGTGTGC CACCTACAAG CTGTGCCACC CCCAGGAGCT GGTGCTGCTC

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    160        170        180        190        200
WT           GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA
E45Q         GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    210        220        230        240        250
```

```
WT          GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC
E45Q        GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC

            ....|....| ....|....| ....|....| ....|....| ....|....|
                   260        270        280        290        300
WT          TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT
E45Q        TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT

            ....|....| ....|....| ....|....| ....|....| ....|....|
                   310        320        330        340        350
WT          CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
E45Q        CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT

            ....|....| ....|....| ....|....| ....|....| ....|....|
                   360        370        380        390        400
WT          CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC
E45Q        CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC

            ....|....| ....|....| ....|....| ....|....| ....|....|
                   410        420        430        440        450
WT          AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
E45Q        AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG

            ....|....| ....|....| ....|....| ....|....| ....|....|
                   460        470        480        490        500
WT          GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT
E45Q        GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT

            ....|....| ....|....| ..
                   510        520
WT          TCTACGCCAC CTTGCCCAGC CC
E45Q        TCTACGCCAC CTTGCCCAGC CC


Pairwise Alignment
Sequence 1: WT
Sequence 2: H156F
Sequence ends allowed to slide over each other
Alignment score: 1038

Identities:   0.9961686


            ....|....| ....|....| ....|....| ....|....| ....|....|
                    10         20         30         40         50
WT          ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG
H156F       ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG

            ....|....| ....|....| ....|....| ....|....| ....|....|
                    60         70         80         90        100
WT          CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
H156F       CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
```

```
             ....|....| ....|....| ....|....| ....|....| ....|....|
                    110        120        130        140        150
WT           AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC
H156F        AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    160        170        180        190        200
WT           GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA
H156F        GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    210        220        230        240        250
WT           GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC
H156F        GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    260        270        280        290        300
WT           TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT
H156F        TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    310        320        330        340        350
WT           CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
H156F        CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    360        370        380        390        400
WT           CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC
H156F        CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC

             ....|....| ....|....| ....|....| ....|....| ....|....|
                    410        420        430        440        450
WT           AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
H156F        AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG


             ....|....| ....|....| ....|....| ....|....| ....|....|
                    460        470        480        490        500
WT           GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT
H156F        GTCCTGGTTG CCTCCTTTCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT

             ....|....| ....|....| ..
                    510        520
WT           TCTACGCCAC CTTGCCCAGC CC
H156F        TCTACGCCAC CTTGCCCAGC CC


Pairwise Alignment
Sequence 1: WT
Sequence 2: Q67V
Sequence ends allowed to slide over each other
Alignment score: 1038

Identities:   0.9961686
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....|
                         10         20         30         40         50
WT              ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG
Q67V            ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG

                ....|....| ....|....| ....|....| ....|....| ....|....|
                         60         70         80         90        100
WT              CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
Q67V            CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        110        120        130        140        150
WT              AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC
Q67V            AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        160        170        180        190        200
WT              GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA
Q67V            GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCGT

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        210        220        230        240        250
WT              GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC
Q67V            GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        260        270        280        290        300
WT              TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT
Q67V            TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        310        320        330        340        350
WT              CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
Q67V            CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        360        370        380        390        400
WT              CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC
Q67V            CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        410        420        430        440        450
WT              AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
Q67V            AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        460        470        480        490        500
WT              GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT
Q67V            GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT

                ....|....| ....|....| ..
                        510        520
WT              TCTACGCCAC CTTGCCCAGC CC
Q67V            TCTACGCCAC CTTGCCCAGC CC
```

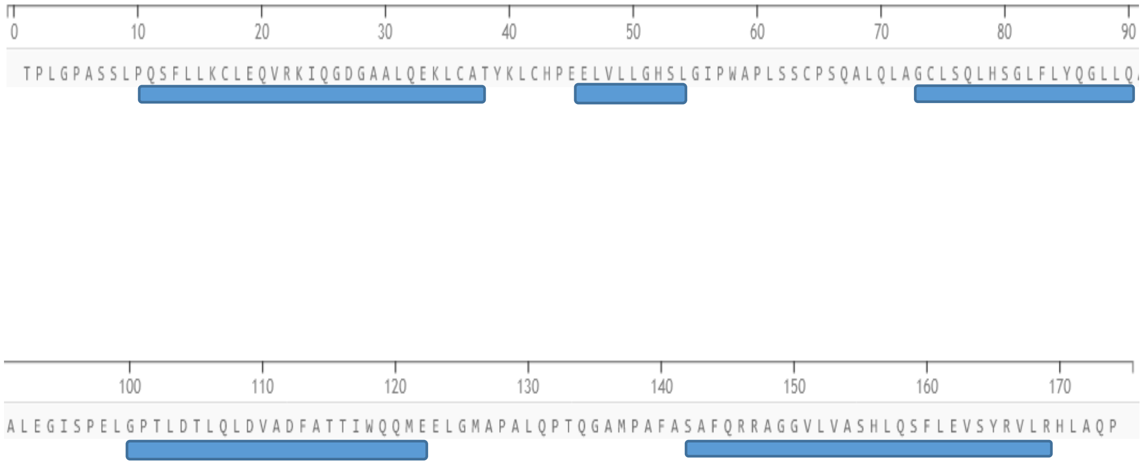
```
Pairwise Alignment
Sequence 1: WT
Sequence 2: P65V
Sequence ends allowed to slide over each other
Alignment score: 1035

Identities:   0.9942529

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     10         20         30         40         50
WT            ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG
P65V          ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     60         70         80         90        100
WT            CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
P65V          CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA

              ....|....| ....|....| ....|....| ....|....| ....|....|
                    110        120        130        140        150
WT            AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC
P65V          AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC

