Investigating Protein Flexibility Using Molecular Dynamics Simulations of α-1 Acid Glycoprotein and Large-Scale Normal Mode Analysis of AlphaFold Models

Bhawna Dixit

Doctoral dissertation submitted to obtain the academic degrees of
Doctor of Biomedical Engineering (UGent) and Doctor of Bioengineering Sciences (VUB)

Supervisors
Prof. An Ghysels, PhD* - Prof. Wim Vranken, PhD**

* Department of Electronics and Information Systems
Faculty of Engineering and Architecture, Ghent University

** Faculty of Sciences and Bioengineering Sciences, Vrije Universiteit Brussel

February 2025

GHENT UNIVERSITY

VUB VRIJE UNIVERSITEIT BRUSSEL

ISBN 978-94-6355-958-4
NUR 928, 927
Wettelijk depot: D/2025/10.500/18

Members of the Examination Board

Chair

Prof. Joris Degroote, PhD, Ghent University

Other members entitled to vote

Prof. Karine Breckpot, PhD, Vrije Universiteit Brussel
Prof. Janez Konc, PhD, National Institute of Chemistry, Slovenia
Prof. Dominique Maes, PhD, Vrije Universiteit Brussel
Prof. Kathleen Marchal, PhD, Ghent University
Prof. Savvas Savvides, PhD, Ghent University
Prof. Vera Van Noort, PhD, KU Leuven

Supervisors

Prof. An Ghysels, PhD, Ghent University
Prof. Wim Vranken, PhD, Vrije Universiteit Brussel

Dedication

To my cherished grandparents, parents, and siblings,
For their unwavering love, support, and encouragement,
To my partner, family, and friends, whose belief in me
has been a constant source of strength,
This work is dedicated to you.

मेरे प्यारे दादा-दादी, नाना-नानी, माता-पिता और भाई-बहन को,
उनके अपार प्यार, समर्थन और प्रोत्साहन के लिए,
मेरे साथी, परिवार और दोस्तों को,
जिनका मुझ पर विश्वास मेरे लिए निरंतर शक्ति का स्रोत रहा है,
यह काम आपको समर्पित है ।

Acknowledgements

I have discovered profound parallels between biomolecules and myself, constantly navigating between two states, two possibilities, and the subtle nuances that lie in between—always existing in duality. Today, that duality is reflected in my life, where I have two countries I now call home, two universities (two HPC accounts), two incredible mentors, two inspiring research teams, and, in many ways, two lives that weave together. As I reflect on this journey, I realize that I have been fortunate to experience the best of both worlds. I want to express my heartfelt thanks to the many people who have been part of this experience—whether in moments of challenge or celebration.

First and foremost, I would like to extend my heartfelt thanks to my supervisors, **Prof. Wim Vranken** and **Prof. An Ghysels**, for their unwavering guidance, encouragement, and constant support throughout my doctoral journey. My sincere gratitude also goes to the members of my thesis committee, **Prof. Dominique Maes**, **Prof. Janez Konc**, **Prof. Savvas Savvides**, **Prof. Kathleen Marchal**, **Prof. Vera Van Noort**, and **Prof. Karine Breckpot**, whose insightful critiques, recommendations, and valuable feedback have been instrumental in shaping this thesis. I would also like to express my appreciation to **Prof. Joris Degroote**, the chair of my Ph.D. defense, for expertly managing the process and offering much-needed support.

I started my Ph.D. journey in 2019 at Vrije Universiteit Brussel, when Prof. Wim Vranken offered me the incredible opportunity to research proteins—an opportunity that would go on to change my life. I would like to express my deepest gratitude to **Prof. Wim Vranken**, for his invaluable guidance, expertise, and constant support throughout this journey. His positive attitude, constructive feedback, and passion for research have made a lasting impact on both my work and my growth as a researcher. A fun piece of trivia: Prof. Wim once shared with me that he had visited Mount Abu, Rajasthan—the very place where I was born but never actually lived. It is one of those delightful coincidences that will always make me smile!

On my first day at the Bio2Byte group at VUB, I met my first batchmates, **Jose** and **Joel**. I also got to know the interns **Rosan** and **Farhan**, as well as **Pathmanaban**, **Francois**, **Malika**, **Aitor**, and **Maite**, all of whom became an important part of my experience. I would like to thank Joel and Jose for their incredible support in helping me get familiar with Python. **Joel**, I will always cherish our after-work football games, even though I was not the best player, your humor, and your remarkable work ethic. **Jose**, I will fondly remember our conversations about food and the time we spent working together on the AlphaFold article. **Pathmanaban**, my friend and mentor from home, you have always been there when I needed you. Your dual life between UGent and VUB, India and Belgium, made everything feel more familiar and accessible. You are truly the longest MVP of Bio2Byte!

Six months later, the onset of the COVID-19 pandemic disrupted many plans. In 2020, I had the opportunity to join the UGent and BioMMeda group, under the guidance of Prof. An Ghysels. This marked the beginning of my journey as part of a joint Ph.D. program. I am deeply grateful to **Prof. An Ghysels** for her expertise, invaluable feedback, and insightful suggestions, all of which greatly contributed to the development of this thesis. Her guidance and encouragement were pivotal in navigating the complexities of my research and ignited my lasting passion for (bio)physics.

I found myself, almost unexpectedly, a bioinformatician surrounded by engineers. I vividly recall my first day—an introductory presentation filled with a whirlwind of questions from engineers (**Wouter**, **Samaneh**, **Sarah**, **Matthias**, **Amith**, **Tim**, **Annette**, **Lise**, **Carlos**, **Jurgen**, **Ghazal**, **Prof. Charlotte**, **Prof. Patrick**, and **Prof. Stefaan**), who, at that time, had little to no knowledge of proteins. But I enjoyed the questions regardless, and our beginning of forging new friendships. Though we didn't always understand each other's work, I hope this thesis brings you all a little closer to understanding proteins. Over time, I was welcomed into the group and surrounded by even more engineers from both

BioMMeda and Medisip, each group with its own hidden geniuses. I want to thank **Yousof**, **Melissa**, **Simeon**, **Jens**, **Jolan**, **Emma**, **Saar**, **Mohammed**, **Hooman**, **Marielle**, and **Milan**, for making my time at BioMMeda enjoyable and memorable with all the nice summer parties. And of course, a huge thanks to **Saskia**, for always keeping everything running smoothly and with a smile!

As time passed, and especially before and after COVID, I had the privilege of sharing some truly memorable moments with both old and new colleagues. I met a new generation of Ph.D. students, interns, and staff members at both Bio2Byte and BioMMeda. From BioMMeda, I would like to thank **Elias**, **Sina** from the molecular modelling group, and others (**Ariana**, **Jessie**, **Jellis** and others) who continuously supported me, even though our time together was brief.

From Bio2Byte, I would like to express my heartfelt gratitude to **David**, **Adrian**, **Konstantina**, **Charlotte**, **Sophie**, **Paolo**, **Danh**, and **Martain**. I will always look back with a smile on the fun times we had - our endless attempts to perfect the coffee machine, lunchtime philosophical discussions, and those drunk conversations. You made this journey unforgettable, and I am truly grateful for all the laughter and support. I would also like to thank the even newer generation with whom I had the pleasure of spending even a brief but impactful time, including **Alexandre**, **Alexander**, **Haochan**, **Golnar**, and **Romain**. I am deeply thankful for all the support Bio2Byte has given me - this PhD thesis would not have been possible without you (literally).

A special thank you goes to **Wouter** and **David**, who have been both my mentors and friends. I am deeply grateful for the way you shaped me into a better molecular modeller, shared your expertise on MD principles, and were always just a call away, right until the very end. A big thanks to **Adrian** for his endless technical support and for our fascinating conversations about photography. I also want to thank **Yousof** for always being there whenever I needed anything—your support has meant a lot.

Lastly, to settle a question many have asked: which team is best? Without a doubt, it is **Bio2Byte** even though it is difficult to pick one.

On a personal note, having never lived away from home, the thought of being 8000 miles apart from my family was initially daunting. However, their unwavering support and belief in me made it possible to embark on this Ph.D. journey. I would like to express my deepest gratitude to my mother (**Anita**), father (**Yogesh**), grandparents (**Barfi**, **Murti**, **Rameshwar Das**, late **Arjun Singh**), and my cherished siblings (**Ekta and Tarun**), brother-in-law (**Gaurav**), as well as my extended family and friends, and my bird (**Cheeni**), for their constant encouragement and love, which allowed me to pursue this dream with confidence. I would also like to thank my partner, **Mike** and his amazing family (**Marleen**, **Peter**, **Senne, Yarne, Elise, Robby, Mark, Tamara**), for making Belgium feel like home in ways I never expected and always being together in sickness and health. My dear friends from home **Nakul**, **Akshay**, **Gaurav**, **Gulshan**, and **Akhilesh**, **Mitalee**, **Manju**, **Kayenat**, and **Nimisha** for keeping up with me on endless video calls and hearing my troubles. My beautiful aunties and uncle (**Asha**, **Akshma**, **Sunita**, **Renu**, **Beena**, and **Vikas**), and my dear cousins, for giving me so much love and support.

Once again, I would like to honor the memory of the beloved family and friends I have lost over these years, who would have been proud to see this thesis come to fruition.

Table of contents

Table of contents **vii**

Abbreviations and symbols **xi**

Summary **xviii**

Samenvatting **xxiv**

1 Introduction **1**

1.1 Cells 1

1.2 Protein sequence and protein structure 3

1.3 Protein folding and dynamics 8

1.3.1 What is the folding code? 10

1.3.2 How do proteins fold rapidly? 12

1.3.3 Predicting native structure of a protein from a sequence 15

1.3.3.1 AlphaFold2: state-of-the-art 16

1.3.3.2 Limitations of AlphaFold2 in capturing protein dynamics 17

1.3.3.3 The advent of AlphaFold3 20

1.3.4 Dynamics and flexibility 21

1.3.4.1 The Functional Significance of Protein Dynamics 23

1.3.4.2 Investigating protein dynamics 24

1.4 Glycosylation 28

1.4.1 Overview of glycosylation 28

1.4.2 Biosynthesis and composition of N-glycans . . 31

1.4.3 3-D structure of glycans 34

1.4.4 Effect of glycans on protein conformation: dynamic and sticky molecular glue 37

2 Research questions **39**

2.1 AGP dynamics 40

2.2 Gradations of protein dynamics 42

3 Methodological background 45

3.1 MD Simulations 45

3.1.1 The algorithm behind MD 46

3.1.2 The force field 47

3.1.3 Periodic boundary conditions 50

3.1.4 Statistical mechanics in MD 50

3.2 Computational and experimental metrics of flexibility 54

3.2.1 Fluctuations from MD simulations 55

3.2.2 Backbone dynamics from DynaMine and ShiftCrypt 57

3.2.2.1 Chemical shifts 57

3.2.2.2 RCI and S^2_{RCI} 59

3.2.2.3 DynaMine 62

3.2.3 Normal mode analysis 63

3.2.3.1 Steps involved in NMA 63

3.2.3.2 Theory of NMA 63

3.2.3.3 Computational parameters in the WEBnma tool 66

3.2.3.4 Computing RMSF from Normal modes 67

3.3 Simulating the AGP protein 68

3.3.1 Simulated systems of AGP 68

3.3.2 Computational details 68

3.3.3 Convergence in glycoprotein MD simulations . 69

4 Effect of glycosylation and mutation on conformational dynamics of AGP 71

4.1 Background and methodology 71

4.2 Contributions 72

4.3 Concluding remarks 73

5 Gradations of protein dynamics: AlphaFold2 vs NMR 101

5.1 Background and methodology 101

5.2 Contributions 102

5.3 Concluding remarks 103

6 Conformation and dynamics of proteins with ambiguous behavior 127

6.1 Background and methodology 127

6.2 Contributions 128

6.3 Concluding remarks 128

7 Conclusion and perspectives 147
7.1 Effect of glycosylation on protein dynamics 147
7.2 Combined effect of glycosylation and mutations on protein's conformational dynamics 148
7.3 Predicting the effect of glycans on protein's backbone dynamics 148
7.4 Convoluting experimental and computational metrics of protein flexibility 149
7.5 AlphaFold vs NMR: Implications for capturing protein dynamics 150
7.6 NMA: strengths and shortcomings in capturing protein flexibility and dynamics 151
7.7 Current limitations 152
7.8 Future perspectives 154

Bibliography 157

Appendices 177
Supporting information: AGP dynamics 177
Supporting information: Gradations of protein dynamics . . 216

List of Publications 254

Abbreviations and symbols

The following list summarizes the most commonly used abbreviations and symbols in this dissertation.

Abbreviations

2-D	Two-dimensional
3-D	Three-dimensional
AFM	Atomic Force Microscopy
AGP	α-1 acid Glycoprotein
CASP	Critical Assessment of techniques for protein Structure Prediction
CS	Chemical Shifts
DNA	Deoxyribonucleic acid
Dol-P	Dolichol-phosphate
EM	Electron Microscopy
ENM	Elastic Network Model
Fuc	Fucose
IDP	Intrinsically Disordered Protein
LLM	Large Language Model
LJ	Lennard-Jones
MD	Molecular Dynamics
mRNA	messenger Ribonucleic Acid
MSA	Multiple Sequence Alignment
NMA	Normal Mode Analysis
NMR	Nuclear Magnetic Resonance
PBC	Periodic Boundary Conditions
PCA	Principal Component Analysis
PDB	Protein Data Bank
PES	Potential energy surface
PTM	Post-translational Modification
RCI	Random Coil Index
RMSD	Root mean square deviation

RMSF	Root mean square fluctuation
RNase	Ribonuclease
SNFG	Symbol Nomenclature for Glycans
smFRET	single-molecule Fluorescence Resonance Energy Transfer
vdW	van der Waals

Symbols

C_α	alpha-carbon
C_β	beta-carbon
μm	micrometer
Å	angstrom
nm	nanometer
s	second
ns	nanosecond
μs	microsecond
ps	picosecond
fs	femtosecond
ϕ	phi torsion angle
ψ	psi torsion angle
ω	omega torsion angle
χ_n	chi as protein side chain torsion angle

Summary

This thesis delves into the concept of protein flexibility, a dynamic and multifaceted phenomenon encompassing a spectrum ranging from complete structural disorder to the motion of protein fragments, with various intermediate conformational states. As inherently complex dynamic entities, proteins exhibit conformational flexibility that allows them to perform a vast array of biological functions. Shaped by factors such as temperature, forces, and vibrations, proteins undergo continuous conformational changes, with their atomic coordinates fluctuating over time due to thermal motion. Despite this, traditional models of protein behaviour have primarily relied on static representations, which, while useful, are simplifications that fail to capture the inherent dynamics of proteins.

This flexibility, driven by the laws of physics and chemistry, is crucial for understanding protein behaviour, especially when considering the physiological context—such as post-translational modifications (PTMs), which are often ignored or overlooked. This thesis adopts two complementary perspectives to investigate protein conformational flexibility and dynamics: a "close-up" view and a "panoramic" view. The close-up perspective focuses on AGP and its mutants, examining the impact of glycosylation on local and global conformational changes, protein backbone flexibility, and glycan dynamics through Molecular Dynamics (MD) simulations. The panoramic view aims to integrate computational predictions of protein flexibility, such as AlphaFold2's pLDDT scores, with experimental data from techniques including Nuclear Magnetic Resonance (NMR) spectroscopy, MD simulations, and Normal Mode Analysis (NMA). By merging computational approaches and experimental metrics of flexibility, this thesis seeks to provide a more nuanced understanding of protein dynamics, addressing the challenges of capturing flexibility on a large scale. Ultimately, this work aims to advance our understanding of protein behaviour,

moving beyond static models to a more comprehensive, dynamic framework.