              ....|....| ....|....| ....|....| ....|....| ....|....|
                    160        170        180        190        200
WT            GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA
P65V          GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCGTGAGCCA

              ....|....| ....|....| ....|....| ....|....| ....|....|
                    210        220        230        240        250
WT            GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC
P65V          GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC

              ....|....| ....|....| ....|....| ....|....| ....|....|
                    260        270        280        290        300
WT            TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT
P65V          TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT

              ....|....| ....|....| ....|....| ....|....| ....|....|
                    310        320        330        340        350
WT            CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
P65V          CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT


              ....|....| ....|....| ....|....| ....|....| ....|....|
                    360        370        380        390        400
WT            CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC
P65V          CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC

              ....|....| ....|....| ....|....| ....|....| ....|....|
                    410        420        430        440        450
WT            AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
```

```
P65V          AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     460        470        480        490        500
WT            GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT
P65V          GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT

              ....|....| ....|....| ..
                     510        520
WT            TCTACGCCAC CTTGCCCAGC CC
P65V          TCTACGCCAC CTTGCCCAGC CC
Pairwise Alignment
Sequence 1: WT
Sequence 2: S8C_Q70C
Sequence ends allowed to slide over each other
Alignment score: 1027

Identities:   0.9904215


              ....|....| ....|....| ....|....| ....|....| ....|....|
                      10         20         30         40         50
WT            ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG
S8C_Q70C      ACACCCCTAG GCCCTGCC-G CTGCCTGCCC CAGAGCTTCC TGCTCAAGTG

              ....|....| ....|....| ....|....| ....|....| ....|....|
                      60         70         80         90        100
WT            CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
S8C_Q70C      CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     110        120        130        140        150
WT            AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC
S8C_Q70C      AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     160        170        180        190        200
WT            GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA
S8C_Q70C      GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     210        220        230        240        250
WT            GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC
S8C_Q70C      GGCCCTGTGC CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     260        270        280        290        300
WT            TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT
S8C_Q70C      TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     310        320        330        340        350
WT            CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
S8C_Q70C      CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
```

```
              ....|....| ....|....| ....|....| ....|....| ....|....|
                     360        370        380        390        400
WT            CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC
S8C_Q70C      CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     410        420        430        440        450
WT            AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
S8C_Q70C      AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     460        470        480        490        500
WT            GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT
S8C_Q70C      GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT


              ....|....| ....|....| ..
                     510        520
WT            TCTACGCCAC CTTGCCCAGC CC
S8C_Q70C      TCTACGCCAC CTTGCCCAGC CC

Pairwise Alignment
Sequence 1: WT
Sequence 2: S12E_Q134H
Sequence ends allowed to slide over each other
Alignment score: 1032

Identities:   0.9923372

Alignment: N:\\~out.tmp

              ....|....| ....|....| ....|....| ....|....| ....|....|
                      10         20         30         40         50
WT            ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGAGCTTCC TGCTCAAGTG
S12E_Q134H    ACACCCCTAG GCCCTGCCAG CTCCCTGCCC CAGGAATTCC TGCTCAAGTG

              ....|....| ....|....| ....|....| ....|....| ....|....|
                      60         70         80         90        100
WT            CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA
S12E_Q134H    CTTAGAGCAA GTGAGGAAGA TCCAGGGCGA TGGCGCAGCG CTCCAGGAGA

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     110        120        130        140        150
WT            AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC
S12E_Q134H    AGCTGTGTGC CACCTACAAG CTGTGCCACC CCGAGGAGCT GGTGCTGCTC

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     160        170        180        190        200
WT            GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA
S12E_Q134H    GGACACTCTC TGGGCATCCC CTGGGCTCCC CTGAGCAGCT GCCCCAGCCA

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     210        220        230        240        250
```

```
WT            GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC
S12E_Q134H    GGCCCTGCAG CTGGCAGGCT GCTTGAGCCA ACTCCATAGC GGCCTTTTCC


              ....|....| ....|....| ....|....| ....|....| ....|....|
                     260        270        280        290        300
WT            TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT
S12E_Q134H    TCTACCAGGG GCTCCTGCAG GCCCTGGAAG GGATCTCCCC CGAGTTGGGT

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     310        320        330        340        350
WT            CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT
S12E_Q134H    CCCACCTTGG ACACACTGCA GCTGGACGTC GCCGACTTTG CCACCACCAT

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     360        370        380        390        400
WT            CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC
S12E_Q134H    CTGGCAGCAG ATGGAAGAAC TGGGAATGGC CCCTGCCCTG CAGCCCACCC

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     410        420        430        440        450
WT            AGGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG
S12E_Q134H    ATGGTGCCAT GCCGGCCTTC GCCTCTGCTT TCCAGCGCCG GGCAGGAGGG

              ....|....| ....|....| ....|....| ....|....| ....|....|
                     460        470        480        490        500
WT            GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT
S12E_Q134H    GTCCTGGTTG CCTCCCATCT GCAGAGCTTC CTGGAGGTGT CGTACCGCGT

              ....|....| ....|....| ..
                     510        520
WT            TCTACGCCAC CTTGCCCAGC CC
S12E_Q134H    TCTACGCCAC CTTGCCCAGC CC
```

S.17. Mutant Primers. Sequences for WT and mutant primers are aligned, with mutated codon highlighted yellow and percentage alignment score titled “Identities”.


0
10
20
30
40
50
60
70
80
90
TPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLSQLHSGLFLYQGLLQ
100
110
120
130
140
150
160
170
ALEGISPELGPTLDTLQLDVADFATTIWQQMEELGMAPALQPTQGAMPAFASAFQRRAGGVLVASHLQSFLEVSYRVLRHLAQP