The introductory **Chapter 1** provides an overview of proteins and glycans, focusing on their synthesis, structure, and dynamics. The summary is divided into two parts:
For **proteins**, the chapter offers a detailed physicochemical description of protein 3-D structures and the principles behind protein folding. A key focus is the folding funnel hypothesis, which explains how proteins navigate unique energy landscapes to achieve their native conformations. The rugged landscape model is discussed in this chapter as the most realistic framework for understanding protein folding, describing a slow, multi-step folding process shaped by kinetic and thermodynamic factors. It illustrates how each protein sequence navigates a unique energy landscape toward its lowest energy or native state. The chapter concludes by describing proteins as ensembles of stable conformers in equilibrium, sensitive to environmental factors like temperature, pH, denaturants, PTMs, and ligand binding. These factors can induce conformational changes, which in turn impact protein stability and function. The chapter emphasizes the unresolved gaps in understanding protein folding and the limitations of tools like AlphaFold2 in capturing protein dynamics and the impact of PTMs, with a focus on glycosylation. It concludes by discussing protein dynamics, ranging from rapid local vibrations to slower, large-scale motions, which are crucial for protein function. The chapter also highlights computational and experimental techniques, along with metrics for assessing protein dynamics and flexibility, obtained through MD simulations, NMA, and NMR spectroscopy.

For **glycans**, the chapter begins with an overview of glycosylation, the process by which glycans covalently attach to proteins, influencing their stability, function, and chemical composition. It then focuses on N-glycosylation, detailing the biosynthesis and composition of N-glycans, and emphasizing variations in glycan structures due to enzymatic modifications. The chapter briefly discusses the role of glycans in diseases, particularly how aberrant glycosylation contributes to conditions like cancer. Next, the chapter explains the 3-D structure of glycans, highlighting the role of glycosidic linkages and the different conformations glycans can adopt. It introduces carbohydrate-Ramachandran (carb-RAMA) plots to explain their flexibility and steric constraints, which illustrate the distribution of glycosidic torsion angles. The chapter also discusses the challenges associated with glycan flexibility and current experimental and computational approaches, such as NMR and mass spectrometry, for resolving glycan structures, as well as MD force fields for probing glycan dynamics. The chapter concludes with a discussion of glycans' role as molecular glue in affecting protein dynamics. By integrating these various aspects, the chapter provides a comprehensive understanding of glycans' biological importance and their complex interplay with proteins.

Chapter 2 presents the key research questions aimed at uncovering protein flexibility and dynamics. It also outlines the objectives and goals of the thesis, addressing these questions, and offers a brief overview of the methodology employed to investigate them.

Chapter 3 provides a comprehensive overview of the methodologies used to investigate protein dynamics and flexibility, in this thesis. It begins with a detailed discussion of MD simulations, explaining the underlying algorithm, force fields, periodic boundary conditions, and the role of statistical mechanics in MD. The chapter then examines both computational and experimental metrics used in this thesis to assess protein flexibility, including fluctuations from MD simulations, NMA, and backbone dynamics analyzed with tools like DynaMine and ShiftCrypt based on chemical shifts from NMR. Additionally, the chapter outlines the theoretical foundations of chemical shifts, RCI, and S^2_{RCI} for measuring flexibility, with a focus on their application in guiding fast-scale predictions and calculating protein flexibility. The chapter concludes with a theoretical description NMA, detailing its steps, computational parameters used in the WEBnma tool, and calculation of fluctuations from normal modes. Finally, the chapter describes the simulation settings and parameters applied in this thesis. The goal of this chapter is to provide a clear understanding of the methods applied in the thesis and ensure the reproducibility of the methodologies of the results discussed in the subsequent chapters.

Chapter 4 explores how glycosylation and mutations impact the conformational dynamics of AGP. AGP plays a crucial role in modulating immune response and inflammation, serves as a potential biomarker for various diseases, including cancer, and has a diverse drug-binding spectrum that influences the pharmacokinetics

of various drugs, making it a therapeutically significant protein of interest. To investigate the impact of glycosylation and mutations, triplicate MD simulations of AGP and eight mutants (with and without glycans) are used to examine backbone flexibility and solvent accessibility, in the context of cancer and drug design. To determine which regions of AGP experience conformational changes due to glycosylation and mutations, and how these changes affect its accessibility, three structural regions of AGP are defined based on its lipocalin fold: the open-end ligand binding-site entrance (LBE), central ligand binding site (LBS), and a hypothetical protein-protein interaction site (hPPI). The impact of glycosylation on solvent accessibility is investigated using probes of varying sizes. Principal Component Analysis (PCA) assesses backbone dynamics, and carb-RAMA plots explore glycan dynamics. The study finds that glycosylation reduces flexibility at the glycosylation site while increasing it in distant regions. Mutations affect local flexibility and induce long-range conformational changes. Glycosylated mutants show similar backbone dynamics, but no consistent biophysical outcomes. A mutation near glycosylation sites influences glycan-protein interactions and modulates both glycan and protein dynamics, affecting AGP's flexibility and solvent accessibility.

Chapter 5 investigates the relationship between protein flexibility of AlphaFold2 models and their NMR ensembles, predicted by computational methods (NMA and MD simulations) and observed through experimental techniques (NMR). For this study, three datasets of AlphaFold2 models are created, including flexibility data from ShiftCrypt, S^2_{RCI}, S^2, and RMSF from MD simulations. NMA was performed on these models, with additional analysis on NMR ensembles for the S^2_{RCI} dataset. Fluctuations were calculated from the lowest eigenvalue normal modes, and Pearson correlations were used to analyze relationships between flexibility metrics at the residue and protein levels. The study reveals that flexibility metrics align well for rigid residues with a single, stable conformation, contrasting with residues that exhibit dynamic behaviour and multiple conformations. This distinction between ordered and disordered regions is clear in the correlations between the parameters, but becomes less evident when examining dynamic residues. The study concludes that the dynamic variations observed by NMR in flexible protein regions are not fully captured by these computational approaches.

Chapter 6 consists of my contribution as a co-author highlighting the effect of PTMs on protein dynamics and conformational behaviour. The study illustrates and addresses the challenges of representing protein biophysical behaviour and dynamics, focusing on conformational states and their functional relevance. The study categorizes protein conformational behaviour into three classes: ordered, disordered, and ambiguous. By analyzing three distinct datasets using interpretable machine learning techniques, the study compares AlphaFold2 features with sequence-based predictions to explore their similarities and differences. Ultimately, it emphasizes the need to explore beyond the traditional two-state model (ordered vs. disordered), advocating for a more nuanced, probabilistic approach that recognizes proteins' dynamic nature and their potential to adopt various conformational states.

Chapter 7 concludes this thesis by summarizing the key findings and contributions of the research, reflecting on its significance, and suggesting potential directions for future research to further advance the understanding of protein dynamics and their functional implications.

Samenvatting

Deze thesis gaat dieper in op het concept van eiwitflexibiliteit, een dynamisch en veelzijdig fenomeen dat een reikwijdte heeft gaande van volledige structurele wanorde tot de beweging van eiwitfragmenten, met verschillende tussenliggende conformaties. Als intrinsiek complexe en dynamische entiteiten vertonen eiwitten conformatiemobiliteit die hen in staat stelt een breed scala aan biologische functies uit te voeren. Beïnvloed door factoren zoals temperatuur, krachten en trillingen, ondergaan eiwitten continue conformationele veranderingen, waarbij hun atomaire coördinaten fluctueren als gevolg van thermische beweging. Ondanks dit zijn traditionele modellen van eiwitgedrag voornamelijk gebaseerd op statische representaties, die hoewel nuttig, vereenvoudigingen zijn die de inherente dynamiek van eiwitten niet vastleggen.
Deze flexibiliteit, gedreven door de wetten van de fysica en scheikunde, is cruciaal voor het begrijpen van het gedrag van eiwitten, vooral in de fysiologische context—zoals post-translationele modificaties (PTM's), die vaak worden genegeerd of over het hoofd gezien. Deze thesis hanteert twee complementaire perspectieven om de conformationele flexibiliteit en dynamiek van eiwitten te onderzoeken: een "close-up perspectief en een "panoramisch perspectief. Het close-up perspectief richt zich op AGP en zijn mutanten, waarbij de impact van glycosylering op lokale en globale conformatiemutaties, eiwit-ruggengraatflexibiliteit en glycaandynamiek worden onderzocht via moleculaire dynamica simulaties (MD). Het panoramische perspectief beoogt het integreren van computationele voorspellingen van eiwitflexibiliteit, zoals de pLDDT-scores van AlphaFold2, met experimentele gegevens van technieken zoals Nucleaire Magnetische Resonantie (NMR) spectroscopie, MD-simulaties en Normal Mode Analyis (NMA). Door computationele benaderingen en experimentele metrieken van flexibiliteit te combineren, probeert deze thesis een meer genuanceerd begrip van eiwitdynamiek te bieden, waarbij de uitdagingen van

het waarnemen van flexibiliteit op grote schaal worden aangepakt. Uiteindelijk is het doel van dit werk het bevorderen van ons begrip van het gedrag van eiwitten, door statistische modellen te overstijgen en naar een meer uitgebreide, dynamische benadering te gaan.

Het inleidende **Hoofdstuk 1** biedt een overzicht van eiwitten en glycanen, met de nadruk op hun synthese, structuur en dynamiek. De samenvatting is verdeeld in twee delen:
Voor **eiwitten** biedt het hoofdstuk een gedetailleerde fysisch-chemische beschrijving van de 3D-structuren van eiwitten en de principes achter eiwitvouwing. Een belangrijk aandachtspunt is de vouwingstheorie, die uitlegt hoe eiwitten unieke energielandschappen doorlopen om hun native conformaties te bereiken. Het robuuste "landschapsmodel wordt in dit hoofdstuk besproken als het meest realistische kader om eiwitvouwing te begrijpen, en beschrijft een langzaam, meerstaps vouwingproces dat wordt beïnvloed door kinetische en thermodynamische factoren. Het illustreert hoe elke eiwitsequentie een uniek energielandschap doorloopt naar de laagste energie- of native staat. Het hoofdstuk eindigt met de beschrijving van eiwitten als ensembles van stabiele conformers in evenwicht, gevoelig voor omgevingsfactoren zoals temperatuur, pH, denaturanten, PTM's en ligandbinding. Deze factoren kunnen conformationele veranderingen induceren, die op hun beurt de stabiliteit en functie van eiwitten beïnvloeden. Het hoofdstuk benadrukt de onopgeloste lacunes in het begrip van eiwitvouwing en de beperkingen van tools zoals AlphaFold2 bij het vastleggen van eiwitdynamiek en de invloed van PTM's, met een focus op glycosylering. Het wordt afgesloten met een bespreking van eiwitdynamiek, variërend van snelle lokale trillingen tot langzamere, grootschalige bewegingen, die cruciaal zijn voor de functie van eiwitten. Het hoofdstuk benadrukt ook computationele en experimentele technieken, samen met metriek voor het evalueren van eiwitdynamiek en flexibiliteit, verkregen via MD-simulaties, NMA en NMR-spectroscopie.
Voor **glycanen** begint het hoofdstuk met een overzicht van glycosylering, het proces waarbij glycanen covalent aan eiwitten worden gehecht, waardoor ze de stabiliteit, functie en chemische samenstelling van eiwitten beïnvloeden. Vervolgens wordt de focus gelegd op N-glycosylering, met details over de biosynthese en samenstelling van N-glycanen, en de nadruk op variaties in glycanstructuren door enzymatische modificaties. Het hoofdstuk bespreekt kort de rol van glycanen bij ziektes, vooral hoe abnormale glycosylering bijdraagt aan aandoeningen zoals kanker. Daarna

wordt de 3D-structuur van glycanen uitgelegd, met de nadruk op de rol van glycosidische verbindingen en de verschillende conformaties die glycanen kunnen aannemen. "Carbohydraat-Ramachandran (carb-RAMA) plots worden geïntroduceerd om hun flexibiliteit en sterische beperkingen uit te leggen, die de verdeling van glycosidische torsiehoeken illustreren. Het hoofdstuk bespreekt ook de uitdagingen die gepaard gaan met glycaanflexibiliteit en de huidige experimentele en computationele benaderingen, zoals NMR en massaspectrometrie, voor het oplossen van glycanstructuren, evenals MD-krachtvelden voor het onderzoeken van glycaandynamiek. Het hoofdstuk eindigt met een discussie over de rol van glycanen als moleculaire lijm bij het beïnvloeden van eiwitdynamiek. Door deze verschillende aspecten te integreren, biedt het hoofdstuk een volledig begrip van de biologische betekenis van glycanen en hun complexe interactie met eiwitten.

Hoofdstuk 2 presenteert de belangrijkste onderzoeksvragen die gericht zijn op het ontdekken van eiwitflexibiliteit en dynamiek. Het schetst ook de doelstellingen en doelen van de thesis, die deze vragen bespreekt, en biedt een kort overzicht van de gebruikte methodologie om deze te onderzoeken.

Hoofdstuk 3 biedt een uitgebreid overzicht van de gebruikte methodologieën voor het onderzoeken van eiwitdynamiek en flexibiliteit in deze thesis. Het begint met een gedetailleerde bespreking van MD-simulaties, waarbij het onderliggende algoritme, krachtvelden, periodieke randvoorwaarden en de rol van statistische mechanica in MD worden uitgelegd. Het hoofdstuk bespreekt vervolgens zowel computationele als experimentele metriek die in deze thesis zijn gebruikt om eiwitflexibiliteit te beoordelen, inclusief fluctuaties uit MD-simulaties, NMA en dynamiek van de eiwitruggengraat geanalyseerd met tools zoals DynaMine en ShiftCrypt op basis van chemische verschuivingen uit NMR. Daarnaast bespreekt het hoofdstuk de theoretische basis van chemische verschuivingen, RCI en S^2_{RCI} voor het meten van flexibiliteit, met de nadruk op hun toepassing in het sturen van voorspellingen op snelle schaal en het berekenen van eiwitflexibiliteit. Het hoofdstuk eindigt met een theoretische beschrijving van NMA, waarin de stappen, computationele parameters die in de WEBnma-tool worden gebruikt, en de berekening van fluctuaties van normale modi worden beschreven. Tot slot worden de simulatie-instellingen en parameters die in deze thesis zijn toegepast beschreven. Het doel van dit hoofdstuk is om een duidelijk begrip te bieden van de toegepaste methoden in de thesis en de reproduceerbaarheid van de methodologieën van de besproken resultaten te waarborgen.

Hoofdstuk 4 onderzoekt hoe glycosylering en mutaties de conformatiemutaties van AGP beïnvloeden. AGP speelt een cruciale rol bij het moduleren van immuunrespons en ontsteking, fungeert als een potentieel biomarker voor verschillende ziektes, waaronder kanker, en heeft een breed spectrum van geneesmiddelbinding dat de farmacokinetiek van verschillende geneesmiddelen beïnvloedt, wat het een therapeutisch belangrijke eiwit maakt. Om de impact van glycosylering en mutaties te onderzoeken, worden triplicaten MD-simulaties van AGP en acht mutanten (met en zonder glycanen) gebruikt om de ruggengraatflexibiliteit en oplosbaarheidstoegankelijkheid te onderzoeken in de context van kanker en geneesmiddelenontwerp. Om te bepalen welke gebieden van AGP conformatiemutaties ondergaan door glycosylering en mutaties, en hoe deze veranderingen de toegankelijkheid beïnvloeden, worden drie structurele regio's van AGP gedefinieerd op basis van zijn lipocalinevouwing: de open ligandenbindingsplaats (LBE), de centrale ligandenbindingsplaats (LBS) en een hypothetische eiwit-eiwitinteractiesite (hPPI). De impact van glycosylering op oplosbaarheidstoegankelijkheid wordt onderzocht met behulp van proeven van verschillende groottes. Hoofcomponentenanalyse (PCA) beoordeelt de dynamiek van de ruggengraat en carb-RAMA-plots verkennen de glycaandynamiek. De studie toont aan dat glycosylering de flexibiliteit vermindert op de glycosyleringsplaats, terwijl het de flexibiliteit in verre gebieden vergroot. Mutaties beïnvloeden de lokale flexibiliteit en veroorzaken lange-afstandsconformatieveranderingen. Geglycosileerde mutanten vertonen vergelijkbare ruggengraaddynamiek, maar geen consistente biophysische uitkomsten. Een mutatie nabij glycosyleringsplaatsen beïnvloedt de interacties tussen glycanen en eiwitten en moduleert zowel de glycaandynamiek als de eiwitdynamiek, wat de flexibiliteit en oplosbaarheidstoegankelijkheid van AGP beïnvloedt.

Hoofdstuk 5 onderzoekt de relatie tussen de eiwitflexibiliteit van AlphaFold2-modellen en hun NMR-ensembles, voorspeld door computationele methoden (NMA en MD-simulaties) en waargenomen door experimentele technieken (NMR). Voor deze studie worden drie datasets van AlphaFold2-modellen gemaakt, inclusief flexibiliteitsgegevens van ShiftCrypt, S^2_{RCI}, S^2, en RMSF van MD-simulaties. NMA werd uitgevoerd op deze modellen, met aanvullende analyses op NMR-ensembles voor de S^2_{RCI}-dataset. Fluctuaties werden berekend uit de laagste eigenwaarde normale modi, en Pearson-correlaties werden gebruikt om de relaties tussen flexibiliteitsmetingen op residu- en eiwitniveau te analyseren. De studie onthult dat flexibiliteitsmetingen goed overeenkomen voor stijve residuen met een enkele, stabiele conformatie, in contrast met residuen die dynamisch gedrag vertonen en meerdere conformaties hebben. Dit onderscheid tussen geordende en ongeordende gebieden is duidelijk in de correlaties tussen de parameters, maar wordt minder evident bij het onderzoeken van dynamische residuen. De studie concludeert dat de dynamische variaties die door NMR in flexibele eiwitgebieden worden waargenomen, niet volledig worden vastgelegd door deze computationele benaderingen.

Hoofdstuk 6 bestaat uit mijn bijdrage als co-auteur die de invloed van PTM's op eiwitdynamiek en conformationeel gedrag benadrukt. Het onderzoek illustreert en behandelt de uitdagingen van het representeren van eiwitbiophysisch gedrag en dynamiek, met de nadruk op conformationele toestanden en hun functionele relevantie. Het onderzoek categoriseert het conformationele gedrag van eiwitten in drie klassen: geordend, ongeordend en ambigu. Door drie verschillende datasets te analyseren met interpreteerbare machine learning-technieken, vergelijkt het onderzoek de kenmerken van AlphaFold2 met sequentiegebaseerde voorspellingen om hun overeenkomsten en verschillen te verkennen. Uiteindelijk benadrukt het de noodzaak om verder te kijken dan het traditionele twee-toestandenmodel (geordend versus ongeordend) en pleit het voor een meer genuanceerde, probabilistische benadering die de dynamische aard van eiwitten erkent en hun vermogen om verschillende conformationele toestanden aan te nemen.

Hoofdstuk 7 sluit deze thesis af met een samenvatting van de belangrijkste bevindingen en bijdragen van het onderzoek, reflecteert op het belang ervan en suggereert mogelijke richtingen voor toekomstig onderzoek om het begrip van eiwitdynamiek en hun functionele implicaties verder te verdiepen.

Chapter 1

Introduction

This chapter provides a comprehensive overview of proteins and glycans. For proteins, it begins with a description of cells, followed by protein synthesis—from DNA transcription to amino acid sequence formation and the development of three-dimensional (3-D) structure—along with a discussion of their physicochemical and biophysical properties. The general protein folding problem is explained in detail, with a particular focus AlphaFold2, its relation to sequence-to-structure prediction and its limitations in capturing protein dynamics. This chapter lays the groundwork for investigating protein flexibility at a broader scale, encompassing a diverse range of proteins. Subsequently, protein dynamics and methods of investigating protein dynamics are also discussed in detail. Moving on to glycans, the chapter covers the current state-of-the-art research, including their role in glycosylation, biosynthesis, and disease. It also examines the structure of glycans, experimental techniques for their analysis, and their dynamic behaviours. This chapter lays the groundwork for understanding glycosylation, glycan biosynthesis, structure, and flexibility, which will be further explored in subsequent chapters, including a detailed analysis of α-1-acid glycoprotein (AGP) and glycoprotein flexibility.

1.1 Cells

Cells, the fundamental units of life, exhibit a high level of structural complexity, containing specialized organelles responsible for distinct cellular functions (Fig. 1.1, a). The nucleus contains the cell's genetic information (Fig. 1.1, b), including chromosomes that regulate cellular activity by directing the expression of genes and the production of proteins (Fig. 1.1, c). These molecular entities serve multifaceted roles as guardians, regulators, promoters, competitors, inhibitors, assemblers, and more, collectively orchestrating the dynamic and intricate biological processes within the cellular world. Notably, proteins emerge as the predominant macromolecular constituents in living cells forming a molecular crowd, comprising more than 50% of a cell's dry weight (Fig 1.2) [1]. Molecular crowding refers to the phenomenon where variety of metabolites—molecules that serve as intermediates in metabolic processes, such as nucleic acids, proteins, carbohydrates, phosphates, alcohols, vitamins, and other cofactors—along with ions, are present at significant concentrations [2]. This crowded environment within cells is a characteristic of biological systems. The crowded environment within cells is known to affect the dynamic properties of proteins, including their conformational dynamics. The synthesis of proteins within cells is governed by the central dogma of molecular biology [3], which outlines the flow of genetic information. This process involves several key steps: the replication of deoxyribonucleic acid (DNA), the transcription of DNA into messenger ribonucleic acid (mRNA), and the translation of mRNA into a sequence of amino acids (Fig. 1.1, d). DNA functions as the repository of genetic information. It consists of two strands forming a double helix, with each strand composed of a sequence of nitrogenous bases: adenine (A), thymine (T), cytosine (C), and guanine (G), a sugar molecule (deoxyribose), and phosphate groups [4]. The sequence of these bases encodes the genetic instructions necessary for the synthesis of proteins. In contrast to DNA, RNA is typically single-stranded. Its backbone consists of alternating phosphate groups and ribose sugars, as opposed to the deoxyribose found in DNA. Each ribose sugar is paired with one of four nitrogenous bases: adenine, uracil (U), cytosine, or guanine. Within cells, various types of RNA are present, including mRNA, ribosomal RNA (rRNA), and transfer RNA (tRNA). The role of mRNA is to transport genetic information from the DNA in the cell's nucleus to the cytoplasm, where the protein synthesis machinery interprets the mRNA sequence and translates each three-base codon into its corresponding amino acid, forming a polypeptide. This process is known as translation [3]. A codon is a sequence of three bases which encodes a single amino acid. For instance, the codon CAG encodes the amino acid glutamine. A polypeptide can serve various functions, such as providing structural support, acting as an enzyme, or interacting with other polypeptides to form more complex proteins.

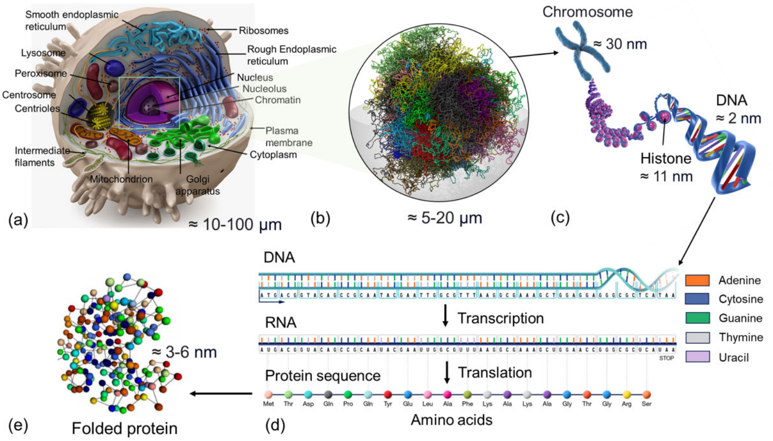


Figure 1.1: From cell to protein Overview of (a) a 3-D structure of a eukaryotic cell, (b) nucleus, and (c) a eukaryotic chromosome model. (d) Central dogma of molecular biology. (e) Folded protein structure, with amino acids as nodes and edges. Average sizes of molecular structures range from μm to nm length scales. The images from a to c are adapted from Ref.[5], and d is sourced from [6].

1.2 Protein sequence and protein structure

Proteins are composed of 20 standard amino acids (Fig. 1.1) [7]. These amino acids are analogous to linguistic alphabets, connected sequentially to form simple words or oligopeptides ($\leq$ 20 amino acids), complex phrases or polypeptides ($\leq$ 50 amino acids) and complete stories or proteins ($>$ 50 amino acids). Each amino acid apart from glycine is chemically composed of a central, asymmetric alpha carbon (C_α), bonded to an amino group ($-$NH2), a carboxyl group ($-$COOH), a hydrogen ($-$H), and a distinctive side chain group ($-$R) [8]. In glycine, the H atom is the R group. The side chain group differentiates the phonetics, shape, and size of the alphabets; these factors determine the biophysical and physico-chemical properties of amino acids, which dictate whether an amino acid is acidic, basic, polar, or nonpolar [8] (Fig. 1.3). These properties enable proteins to fold into specific 3-D shapes or folds, allowing them to perform particular functions. In a protein, these amino acids are linked together via peptide bonds, which are covalent linkages established between the carboxyl group of one amino acid and the amino group of another

(Fig. 1.2). During the peptide bond formation, the removal of water occurs, leaving behind what is referred to as an "amino acid residue" for each constituent amino acid [8]. The standard amino acids are denoted by a single uppercase letter or a three-letter abbreviation; for example, glycine is represented as G or Gly. [9] (Fig. 1.3). However, in this thesis, 'Gly' is reserved for glycosylation to avoid ambiguity with glycine.

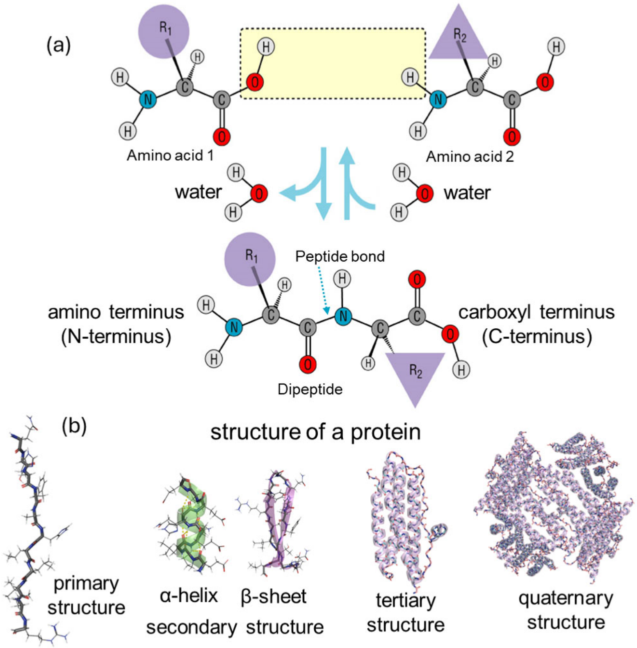


Figure 1.2: (a) Formation of a peptide bond between two amino acids resulting in loss of water and formation of a dipeptide. The image is adapted from Ref.[10]. (b) Protein structure classification: primary structure of a protein depicted by an oligopeptide containing several amino acids, folded into secondary structure elements such as an α-helix, and a β-sheet. Fully folded, tertiary and quaternary structure of protein composed of secondary structure elements depicted by backbone amino acid residues. The image is generated in PyMol.

To build an understanding of the structure and function of proteins, we start with the primary structure, which is the sequential arrangement of amino acids linked by peptide and disulfide bonds to form a polypeptide. The primary structure requires discerning the relationship between nearest neighbor amino acids [11]. The primary structure holds necessary information for dictating the protein's function (Fig. 1.2) (c). It consists of a "backbone" or main chain, incorporating N, C_α, and C atoms from each amino acid in the sequence. The side chains, determined by the identity of each amino acid, diverge from this protein backbone. Due to the thermal motion and kinetic energy of the atoms, both the backbone and, the side chains undergo constant movements [12]. The backbone consists of semi-rigid peptide planes linked at C_α atoms [13]. The angle of the peptide bond which forms the plane is known as ω which is represented by $C_\alpha{-}C{-}N{-}C_\alpha$. These planes maintain a degree of rigidity due to special electron sharing among the atoms within the plane [14] (Fig. 1.4). Thus, it is reasonable to assume the $C{-}N$ dihedral angle in the middle of the plane is fixed in a flat conformation [14]. The only remaining degrees of freedom are two dihedral angles per amino acid residue, named phi (ϕ) and psi (ψ) [15]. The ϕ dihedral angle represents the angle between $C{-}N{-}C_\alpha{-}C$, while ψ is the dihedral angle between $N{-}C_\alpha{-}C{-}N$. The side chain dihedral angles are denoted as χ_n, (chi, where n = 1, 2, 3, 4) angles are measured along $N{-}C_\alpha{-}C_\beta{-}X$ axis, until the side-chain branch, where X represents any non-hydrogen atom present on the side chain [16] (Fig. 1.4). In principle, the rotations around ϕ and ψ could take on any values. However, due to "steric" hinderances or van der Waals (vdW) clashes among the atoms, ϕ and ψ values are constrained within statistically observed favorable or allowed regions described by a Ramachandran plot [15]. The regions are deemed "allowed" as, under the consideration of standard vdW radii for peptide atoms which do not lead to collisions. The rotations around $C{-}N$ bonds enable preferred spatial arrangements with minimal steric clashes between atoms, also known as conformations [17]. The coplanarity of peptide bonds allows for only two different conformations: trans and cis. In the trans conformation ($\omega \cong \pm 180°$), the alpha carbons (C_α) are positioned on opposite sides of the $C{-}N$ bond. In the cis conformation ($\omega \cong \pm 0°$), the alpha carbons are on the same side of the $C{-}N$ bond [17].

The backbone typically adopts 3-D folding patterns of three main types of secondary structures formed through hydrogen bonds (H−bonds) between the O and N atoms: α-helices, β-strands, loops or turns (Fig. 1.4, b) [19]. The secondary structure of protein can be explained in terms of bond distances, bond angles, semi-rigid peptide planes, restricted rotation, and non-covalent bonds [11]. These secondary structures result from the polypeptide backbone adopting ϕ and ψ dihedral angles, which lead to the formation of regular, repetitive patterns [19]. Certain regions of the protein chain do not adopt regular secondary structures or exhibit consistent H-bonding patterns. These regions, referred to as random coils [20], can be found in two main locations within proteins: at the protein

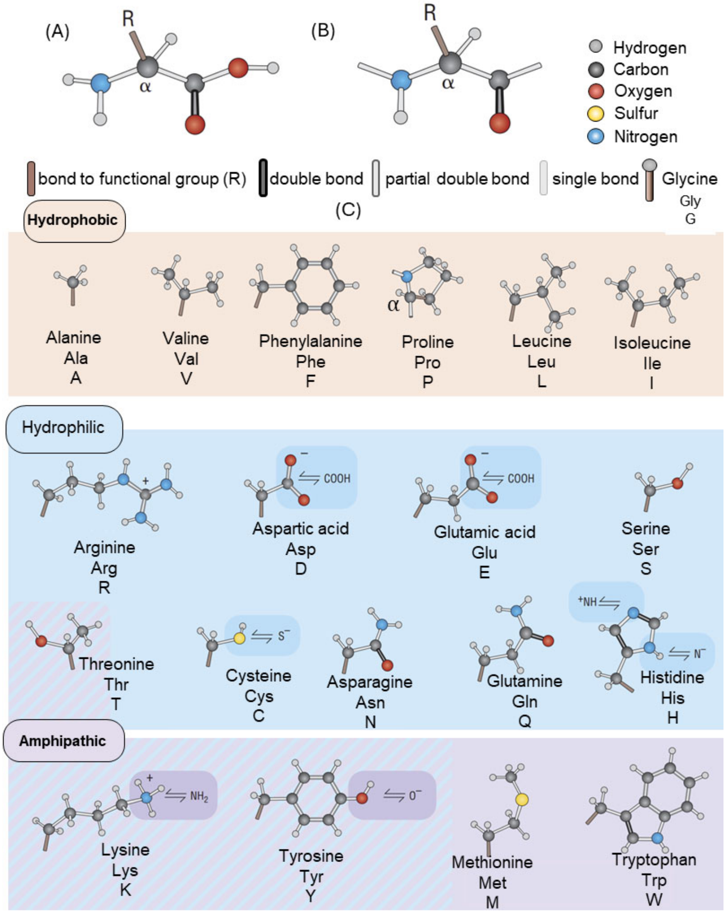


Figure 1.3: Amino acids structure and classification (A) The chemical structure of an amino acid is shown in its neutral form. At physiological pH, (pH ∼ 7) the amino and carboxylic acid group ionize to NH3$^+$ and COO$^-$. The backbone, common to all amino acids and consists of the amino group ($-$NH2), asymmetric C_α, and a carboxyl group ($-$COOH). Except for glycine, R group is simply a hydrogen atom, amino acids are chiral meaning they have a left-right asymmetry. The structure shown here represents the L-configuration, which is the predominant form found in proteins. (B) An amino acid residue within a polypeptide chain is depicted with its side chain (R group). (C) Different amino acids are distinguished by their unique R groups, which influence their chemical properties. The 20 side chains that occur in protein are depicted. Proline is unique because its side chain forms a cyclic structure by bonding back to the nitrogen of the backbone. The configuration about the C_α is L for most amino acids in proteins. The image is adapted from [10]

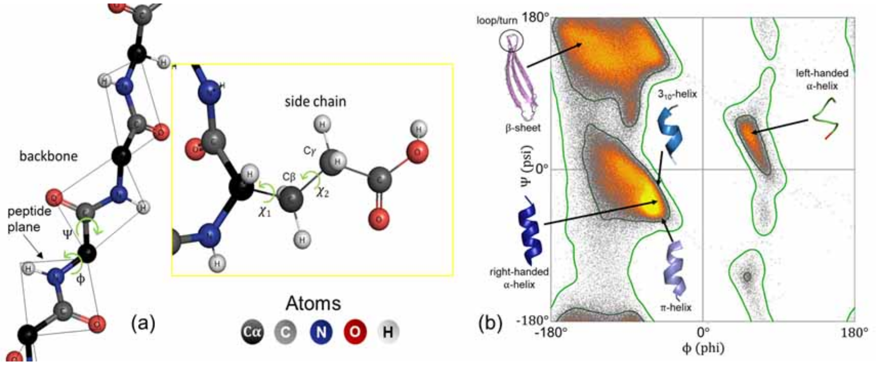


Figure 1.4: Structure of protein backbone (a) Protein backbone representing semi-rigid peptide planes connected at C_α atoms with rotations shown around $N-C_\alpha$ and $C_\alpha-C$ bonds depicted by ϕ ($C-N-C_\alpha-C$) and ψ ($N-C_\alpha-C-N$) dihedral angles, generated by PyMol. The dihedral angles of sidechain are also shown as rotations around $C_\alpha-C_\beta$ as χ_1 ($N-C_\alpha-C_\beta-C_\gamma$), and $C_\beta-C_\gamma$ as χ_2 ($C_\alpha-C_\beta-C_\gamma-C_\delta$). (b) Ramachandran plot demonstrating commonly observed ϕ and ψ values (grey) for proteins, and sterically allowed regions (orange yellow) for common secondary structure elements β-sheet, 3_{10}-helix, π-helix, right-handed and a left-handed α-helices of a protein (image sourced from Ref. [18]).

termini, and in loops, which are unstructured segments connecting the regular secondary structure elements. The loops or turns are the flexible linkers of helices and strands which have low degree of regularity and resemble a loop shape [21, 22]. A turn is a loop which allows the peptide chain to deviate 180° from its direction [22]. The formation of α-helices and β-sheets is primarily driven by H-bonding between the backbone atoms of the polypeptide chain, rather than the side chains. The α-helix is a spiraling coil while β-pleated sheets involve side-by-side β-strands held by H$-$bonds, making up the core of many proteins [23]. Other less common helices are 3_{10}-helix, π-helix, and left-handed α-helix (Fig. 1.4, b) [19, 24]. Each of the backbone residues belonging to helices and sheets can be represented as a point on the Ramachandran plot. In this plot, the angles are distributed across multiple regions depending on their sequence length, adjacent secondary structures, and molecular packing [25].

A tertiary structure of a protein is formed when R-groups from one or more secondary structure elements in a polypeptide chain interact, leading to a 3-D super-fold in space (Fig. 1.2) (c)[11]. The R$-$group interactions include H-bonding, ionic bonding, dipole-dipole interactions, London dispersion forces, and hydrophobic interactions. Polar R-groups engage in H$-$bonds and dipole-dipole interactions, while hydrophobic interactions involve the clustering of nonpolar, hydrophobic R groups internally. Additionally, disulfide bonds, a special type of covalent bond formed between cysteines, contribute significantly to tertiary structure, acting as strong molecular "safety pins" to hold parts of the polypeptide together [26]. Several proteins consist of a single polypeptide chain and exhibit only three levels of structure. Nevertheless, certain proteins are composed of multiple polypeptide chains, referred to as subunits. The assembly of these subunits results in the formation of the protein's quaternary structure (Fig. 1.2, c) [27].

1.3 Protein folding and dynamics

Protein folding is the intramolecular process constituting a biological self-assembly, wherein a protein's unfolded primary sequence is folded into its 3-D, typically biologically active, highly ordered structure under physiological conditions [28, 29]. *In vivo* protein folding often begins co-translationally, starting at the N-terminal as the protein is synthesized [30, 31]. This means that partially translated polypeptide chains begin exploring conformational space at the earliest opportunity, dictated by the time it takes for translation, which can extend to 10 s or more. Experimental evidence indicates that single domain globular proteins can spontaneously fold on a timescale ranging from μs to hours [32]. Within a test tube (*in vitro*), proteins are generally subjected to folding conditions from an unfolded ensemble. *In vivo*, the ribosome plays a crucial role in synthesizing proteins and also in preventing premature folding or misfolding during the folding process [30]. However, many proteins are unable to fold properly within the limited cellular timescales and require the assistance of specialized helper proteins known as chaperones to achieve their correct conformations [33]. Only some chaperones provide a hydrophobic folding environment, while other chaperone-like proteins assist folding through bond rearrangements rather than encapsulation. In addition to the well-ordered globular proteins, a significant portion of the proteome comprises of Intrinsically Disordered Proteins (IDPs), which lack a highly-ordered 3-D structure and instead exist as dynamic ensembles [34]. Schlessinger and colleagues describe disordered regions in proteins as segments that, when unbound to other molecules, fail to form a regular 3-D structure [35]. Instead, these regions remain disordered, sampling a wide range of conformations within their conformational space. Here, dynamic ensembles refer to states where atomic positions and backbone dihedral angles fluctuate significantly over time, without fixed equilibrium values [36]. Conversely, well-folded proteins exhibit dihedral angles that deviate minimally from their equilibrium positions, with only occasional, cooperative conformational changes [37]. IDPs share dynamic properties with the native and non-native states observed in globular protein folding [38]. IDPs are highly flexible, and despite this flexibility, they are not typically detected as misfolded by the cell's quality control systems. Kulkarni et al. described this phenomenon as the order/disorder paradox, suggesting that IDPs exploit their ability to transition from disorder to order when binding to biological targets (coupled folding and binding), allowing them to evade the cell's surveillance mechanisms [34]. As the nascent IDP emerges from the ribosome, it begins folding and is "recognized" by the chaperone as misfolded. However, by transitioning from disorder to a temporarily folded state upon binding via discrete amino acid sequence motifs, termed molecular recognitions regions, the IDP deceives the chaperone and avoids degradation.

Therefore, it is evident that protein folding is a complex process that varies widely across different proteins. Some proteins can spontaneously unfold and refold, while others cannot achieve folding without assistance. Folding pathways, therefore, are highly protein-specific. Once folded, proteins may undergo post-translational modifications (PTMs), where covalent groups such as glycosyl, methyl, phosphoryl, or acetyl are attached to specific amino acids, further influencing their structure and dynamics [39, 40]. Among the numerous PTMs—estimated at around 400—glycosylation stands out as particularly significant due to its critical role in regulating protein dynamics and is known for "outfitting" proteins for fold-function balance [40, 41]. Glycosylation is a process of covalent attachment of sugars or glycans to macromolecules usually proteins or lipids [42]. Glycosylation impacts more than half of eukaryotic proteins and is known to exert significant effects on the thermodynamic, kinetic, and structural features of proteins, extending beyond the influence of their primary sequence. Additionally, inhibiting or suppressing glycosylation often leads to protein aggregation or misfolding, resulting in nonfunctional states [39]. Thus far, it seems that the cellular machinery is primarily focused on one key task for proteins: preventing their misfolding. But why? A misfolded protein typically assumes a structure that deviates from the energetically optimal state (lowest free energy), essentially becoming kinetically trapped. For instance, failing to complete your Ph.D. within the expected timeframe can resemble a common scenario of being kinetically trapped, hindering progress towards your desired goal of becoming a scientist. So, chaperone proteins (Ph.D. promoters) act as hydrophobic boxes, decorated with hydrophobic groups. Misfolded proteins enter these chaperones, where hydrophobic interactions with the interior reduce the energy barriers between folded and misfolded states. This accelerates the process of exploring different conformations and, ultimately, facilitates proper folding [14].

Numerous experiments and theoretical studies have focused on spontaneous protein folding. Although these experiments do not fully mimic the cellular folding process, their simplicity has garnered significant theoretical interest. Consequently, the fundamental aspects of spontaneous protein folding are discussed in terms of thermodynamics and kinetics. The protein folding process is governed by three key questions [28].

1. What is the folding code that allows an amino acid sequence to determine a protein's native structure?
2. What enables proteins to fold rapidly?
3. How to predict the native structure of a protein from its amino acid sequence?

Initially, the thermodynamics of folding will be explained, focusing on how the native structure is determined by the sequence and microenvironment. Next, the kinetics of the folding process and the associated energy landscapes that allow proteins to fold rapidly will be discussed. Finally, the current tools for predicting the 3-D folded structure from the amino acid sequence will be briefly discussed. In the following sections, the role of glycosylation in protein folding will also be discussed.

1.3.1 What is the folding code?

Many classical experiments about protein structure and folding adhere to Anfinsen's dogma or the thermodynamic hypothesis, which postulates that, under specific physiological conditions, a protein will adopt a unique and stable conformation determined by total interatomic interactions, in which the Gibbs free energy (ΔG) of the whole system is lowest [43]. In the context of protein folding, ΔG represents the energy difference between the folded (native) state and the unfolded (denatured) state of a protein. Anfinsen's experiments commonly lead to the conclusion that the sequence of amino acids contains all the necessary information for achieving the native state. Many efforts have been dedicated to investigating the kinetically accessible folding routes (discussed in the next section) under this hypothesis. Specifically, Levinthal emphasized that proteins fold within a remarkably short timeframe compared to the exhaustive search required, implying that only a limited sets of conformations are sampled during folding, thus giving rise to a kinetic pathway [44]. Various well-known theories have been proposed to elucidate the process of protein folding. The framework theory and similar diffusion-collision theory suggest that the formation of secondary structures serves as the initial step and basis for overall folding [45]. In contrast, the nucleation condensation theory highlights the importance of specific global contacts as the starting point for both secondary structure formation and overall folding. Conversely, the hydrophobic hydration theory posits that the general repulsion between hydrophobic residues and the surrounding water environment drives the redistribution of polar and non-polar residues, ultimately leading to global folding [46]. In the more recent funnel theory, protein folding kinetics and thermodynamics are depicted as funnel-shaped, with a gradual decrease in both conformational space (entropy) and energy (enthalpy), and numerous kinetic traps along the folding pathway [47]. The funnel theory will be discussed in further sections. However, this theory does not specify the exact driving force behind protein folding. Interestingly, sixty-five years later, a recent study aimed to replicate Anfinsen's experiment by reducing disulfide bonds in ribonuclease (RNase), denaturing the protein with urea, and then refolding it under different conditions [43]. The study found that removing urea before promoting disulfide bond formation resulted in a fully functional, native protein, while promoting disulfide bond formation before removing urea led to an inactive mixture of protein forms. However, the study revealed that native disulfides play a pivotal role in shaping the secondary and tertiary structures, countering the common belief that these structures form independently and are then stabilized by native disulfides [48]. Moreover, contrary to Anfinsen's previous findings, spontaneous re-oxidation of reduced RNase did not fully restore enzymatic activity; instead, complete recovery required the use of a reshuffling solution such as GSH/GSSG (Glutathione) [48]. They also concluded that the continuous breaking and reformation of disulfide bonds, facilitated by a small amount of reducing agent or a GSH/GSSG mixture was essential for the correct folding of proteins [48]. This process preceded the final formation of secondary and tertiary structures. Under favourable conditions, approximately 50% of RNase spontaneously achieved its native conformation and proper disulfide bonds, while the remaining 50% adopted an energetically unfavourable state with incomplete structure and low activity. Reshuffling steps were required to convert the protein into its native form. Despite the current evidence, the general mechanism of *in vivo* protein folding remains unclear, particularly in distinguish between the differences and similarities in the folding routes and rates of different proteins. However, Anfinsen's thermodynamic hypothesis remains the most widely accepted basis of protein folding.

Returning to the thermodynamic hypothesis, the term-'native or equilibrium structure,' often associated with stability, may give the impression that its conformation is static; however, it undergoes constant motion and fluctuations, statistically balancing among numerous dynamic conformational states [14]. Here "stability" refers to a protein's potential to survive over time, influenced by kinetic changes associated with its conformation [49]. The thermodynamic hypothesis only argues that the native state is the lowest energy state within a specific conformational space including all kinetically accessible states [44]. Due to the immense size of the conformational space, testing the hypothesis becomes computationally challenging, as an exhaustive search of the entire conformational space is impractical. While these pathways lead to low-energy conformations relative to other accessible states, given the enormous size of the conformational space, there is no guarantee that these low-energy conformations are at global energy minimum. Large regions of conformational space may be kinetically inaccessible, potentially harboring more stable states [44]. Given that observed protein-folding times typically span the μs to s range [50], it implies that the Levinthal paradox provides a description of how proteins explore the conformational space. It also explains why they do not fold by randomly searching through all possible conformations.

This prompts the question: "How do proteins manage to reach their native folded states so efficiently and quickly, despite the vast complexity of their conformational landscapes?" .

1.3.2 How do proteins fold rapidly?

As a solution to the Levinthal's paradox of protein folding, several models of protein folding were proposed to elucidate whether secondary or tertiary structures fold first, the hierarchical nature of protein folding, the existence of folding nuclei, and related questions [28]. Some examples include the diffusion-collison model, nucleation-condensation model, zipping-and-assembly model, jigsaw puzzle model, hydrophobic collapse model, and the folding funnel hypothesis [51]. These models are not mutually exclusive in a sense that proteins could fold through multiple pathways or mechanisms operating simultaneously or sequentially leading to major advances in experimental and computational techniques. In this section the focus will be mainly on folding funnel hypothesis which describes a complete scenario and kinetics of protein folding.

In essence, protein folding can be conceptualized as navigating a funnel-shaped energy landscape through numerous parallel pathways, transitioning from non-native conformations to the native states located at the bottom of the funnel (Fig. 1.6) [47]. Like a ball rolling down a hill, proteins follow energy gradients, accumulating favorable interactions as they compact and search for their lowest-energy conformation. At any given point, the protein exists as an ensemble of conformations and might become temporarily trapped in local energy minima [52]. Here, the protein folding kinetics are described by the energy barriers between distinct conformational states in the folding pathway. In this context, the folding time of a protein is also related to the kinetics. On the contrary, free energy as discussed above relates to the thermodynamics of conformational states which includes relative populations or probabilities of conformational substates [51]. In kinetics, the folding funnel is a quantitative depiction of protein conformational space, encompassing the folded native state (typically regarded as a global free energy minimum), collections of different conformational substates near or distant from the native state, various unfolded or denatured states, and a broad spectrum of folding intermediates, such as molten globules, transition states, and trapped non-native states [52, 53]. Every unique protein sequence consists of its own funnel. Based on the biophysical properties, folding rate, native structure of the proteins, and the folding energy landscapes, the funnel pathways are divided into following distinct types: 1) ideal protein folding funnel with smooth surface, 2) rugged landscape with hills, 3) moat landscape, 4) champagne glass landscape, and 5) Levinthal's golf course [51]. The smooth funnel (Fig. 1.5, a) suggests fast folding with single-exponential kinetics, allowing denatured states to easily transition to a native conformation. With no traps or bumps, this model illustrates how a large number of denatured conformations can roll down through various routes to form fewer compact conformations. In contrast, the rugged landscape (Fig. 1.5, b) is marked by energy barriers and traps, resulting in slow folding characterized by multiple-exponential kinetics. The moat landscape (Fig. 1.5, c) introduces kinetic traps that require molecules to navigate through intermediates, while the champagne glass landscape (Fig. 1.5, d) depicts how conformational entropy creates free-energy barriers that delay folding. Lastly, the Levinthal's golf course illustrates the challenges of a random search, where finding the native state resembles a ball rolling aimlessly on a vast course before locating the hole. Together, these funnel models highlight the complexities and dynamics inherent in protein folding kinetics. Among all these models, the rugged landscape is the most realistic for most proteins, representing the multi-exponential slow folding process. In unfavorable conditions, the funnel becomes shallow, causing polypeptide chains to remain near the top surface of the funnel for extended periods [52]. This results in conformational heterogeneity among unfolded proteins and leads to very slow folding due to the shallowness of the funnel. In essence, because the polypeptide chains can not efficiently move toward a single, stable native state, they remain trapped in diverse, less organized states, leading to a distinct conformations rather than a uniform one. However, in favourable conditions, the slope of the funnel becomes steep leading to native state.

Next, with respect to the speed limit of folding, a small globular protein in an ideal smooth folding funnel has a folding rate of about $\sim 100\,N/s$, where N denotes the number of amino acids [54]. However, the majority of natural proteins fold at least two to three orders of magnitude slower than this predicted empirical relationship, largely because of the roughness present in the energy landscape [54]. Current insights into protein folding and behaviour primarily stem from experiments conducted on proteins in solution like nuclear magnetic resonance (NMR) spectroscopy or the detailed structures obtained through X-ray crystallography for individual molecules [27, 53]. Other experimental methods such as atomic force microscopy (AFM), mutational analyses, hydrogen exchange, cryo-electron microscopy (cryo-EM) and single-molecule approaches such as small-molecule fluorescence resonance energy transfer (smFRET) have also been pivotal in exploring the timescales of folding events across various proteins and probing the protein motions [55]. The computational methods include the use of molecular dynamics (MD) simulations (discussed in the next section). Despite six decades of advancements in this area, the knowledge gap remains regarding the general principle elucidating the folding routes and rates of various proteins, while accounting for random thermal motions [50, 56]. Despite these challenges, a few general observations highlight that proteins tend to fold in units corresponding to secondary structures (discussed in previous sections) before progressing towards more global structures. The conformational stability of a protein increases as its partial structures grow during the folding process. Despite the seemingly blind nature of the folding process, it can occur rapidly due to a divide-and-conquer strategy, moving from local to global structures [57]. Funneled landscapes imply that to attain a native structure, distinct protein molecules consisting of identical sequences could adopt diverse microscopic pathways, implying that certain pathways are visited more frequently than others [50]. At the bottom of the funnel, stable conformers exist as ensembles of states in equilibrium rather than a single rigid conformation. Perturbations such as temperature, pressure, pH, denaturants, mutations, PTMs, or ligand binding can disrupt this equilibrium, driving conformational transitions and redistributing state populations [58].

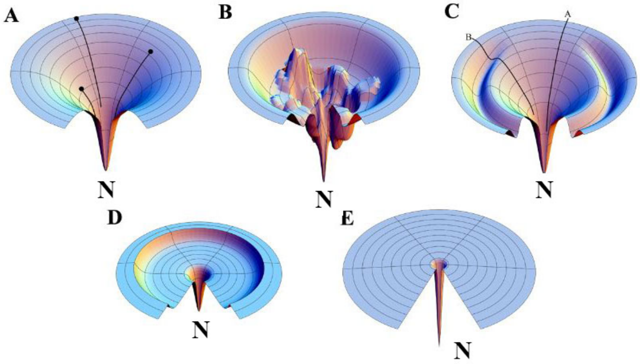


Figure 1.5: Models of protein folding funnels illustrating (A) ideal protein folding funnel with smooth surface, (B) rugged landscape with hills, (C) moat landscape, (D) champagne glass landscape, and (E) Levinthal's golf course. N represents the native state. The reference image is taken from Ref.[51].

1.3.3 Predicting native structure of a protein from a sequence

Finally, the challenge of predicting a protein's native structure from its amino acid sequence remains a key computational problem. Successfully addressing this challenge would (i) accelerate drug discovery, and (ii) enhance the annotation of protein functions based on genomic sequences [28]. As the number of experimentally determined structures grows in the PDB, protein structure prediction has increasingly become a challenge of inference and machine learning, in

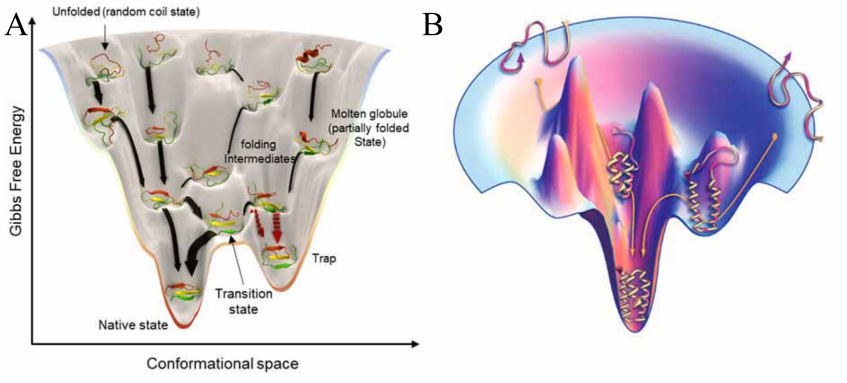


Figure 1.6: Protein folding pathway (A) The protein folding funnel landscape illustrates the folding trajectory from unfolded random coils to its lowest energy state, encompassing potential conformational states from Ref.[59]. As the protein initiates folding, the free energy and accessible conformational states decrease. The "red arrow" represents a local energy minimum that may temporarily trap the protein in a metastable state. The free energy reaches its minimum at the bottom of the funnel, and the protein assumes a singular conformational state referred to as the 'native state.' (B) A 3-D protein folding landscape is shown. (image sourced from Ref.[50]).

addition to protein biophysics. In this context, recent advancements by AlphaFold2, RoseTTAFold, and ESMfold have made significant strides in partially resolving the structure prediction problem, though it is important to recognize the limitations and varying degrees of accuracy associated with these predictions [60–62]. AlphaFold2 and RoseTTAFold utilize multiple sequence alignments (MSAs) to leverage homologous sequences and co-evolutionary relationships to infer protein structures [60, 61]. In contrast, ESMFold, employs large language models (LLMs) trained on single sequences to directly predict protein structures from the primary sequence, bypassing the use of MSAs [62]. These tools leverage known protein data from the Protein Data Bank (PDB) and have shown superior performance compared to traditional protein modeling techniques [60, 62, 63].

1.3.3.1 AlphaFold2: state-of-the-art

Among all the current state-of-the-art tools for sequence-to-structure prediction, AlphaFold2 has achieved unmatched accuracy in modeling the native state of monomeric proteins [60], as evidenced by its performance in the 14th edition of the "Critical Assessment of Techniques for Protein Structure Prediction" (CASP14) [64]. Therefore, AlphaFold2 will be discussed in detail in this section. AlphaFold2 predicts atomic coordinates and provides per-residue confidence estimates on a scale from 0 to 100, with higher scores indicating greater confidence and lower scores indicating less certainty [60]. This score is termed as pLDDT and is based on a per-residue lDDT-C_α metric. The lDDT measure, which stands for Local Distance Difference Test C_α, evaluates the accuracy of predicted atomic positions in protein structures by comparing them to reference structures [65]. The pLDDT assigns high scores for well-predicted regions, even if the overall prediction does not align perfectly with the true structure. This is especially important for evaluating multi-domain predictions, where individual domains may be accurately predicted, but their relative positions may not be. AlphaFold2 predicts a protein structure by integrating evolutionary, physical, and geometric constraints into a deep learning framework. It features a transformer-based Evoformer to process MSA and residue interactions, a structure module for end-to-end 3-D prediction, and equivariant attention for spatial accuracy. Using self-distillation and iterative recycling, AlphaFold2 refines predictions, achieving unprecedented accuracy, as demonstrated in CASP14. Despite its advancements, AlphaFold2 has certain limitations, particularly in modeling protein dynamics. In addition to this challenge, PTMs such as glycosylation, add another layer of complexity to protein structure, function, and dynamics. However, these modifications are often poorly represented in structure prediction methods like AlphaFold2.

1.3.3.2 Limitations of AlphaFold2 in capturing protein dynamics

The limitations in capturing protein dynamics arises from two main factors: 1) its training dataset, which primarily consists of single conformers or static structures derived from X-ray crystallography and cryo-EM, rather than multiple conformers from NMR ensembles, and 2) the nature of X-ray crystallography and cryo-EM itself, where proteins samples are prepared at cryogenic temperatures affecting protein packing, that does not capture the inherent flexibility of proteins in their biological environment. These limitations are discussed below in detail.

To validate the claim that AlphaFold2's training dataset, which mainly consists of static structures from techniques like X-ray crystallography and cryo-EM, may not fully capture protein dynamics, the PDB70 dataset was analyzed. Specifically, the average temperatures (in K) at which intensities were measured, as well as the experimental methods used, were examined to provide a comparative perspective. AlphaFold2 relies on PDB70 to retrieve structural templates for its predictions, and according to the supporting information from Ref. [60], the PDB70 dataset was collected from (https://doi.org/10.5281/zenodo.7953087). PDB70 contains a total of $91,032$ unique PDB IDs. Data regarding the experimental method and the temperature was retrieved via the REST API for only $78,901$ of these IDs. Of these $78,901$ IDs, $49,021$ lacked temperature information, and $7,477$ of those $49,021$ had no temperature data but did include an experimental method. The remaining $41,544$ IDs were marked as obsolete. The analysis of the PDB70 dataset revealed that a large portion of the protein structures were resolved at cryogenic temperatures, which aligns with the conditions typically used in X-ray crystallography and cryo-EM (Fig. 1.7). However, a substantial number of entries had missing temperature data, and a portion of the dataset was marked as obsolete, indicating the lack of both temperature and experimental method information.

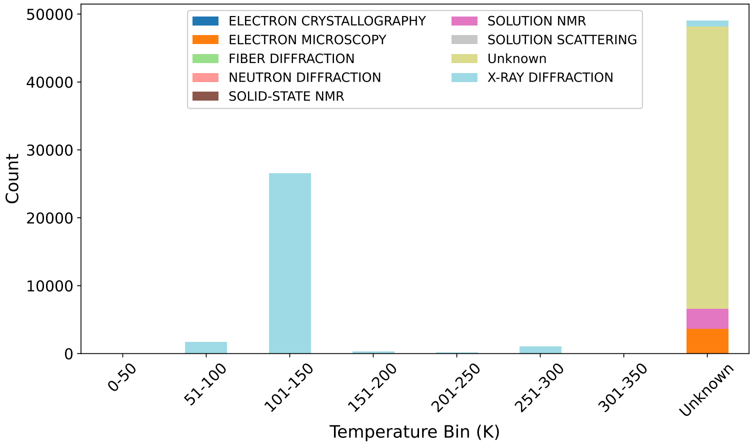


Figure 1.7: Temperatures and experimental methods in PDB70 dataset The plot visualizes the temperature distribution and experimental methods for protein structures in the PDB70 dataset. The x-axis represents the temperature in K, while the y-axis displays the count of proteins at each temperature. The categories on the x-axis represent different experimental methods used to resolve these structures, with a separate category for "Unknown" temperature data, indicating missing temperature information. The "Unknown" experimental method category likely indicates obsolete entries for which no experimental method data is available.

Next, AlphaFold2 operates under the assumption that the true structure exists as a single conformation [66]. While this assumption aligns with the majority of experimentally determined structures, it does not fully capture the inherent flexibility of proteins in their biological context, where they adopt multiple conformational states essential for their thermodynamic stability and function, such as G-protein-coupled receptors, transporters, and kinases, which undergo significant conformational changes [67, 68]. Some specific examples include dynamic targets like T1027 and T1088 where model's tendency to converge on a single best-fit structure rather than capturing a range of conformational states lead to inaccuracies [66]. As an instance, in T1027, the NMR structure places the N-terminal helix in a pocket within the protein core, while the C-terminal region is disordered. In contrast, the AlphaFold2 model shows the N-terminal region as disordered and instead features a newly formed C-terminal helix that packs into the core [66]. Additionally, for a curated set of *apo-holo* conformer pairs, AlphaFold2 correctly predicts the *holo* form of a protein in nearly 70% of cases, but fails to capture the observed conformational diversity, showing similar errors for both conformers due to the inability of AlphaFold2's pipeline's to sample the expected structural heterogeneity [67].

Secondly, prior to X-ray and cryo-EM data collection, the sample is prepared in solution at room temperature and then rapidly cooled to cryogenic temperatures (generally 100 K). Rapid cooling to cryogenic temperatures during techniques like cryo-EM can hinder the natural flexibility of proteins by freezing them in a less dynamic, static state. This is because the cryo-cooling process, during crystallography or cryo-EM, which takes up to a second, is too slow to capture the room-temperature equilibrium distribution of protein and solvent configurations, which may not be visible in B-factor analysis or standard model refinement [69]. B-factors, which measure atomic displacement, are affected by both thermal motion and static disorder. While thermal fluctuations dominate at higher temperatures, studies have shown that static disorder is often the primary contributor, especially at lower cryogenic temperatures, where protein dynamics are significantly reduced [70]. Thus, preventing the preservation of transient, higher-energy conformations and structural rearrangements that occur in biologically relevant, room-temperature conditions, potentially missing important conformational states that are crucial for the protein's function [71]. As a result, the resulting structures may subtly differ from the protein's true, functional state in its native environment. Therefore, while cryogenic temperature-based structures may miss key dynamic features, they offer high resolution, whereas room or physiological temperature studies could better represent protein function but are limited by technological constraints—limitations that are reflected in AlphaFold2's training.

These limitations highlight the need for a deeper understanding of how computational models like AlphaFold2 capture protein dynamics. This is addressed through a comprehensive analysis of AlphaFold2's predictions, comparing them to experimental NMR data, as discussed in **Gradations in Protein Dynamics** (outlined in the Research objectives). The chapter delves into how these predictions aligns with in-solution flexibility, and other computational methods of flexibility, providing insights into the challenges and potential improvements. The findings, presented in Ref.[72] (a published contribution of this thesis), provide a comparison between computational models and experimental observations.

1.3.3.3 The advent of AlphaFold3

To address the limitations of complex prediction (such as predicting protein with glycosylation or other ligands) in AlphaFold2, an updated version, AlphaFold3, has been introduced. AlphaFold3 extends the capabilities of its predecessor by incorporating protein complexes alongside nucleic acids, small molecules, ions, and PTMs to predict the joint structure of complexes directly from protein sequences [73]. AlphaFold3 uses confidence measures to predict errors in atom positions and pairwise representations in its predicted structures. Unlike AlphaFold2, which calculates errors during training, AlphaFold3 uses a "diffusion rollout" approach, predicting the full structure step-by-step. This method helps AlphaFold3 generate more accurate predictions. The model then calculates three types of error metrics: pLDDT, PAE (predicted aligned error), and PDE (error in the distance matrix of the predicted structure as compared to the true structure) to assess prediction accuracy. In this AlphaFold3 algorithm, template and genetic searches are performed, and the results, along with a conformer search, are provided as inputs to the structure template and MSA modules. The MSA module has been scaled down compared to AlphaFold2. The Evoformer module is replaced by a Pairformer module, which processes only single and pair representations, excluding MSA data. The model now uses a diffusion model instead of the previous structure model. Unlike conventional methods that predict a single structure, the diffusion model generates multiple possible structures, offering more accurate predictions with reduced uncertainty [73]. However, AlphaFold3, like its predecessors, struggles with proteins lacking evolutionary data, disordered regions, and physiological context. It does not handle chemical modifications, ligands, or ions well, and may produce unrealistic structures and produces atom clashes, chirality errors, and issues with modeling conformational changes.

Therefore, **Gradations in Protein Dynamics** also extends the discussion to AlphaFold3 by comparing the pLDDT scores between AlphaFold2 and AlphaFold3, further exploring their abilities to capture protein dynamics. These comparisons highlight the progress made in protein structure prediction underscores the need for both computational and experimental approaches to move beyond the simplification of assuming a static fold, aiming to better capture conformational dynamics, transitions, and protein interactions [74].

1.3.4 Dynamics and flexibility

Proteins are therefore "complex dynamic entities" influenced by temperature, forces, displacements, vibrations, and time. As highlighted by Zuckerman, proteins do not know biology, and at their core, they are simply molecules governed by the principles of physics and chemistry [14]. As described by physicist Richard Feynman, their primary role is spontaneous “wigglings and jigglings” or fluctuations to attain highly evolved functions [75]. The time-dependent fluctuations of a protein’s atomic coordinates can be described as protein dynamics. Protein dynamics can occur on a range of length scales and timescales dictated by the complexity of motion and structure [76, 77] (Fig. 1.8).

The protein dynamics span from the vibrations of individual chemical bonds at a sub-angstrom level and occurring within femtoseconds (fs) (1 fs = 10^{-15}s), to the complex rearrangement of domains and subunits spanning tens to hundreds of angstroms over μs to ms. Thus, the broad dynamical spectrum of these atomic vibrations ranges from rapid local motions to slow collective distortions across large molecular regions. For instance, motions of methyl groups in the side chains typically occur on the ps timescale [77]. The active site residues of an enzyme exhibit dynamics on the ps-ns timescale, and larger domain movements in proteins generally occur on the μs-ms timescale [77]. Motions occurring on the order of 10^{-14} to 10^{-13} s represent the fastest motions in proteins, arising from collisions between neighboring atoms and localized or collective vibrations. These vibrations are characterized by frequencies ranging from $3,000$ to 300 cm^{-1} [80]. Local structural transitions in which the initial and final states are separated by small energy barriers

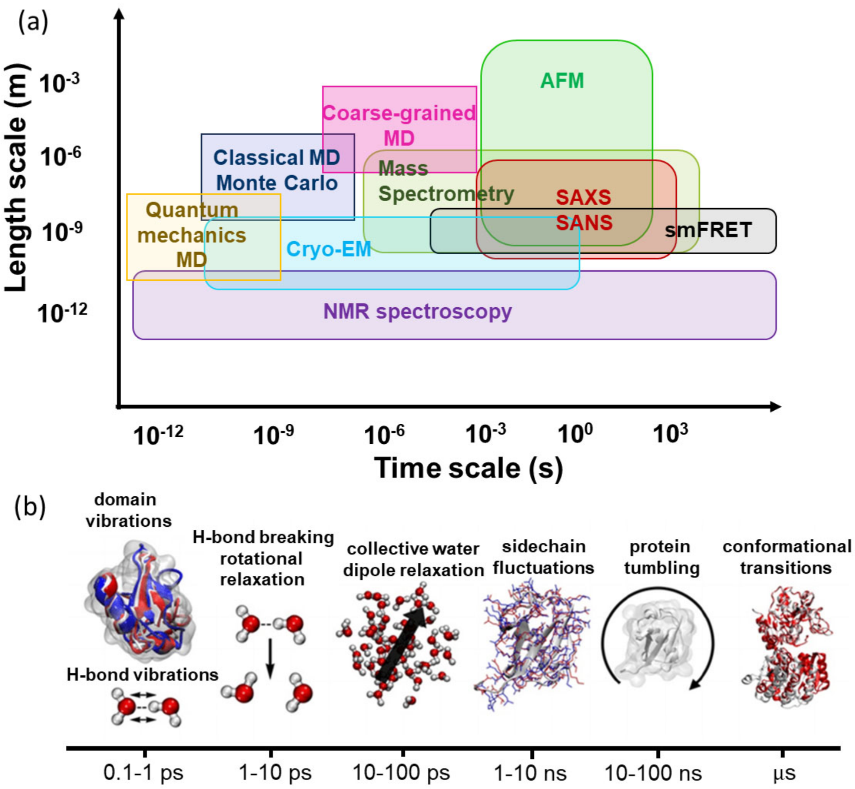


Figure 1.8: Protein dynamics (a) The description of protein dynamics is provided depicting a range of experimental and computational methods to probe the multiscale dynamics of proteins and their complexes (highlighted in coloured boxes). The methods shown in the figure are MD simulations, atomic force microscopy, small angle X-ray scattering, small angle neutron scattering, EM, NMR spectroscopy, and smFRET. The time scales are shown on the x-axis from ps to s highlighting local to global motions. The black arrows indicate the direction of motions occurring within the proteins. The y-axis shows different length scales. The image is taken from Ref.[78]. (b) Proteins motions are shown with intramolecular vibrations, hydrogen bond rearrangements and diffusion in water, dipole relaxation, sidechain fluctuations, rotational tumbling, and large conformational changes like hinge motions (µs and beyond) [79].

can also be very fast. The average effects of fast motions influence the dynamics of slower processes occurring in proteins [80]. For slow dynamics of proteins, there are two descriptions: "infrequent" processes consisting of rare conformational changes occurring over timescales ranging from less than 10^{-4} s to more than 1 s, and "intrinsically complicated" processes, characterized by extensive conformation sampling across configuration space and occurring over timescales ranging from about 10^{-11} to more than 10^{-7} s [80]. These dynamics involve events such as molecule movement, large-amplitude side chain movements, and global protein surface rearrangements, with varying contributions from steric constraints, solvent interactions, and dynamic rearrangements of numerous atoms.

1.3.4.1 The Functional Significance of Protein Dynamics

Understanding protein dynamics is essential, as motion underlies key biological functions like enzyme catalysis, ligand binding, and signal transduction [81]. Even subtle conformational changes can impact function, influencing reaction rates and binding affinities [82]. Moreover, altered dynamics are linked to diseases such as cancer [83], highlighting the importance of studying protein motion for therapeutic design and biomolecular engineering. Several examples illustrating the significance of dynamics are discussed below.

In a study, Smith et al.[83] demonstrated that KRAS exists in a dynamic equilibrium between an open, inactive conformation (state-1) and a closed, active conformation (state-2). Mutations can subtly disrupt this equilibrium, influencing KRAS activity, including its role in uncontrolled cell proliferation, making it a critical driver in many cancers. These mutations also affect the protein's response to inhibitors. As a GTPase, KRAS is a key molecular switch in cell signaling, cycling between an active GTP-bound state and an inactive GDP-bound state. The primary binding sites on KRAS—known as the switch-1/2 and switch-2 pockets—are considered "cryptic" because they remain inaccessible in the protein's most stable and predominant conformation, only becoming transiently exposed through dynamic fluctuations. In figure1.9, the functional mechanism of HRAS is shown. HRAS exhibits >90% sequence similarity and high structural similarity with KRAS, with their functions being highly conserved across most RAS isoforms. The switch-1 region, a flexible loop near its N-terminus, undergoes conformational changes that regulate effector binding. Mutations in KRAS alter the equilibrium between the inactive state (state-1) and the active state (state-2), highlighting the importance of understanding these conformational dynamics for developing targeted cancer therapies.

Myoglobin plays a vital role in oxygen storage and transport, with its ability to bind and release oxygen relying on its dynamic

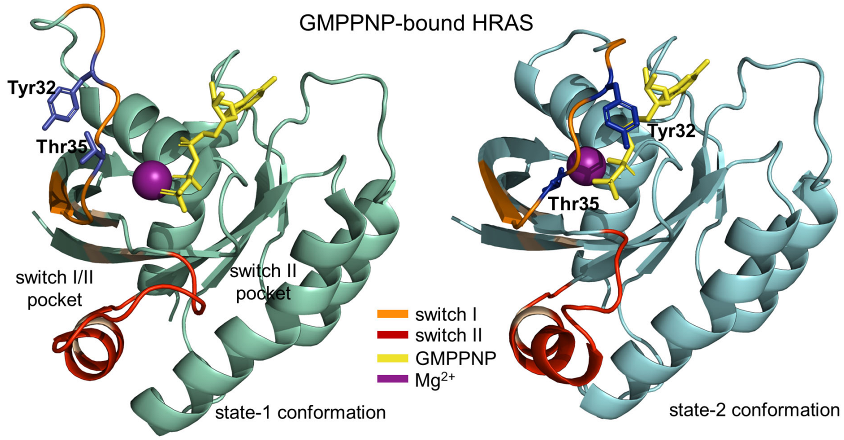


Figure 1.9: Conformational switch of HRAS GMPPNP-bound HRAS crystal structures with Mg^{2+} ion are shown in state-1 conformation (green, PDB id: 4efl), and state-2 (cyan, PDB id: 5p21) conformation. The side chains of Tyr32 and Thr35 are highlighted in blue, and switch I/II and switch II binding pockets are labeled. The image is generated in PyMol.

behavior [84]. The heme group in myoglobin binds oxygen and small molecules like carbon monoxide (CO), but its functionality requires it to undergo subtle conformational changes to allow these ligands to enter and exit the heme pocket. The distal histidine side chain typically obstructs the ligand's path, but dynamic rearrangements, including protein relaxation and shifts in conformational equilibrium, open temporary channels for ligand binding and unbinding [85]. Studies have shown that dynamic processes, such as photodissociation (where a laser breaks the bond between CO and iron in the heme), lead to heterogeneous, non-exponential rebinding behaviours [86]. These events occur due to different substates within the protein, each with distinct intrinsic rates. The protein's relaxation following dissociation is essential for CO escape from the pocket, and these conformational changes are crucial for oxygen binding specificity. This dynamic flexibility in myoglobin further emphasizes the critical role of protein dynamics in its function.

1.3.4.2 Investigating protein dynamics

Next, for investigating protein dynamics with these varying length and time scales, diverse experimental and theoretical methods and representations are suitable as shown in the Figure (Fig. 1.8). Protein dynamics, therefore, capture the movement or trajectory of a moving protein in time. The full description of the biophysical behaviour is contained in the ensemble of these trajectories. This concept, which illustrates how protein motion can be understood in terms of dynamic trajectories, is effectively explained by Zuckerman [14], and I will adapt this explanation for our discussion. For simplicity, consider a scenario where Anfinsen is working with many protein molecules, say on the order of a million, contained within a test tube. By diluting the solution adequately, Anfinsen ensures that these molecules do not interact with each other. Now, he could capture snapshots of these molecules at a particular moment in time using experimental techniques. These snapshots essentially encapsulate all potential conformations of the protein molecules, which are statistically distributed, thus forming an ensemble. This ensemble serves to provide an average representation of the behaviour or properties of the protein molecules. Alternatively, Anfinsen could zoom in on a single protein molecule and observe it over an extended period to measure its average properties. In this scenario, the individual protein molecule naturally undergoes movement, fluctuation, and dynamic behaviour, adopting different conformations for varying durations of time. These varying fractions of time during which the protein molecule occupies different conformations also constitute an ensemble. The fundamental premise here is that all the protein molecules within the solution are identical. Therefore, over time, they are expected to experience the same populations of conformations in different, random orders. Consequently, the snapshot obtained from Anfinsen's test tube should capture the molecules with precisely the same distribution as each molecule experiences individually over time. In essence, experimental studies combine both time-averages and ensemble averages. While most experiments involve studying large numbers of molecules simultaneously (such as those in a solution), it is important to note that any measurement, even those seemingly taken at a single time point, effectively averages over a specific window of time determined by the speed of the instrument [14].

Thus, a widely used theoretical technique to generate the conformation ensembles in molecular biophysics is all-atom MD simulations. Despite its significant computational cost, MD simulations provide a detailed picture of atomic movements (Fig. 1.10). These simulations utilize detailed force fields to calculate interactions among all atoms in a system, including the protein or complex under study, as well as surrounding water and ions. Force fields are parametrized functions that are used to predict the potential energies of protein structure using empirical potentials that describe the electrostatic and vdW interactions, vibrations of angles and bonds, and internal rotations of torsions [87]. These force fields model interactions by numerically solving Newton's equations of motion over numerous time steps, typically ranging from 10^7 to 10^{10} steps of 1-2 fs each [80]. The high number of time steps can be invaluable for investigating local motions within the context of drug design, as well as the effects of PTMs and mutations. However, simulating the collective motions of large macromolecular assemblies over longer timescales often requires dedicated supercomputers or sophisticated enhanced sampling algorithms [76]. In such cases, alternative approaches may be preferred or necessary to accurately capture these global motions. These global or collective motions are characterized by large-amplitude and low-frequency involving a significant region of a protein. Due to the large scale of these motions, a coarse-grained description is sufficient to investigate their dynamics compared to an all-atomistic detailed description [76]. For instance, these dynamics can be modeled at the level of amino acid residues, representing them with a single node based on the carbon atom or an average of all atoms within the residue.

There are two methods to analytically evaluate the fluctuations associated with global motions or conformational variability rapidly, which are Principal component analysis (PCA) and Normal mode analysis (NMA) using linear algebra [81, 88](Fig. 1.10). PCA is a method that sifts through a collection of conformers obtained from experiment or simulations, to pinpoint and extract those displaying the most significant fluctuations. PCA is a dimensionality reduction technique that is carried out as a postprocessing step of the simulation trajectory, or ensembles, while NMA is used to simulate protein movements and generate data on protein flexibility. Therefore, NMA is a method that is applied to a single conformer which is assumed to be representative of the equilibrium structure, to generate conformers around equilibrium. However, these methods have their limitations. Global motions may result in unrealistic deformations in bond lengths and angles due to their inherent lack of atomic-level detail. To correct these limitations, these methods are often combined with MD simulations [76]. However, MD simulations can only explore limited timescales due to limitations associated with force fields and high-computational cost leading to insufficient sampling of conformational dynamics [89]. Experimental methods such as NMR spectroscopy are often combined with theoretical methods to provide information about the dynamics of proteins within ps to ms timescales [90]. NMR can also be used to probe the dynamics of partially folded or unfolded states of proteins due to its high resolution [91].

These techniques and their parameters will be discussed in detail in the following chapters. Thus far, it has been established that overall, all proteins are intrinsically dynamic which is essential for their function. *However, at the level of a protein structure, the next big question is, are these dynamics uniformly distributed throughout the structure or are these only present in some parts of the proteins?*

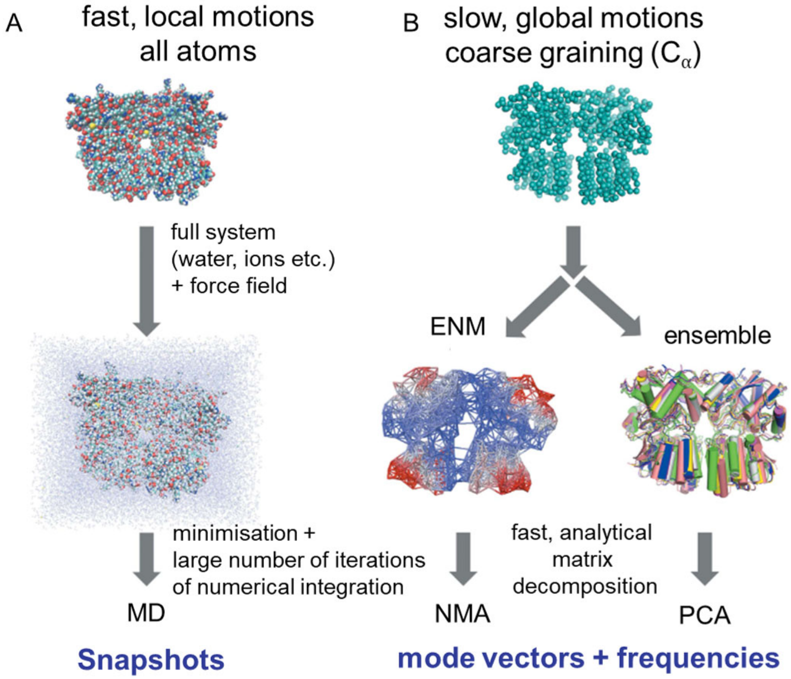


Figure 1.10: Different methods to study protein motion (A) An MD system containing all-atom model of glutamate receptor N-terminal domain dimer (GluA3) protein, surrounding water and ions is shown. A specialized force-field is applied to the system for extensive energy minimization, and to run multiple iterations of MD simulations to capture fast and local motions. (B) A coarse-grained representation of GluA3 protein is shown, where each node corresponds to a C_α atom of each amino acid residue of the protein. To capture slow and global motions, NMA and PCA methods are applied using fast, analytical matrix decomposition to extract mode vectors. The image is taken from Ref.[76].

The answer to the previous question lies in the flexibility of a protein. Previous research has demonstrated that protein dynamics vary across different regions of the protein, particularly highlighting that sites involved in protein-protein interactions are highly dynamic [92, 93]. Thus, it is crucial to emphasize that while proteins exhibit flexibility due to their dynamics, not all dynamic proteins are inherently flexible [94]. Consequently, a protein with highly dynamic behaviour may not always display high flexibility, although this correlation is often observed [94]. Here, flexibility of a protein refers to its ability to adapt its structure or conformation due to the changes in its environment such as binding to ligands. Often, protein flexibility is interchangeably used with protein stability, though this relationship is debated among structural biologists and protein chemists [95]. For clarity, I will use the term 'conformational flexibility' hereafter when referring to flexibility. Based on different timescales, distinct measures of protein conformational flexibility can be observed, measured, computed or predicted. At the protein residue level, the conformational flexibility due to the local motions can be assessed using root-mean-square fluctuation (RMSF) [96]. For a single residue, RMSF can be described as a time-averaged measure of fluctuation which is computed from MD trajectory after the removal of translational and rotational movements [97]. From the experimental methods such as NMR, conformational flexibility can be computed from Random Coil Index (RCI) which is a metric computed from empirically derived relationship between NMR derived chemical shifts and protein mobility [98]. In addition, crystallographic temperature factors or B-factors can be related to positional fluctuations of atoms due to thermal vibration and static disorder[96]. Also, a protein's intrinsic flexibility based on a single protein structure can be computed using NMA at finite temperature [81]. Sequence based machine learning approaches based on protein's evolutionary information [35] and backbone dynamics data from experiments can also be used to predict per-residue flexibility such as DynaMine, and ShiftCrypt [99–101]. The computational metrics of flexibility such as RMSF computed from DynaMine, MD simulations, PCA, and NMA are addressed in detail in the Methods section.

1.4 Glycosylation

1.4.1 Overview of glycosylation

Out of the various types of covalent PTMs that a protein undergoes, glycosylation is most prevalent, most diverse in terms of both composition and the amino acids that are modified [102]. The process of glycosylation includes the specific attachment of glycans (also known as sugars, or saccharides) to proteins and lipids [42]. The chemical composition of glycans comprises monosaccharides covalently linked by glycosidic linkages (bonds), forming both linear and branched chain-like structures [103]. The diversity of glycans arises from both biological as well as chemical factors. Chemically, the monosaccharides can combine in various ways differing in their sequence such as N-acetylglucosamine (GlcNAc) or Mannose (Man), chain length, linkages, where linear glycan chains branch into multiple branches (branching points), and anomeric effects (α or β) (Fig. 1.11, A) [103, 104]. Monosaccharides undergo cyclization to form ring structures (pentose, hexose) and acquire an asymmetric center derived from the C atom containing carbonyl group, known as anomeric Carbon (C1). The cyclization reaction thus results in the formation of two stereoisomers since hydroxyl group of C1 can orient in two possible orientations α or β [105]. The glycosidic linkage is formed via the hydroxyl group of the anomeric C between two monosaccharides resulting in either α or β linkages, which are determined by the position of the glycosidic oxygen relative to the anomeric carbon and the ring structure (Fig. 1.11, A) [105]. Under most conditions, after the formation of a glycosidic bond, the configuration of glycosidic linkage remains stable [106]. Like polypeptides consisting of N- and C-termini, a glycan chain has a reducing end and a non-reducing end, which together describe the glycan's polarity [107]. The reducing end of the glycan chain has a free anomeric center that is not engaged in a glycosidic bond and thus retains the chemical reactivity of the aldehyde. However, it continues to be referred to as the reducing end even when it is engaged in a linkage to another hydroxylic compound, such as the hydroxyl group of serine or threonine in glycoproteins. Structures are commonly written from the nonreducing end on the left toward the reducing end on the right (Fig. 1.11, B).

Additionally, glycans can undergo further modification through covalent attachment of sulfate, phosphate, acetyl, or methyl groups, leading to a vast array of potential glycans, including oligosaccharides and polysaccharides, from a limited number of monosaccharides [112]. Monosaccharides are the simplest form of carbohydrates consisting of single sugar units, polysaccharides are long chains of monosaccharide units bonded together, and oligosaccharides are carbohydrates composed of a few (typically 3-10) monosaccharide units. Biologically, the diversity arises from glycans being the secondary gene products meaning not directly encoded by the DNA, unlike proteins, which are primary gene products [113]. This results in glycosylation being species- and cell- or tissue-specific, influenced by the protein's structure and the site of its covalent attachment [114]. Consequently, the glycosylation pattern of any protein is determined by the cell or tissue in which it is produced, with the polypeptide itself encoding information that directs its glycosylation pattern [103]. There exist two

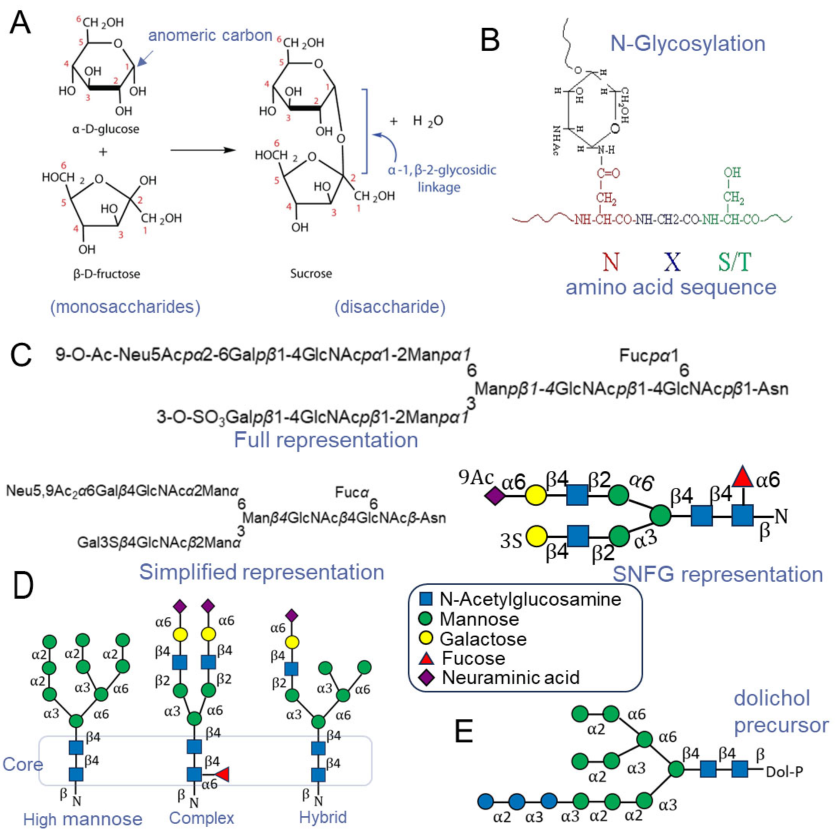


Figure 1.11: Chemical structure and nomenclature of glycans N-glycosylation and glycans. (A) Formation of α-1, β-2-glycosidic linkage of sucrose between α-D-glucose and β-D-fructose to (Ref. [108]). (B) Formation of N-glycosidic linkage between a glycan and Asn residue on the N-X-T amino acid sequence or N-glycosylation site of a peptide (Ref. [109]). (C) Representation of a glycan chain sequence in a full IUPAC representation, a simplified representation, and a Symbol Nomenclature for Glycans (SNFG) representation (Ref. [104]). Core fucosylation of the glycan chain is shown by the attachment of fucose at N-linked N-acetylglucosamine (GlcNAc). (D) Types of N-glycans. The three different types (high mannose, complex and hybrid) share a common core structure including the first two GlcNAc residues and the first three mannose residues. Image is adapted from Ref. [110]. (E) A dolichol-oligosaccharide precursor required for the initiation of glycan biosynthesis is shown (Ref. [111]).

primary types of glycosylation: N-glycosylation and O-glycosylation. Alongside these, there are three less common forms of glycosylation, known as glycation (or nonenzymatic glycosylation), C-glycosylation (C-mannosylation), and glypiation (or glycosylphosphatidylinositol anchoring)[115]. Specifically, in this chapter, the focus will be on N-glycosylation. N-glycosylation is the covalent addition of a glycan moiety to the amide side chain of an Asn residue within any of the specific consensus sequences: N−X−S and N−X−T (and in some rare cases, N−X−C), where X could be any amino acid except Pro (Fig. 1.11, B) [116]. These tripeptide sequences are known as sequon, and the site of attachment is known as glycosylation site of a peptide or a protein. Each glycosylation site of a protein can potentially attach to different glycans, also termed as site heterogeneity. This potentially leads to functional diversity as it causes microheterogeneity within the entire molecule, resulting in distinct subsets, or glycoforms, of a glycoprotein that possess varied physical and biochemical properties [117]. Thus, a set of glycoforms emerges due to protein glycosylation, consisting of a similar backbone, but varying in the arrangement or structure of their individual glycan units or both. Despite, the high variability and complexity of glycans, the composition of individual glycome resulting into different glycoforms at the cellular level indicates stable regulatory cellular mechanisms [113]. As a result, glycoform patterns (glycosylation patterns) can differ based on the organism, tissue, cell type, and their physiological state. Unlike proteins, which can result in six possible trimers from a combination of three different amino acids, three different monosaccharides (for example, three different hexoses) could result in approximately 1000 to 27000 unique tri-saccharides. Despite the astronomical size of possible combinations of glycans, there are limited number of naturally occurring monosaccharides and their possible combinations.

1.4.2 Biosynthesis and composition of N-glycans

N-glycosylation occurs in both prokaryotes and eukaryotes, with several differences in their biosynthetic pathway [118]. In eukaryotes, the primary biosynthetic pathways responsible for producing mature glycoproteins involve glycosyltransferases (GTs), glycosidases, and enzymes that modify carbohydrates [119]. The pathway consists of three major steps: (1) the formation of the lipid-linked oligosaccharide (LLO) donor, (2) the co-translational transfer of the glycan onto Asn-X-Ser/Thr (N-X-S/T) nascent polypeptide chain [120], and (3) processing of the $Glc_3Man_9GlcNAc_2$ oligosaccharide chain in the endoplasmic reticulum (ER) and Golgi [121, 122]. This pathway is illustrated in detail in Fig. 1.12.

Initially, membrane-embedded glycosyltransferases (GTs) on the cytoplasmic side of the ER sequentially attach monosaccharides from soluble nucleotide carriers to dolichol-phosphate (Dol-P) molecules

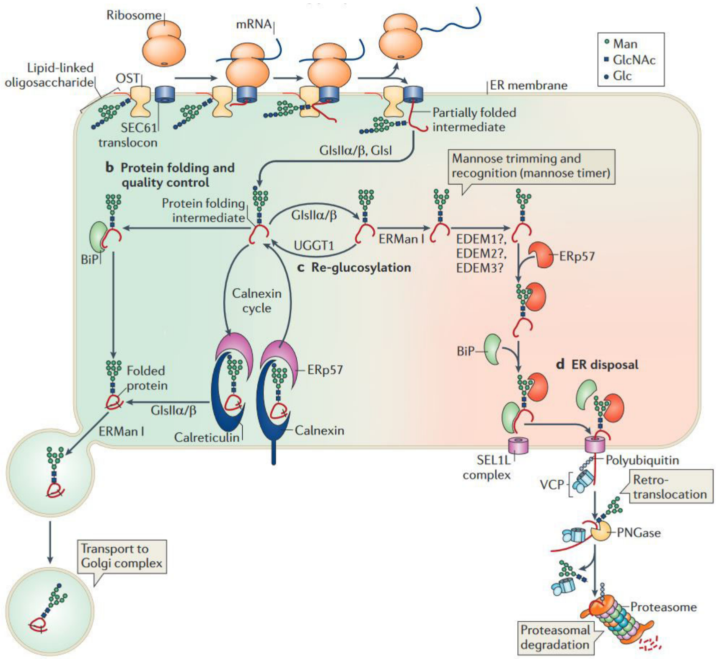


Figure 1.12: Process of protein N-glycosylation and fine-tuning during glycoprotein biosynthesis (a) During biosynthesis, nascent polypeptides are translated and translocated through the SEC61 pore. Simultaneously, glycan is transferred from a LLO to N-X-S/T sequon by OST. One cleft in the STT3 subunit of OST scans for acceptor sequons, while another cleft binds the glycan donor. (b) Immediately after transfer, glycan trimming begins with the removal of glucose residues by α-glucosidase I (GIsI) and the α-glucosidase II α-β heterodimer (GIsIIα/β). Folding intermediates containing $Glc_1Man_9GlcNAc_2$ structures interact with the lectins calnexin or calreticulin, along with ERp57, facilitating proper folding. Upon dissociation from lectins, further glucose cleavage occurs. Additional chaperone assistance is provided by ATP-driven BiP (GRP78). Correctly folded glycoproteins are then packaged for transport to the Golgi. (c) Folding sensor UDP-Glc:glycoprotein glucosyltransferase (UGGT1) recognizes incompletely folded glycoproteins, which are reglucosylated by adding a glucose residue back to the glycan structure. These proteins are reintegrated into the calnexin cycle for further folding attempts. (d) Terminally misfolded glycoproteins undergo ER disposal through mannose trimming by ER α-mannosidase I (ERManI) or GolgiManIA/B/C enzymes. This process is followed by the action of ER degradation-enhancing α-mannosidase-like proteins (EDEM1-3), which bind trimmed glycans and facilitate translocation into the cytosol. The peptide is then deglycosylated by cytosolic PNGase and degraded by the proteasome. Green spheres represent mannose, blue squares N-acetylglucosamine (GlcNAc) and blue spheres glucose residues. The image and caption is adapted from Ref. [111].

(Fig. 1.11, E) [111]. This process results in the formation of Dol-P-oligosaccharides, which are then transported across the membrane by flippases [119]. On the luminal side of the ER, additional membrane-embedded GTs continue to modify the oligosaccharides. Simultaneously, monosaccharides for glycosylation in the ER lumen are also linked to Dol-P carriers. Subsequently, the fully assembled oligosaccharide is transferred *en bloc* to an acceptor protein within the ER lumen by N-OST [123]. Following this transfer, the oligosaccharide undergoes trimming, starting with the removal of two terminal glucose residues, resulting in $\mathrm{Glc_1Man_9}$-$\mathrm{GlcNAc_2}$-Asn-linked protein [120]. In this state, the newly synthesized glycoprotein enters the calnexin/calreticulin cycle. Calnexin (membrane-bound) and calreticulin (soluble) are ER-resident lectin proteins that specifically interact with mono-glucosylated glycoproteins to aid in their proper folding [124]. Once the glycoprotein attains its native conformation, it exits the calnexin/calreticulin cycle and continues through the secretory pathway [125]. Alternatively, if further folding assistance is needed, the glycoprotein may undergo re-glucosylation and return to the calnexin/calreticulin cycle [125]. After the remaining glucose residue is removed, one mannose is trimmed, resulting in $\mathrm{Man_9GlcNAc_2}$-Asn-linked protein [120]. This processed high-mannose glycan then acts as a substrate to assist in the formation of diverse structures of glycans substituted with other glycan moieties such as Galactose (Gal), Fucose (Fuc), and Sialic acids (Neu5Ac or Neu5Gc) (Fig. 1.13) (A) [116]. Based on mannose substitutions and processing of these initial N-glycans structures, there are three major types of N-glycans: high-mannose, hybrid, and complex glycans (Fig. 1.11) (D) [126]. The high mannose type of N-glycans demonstrate minimal processing of mannoses, while hybrid N-glycans show GlcNAc attached to the $\alpha 3$ arm of mannose [111]. The hybrid N-glycans can be extended by the addition of other glycan moieties [126]. The processed $\alpha 6$ arm of hybrid N-glycan results in a complex N-glycan structure [111]. Both complex and hybrid N-glycans can contain two or more branches, termed as antennae. The N-glycan structures of hybrid can be mono-antennary or bi-antennary, while complex glycans can be multi-antennary [111]. Subsequently, the complex and hybrid N-glycans in Golgi can be post-processed by a fucosyltranferase in the Golgi, resulting in a core-fucosylation modification [126]. This modification leads to the addition of a Fuc moiety in an $\alpha 1$-6 linkage to the GlcNAc linked to Asn (Fig. 1.11, B). Core-fucosylation has demonstrated an increased expression in tumorigenic tissues in comparison with the healthy tissues, indicating its potential as a biomarker of cancer [127–129]. Apart from core-fucosylation, the fundamental changes in the glycosylation patterns of cell surface and secreted glycoproteins, also known as aberrant glycosylation, have been shown in various disorders and diseases [130]. These include autoimmune diseases, rare congenital disorders of glycosylation, and especially cancer including carcinogenesis, tumor progression, and metastasis [130–132]. However, it remains unclear for most diseases whether aberrant glycosylation is the cause, or a consequence of the disease. Currently, most of the tumor biomarkers used in cancer diagnosis are glycoproteins, specifically serum glycoproteins such as AGP, α-fetoprotein, and Thyroglobulin [133]. Given their significance in cancer and other diseases, it is essential to identify the structure, conformation, and dynamics of these glycoproteins and their associated glycans. Understanding the key structural details at the atomic and molecular levels, along with their key interactions, is crucial for effectively developing targeted therapeutic molecules.

1.4.3 3-D structure of glycans

In a 3-D structure of a disaccharide within a glycan, the glycosidic linkage exhibits the highest conformational flexibility due to axial and equatorial nature of the glycosidic bonds with respect to the ring structure of the saccharide [134]. In N-glycans, monosaccharides can be linked via 1→1, 1→2, 1→3, 1→4, 2→3, 1→6, or 2→6 glycosidic bonds, where the numbers denote the carbon atoms of the two monosaccharides involved [134]. The relative orientations of these two monosaccharides are described by two torsion angles, ϕ and ψ, for 1→1, 1→2, 1→3, 1→4, and 2→3 linkages, and by ϕ, ψ, and ω for 1→6 and 2→6 linkages [134]. An example of a disaccharide consisting of α-Neu5A and β-Gal linked with 2→6 glycosidic bond is shown in Fig. 1.13, B. The three torsion angles ϕ (O6-C2-O6-C6), ψ (C2-O6-C6-C5), and ω (O6-C6-C5-O5) of α-Neu5A-(2→6)-β-Gal are shown. The ϕ angle is characterized by the exo-anomeric effect [135]. This effect, an extension of the anomeric effect, causes ring substituents to adopt a gauche conformation, despite steric hindrance typically favoring an anti-periplanar conformation. The ψ angle is characterized by steric interactions and H-bonding between residues and with the solvent [136]. The ω angle can adopt three staggered rotamers based on steric interactions, referred to as gauche-trans (gt), trans-gauche (tg) and gauche-gauche (gg) [136] (Fig. 1.14).

Thus, the high flexibility of glycosidic linkages, combined with the rigidity of the monosaccharide units within the glycan molecule, leads

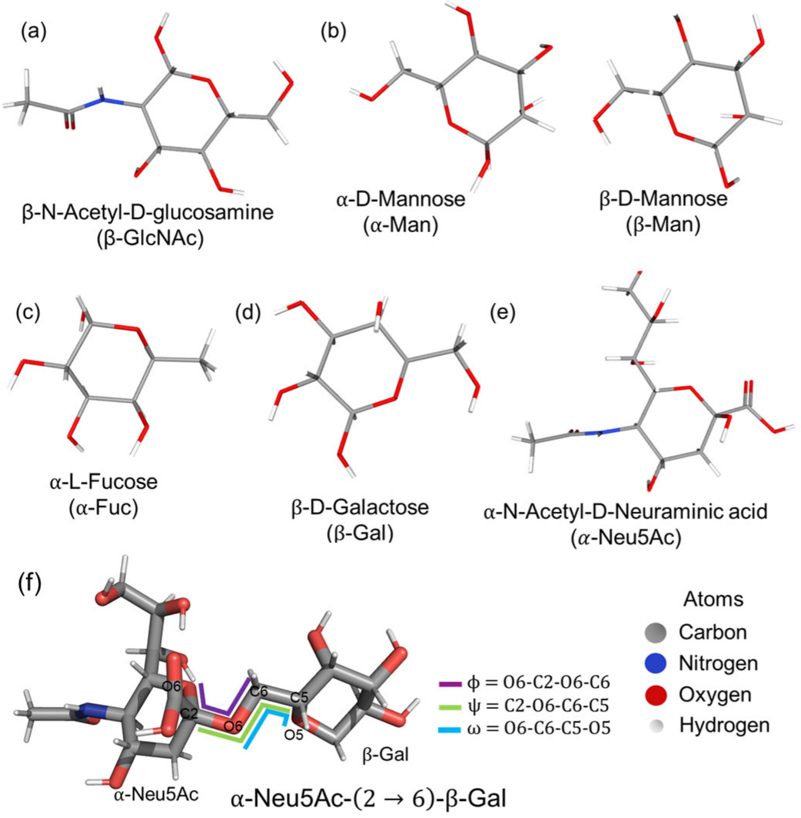


Figure 1.13: 3-D structure of glycans saccharides (a-e) examples of monosaccharides commonly found in N-glycans with their standard names, and SNFG names in brackets (source PubChem) with a wire-frame representation, (f) simplified 3-D representation of a disaccharide α-Neu5Ac-(2→6)-β-Gal with three torsion angles ϕ (purple), ψ (green), and ω (skyblue). The atoms are coloured as follows: C (grey), N (blue), O (red), and H (H). The image is generated in PyMol.

to multiple distinct conformations in solution [87]. The conformational states of N-glycans can be characterized using carb-RAMA plots (1.15) containing the statistically significant distribution of torsion angles of glycosidic linkages similar to the Ramachandran plots of a protein [138]. Glycans demonstrate multiple minima in their conformational phase space which are separated by different energy barriers of greater than few k_BT [139]. Within the typical timescale of conventional MD simulations, which spans only a few tens of ns, the transitions of glycosidic torsion angles between different conformational states occur rarely [139]. Therefore, to comprehensively

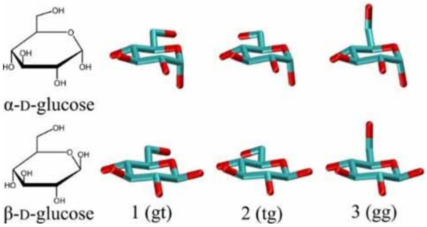


Figure 1.14: Conformers of glucose (1) gauche-trans (gt), (2) trans-gauche (tg), and (3) gauche-gauche (gg) conformation for α-D-glucose and β-D-glucose. The 3D conformers are visualized in a stick model with C atoms (turquoise) and O atoms (red). The image is taken from Ref. [137].

capture the various conformations of N-glycans, extensive sampling of their conformational space is required. Experimental approaches such as mass spectrometry, chromatography, lectin or antibody binding assay, cryo-EM, NMR spectroscopy, and X-ray crystallography are used to resolve the complex 3-D structure of glycans, yet their high flexibility poses significant challenges [111, 140–142]. For instance, the heterogeneity of glycans on protein surfaces disrupts crystal packing, and their high flexibility prevents the electron density maps of these glycans from being resolved [143]. To tackle these challenges, experimental techniques are often integrated with molecular modeling and simulations, such as Monte Carlo and MD simulations, to investigate glycan conformations in their native environment [144]. Significant research has been devoted to developing force fields for the all-atom MD simulations of glycans including CHARMM36m, GLYCAM06, OPLS-AA-SEI, and GROMOS 53A6GLYC [145–148]. MD simulations of glycans using these force fields, along with proper treatment of solvation effects, can help identify the most populated rotameric states in the conformational space of glycans in solution, provided that the sampling is sufficient (Fig. 1.15) [149]. While accurately predicting state populations for highly flexible glycosidic linkages with three torsion angles (ϕ, ψ, and ω) remains challenging, predictions for the more common linkages with two torsion angles (ϕ and ψ) have reasonable accuracy, as these linkages often significantly populate only one rotameric state [87, 150]. In summary, the flexibility of glycans comes from two main factors: transitions that last very long (on the ns timescale) between stable conformations and rapid motions (on the ps timescale) around these stable conformations [87].

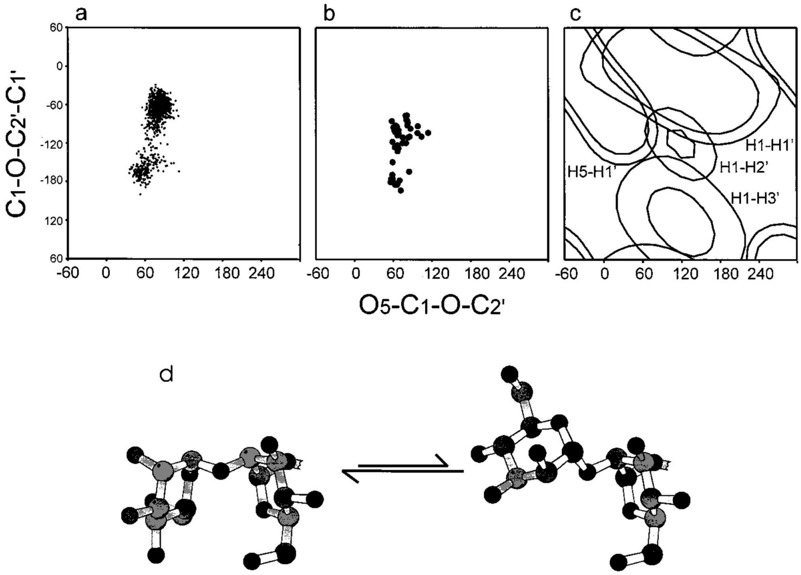


Figure 1.15: Conformations of αMan1→αMan2 glycosidic linkage in Man$_9$GlcNAc$_2$ The subplot (a) shows the ϕ and ψ distributions as carb-RAMA plots of an unrestrained MD simulation of Man$_9$GlcNAc$_2$ where each datapoint corresponds to a snapshot of the glycan structure at 1 ps interval each for a MD trajectory of 1000 ps. The subplot (b) shows the discrete conformers of the Man$_9$GlcNAc$_2$ obtained from X-ray crystallography. The subplot (c) shows the distance contraints obtained from the solution NMR data depicting with four interproton distances which sufficiently describe a single conformation of a glycosidic linkage. The figure (d) shows two different conformations of αMan1→αMan2 linkage in a dynamic equilibrium inclining slighly towards the left conformation. These conformers are dynamic however show limited oscillations for ϕ and ψ. The figure is taken from Ref.[136].

1.4.4 Effect of glycans on protein conformation: dynamic and sticky molecular glue

Various studies have shown that glycosylation of protein can alter its kinetic, thermodynamic, as well as structural properties [41, 136, 151, 152]. An H/D exchange NMR study by Joao and coworkers showed that thermostability of a glycosylated RNAse increased due to decrease in its overall structural dynamics due to glycans as distant as 30 Å from the glycosylation site [153, 154]. Their study indicated that these local effects might influence the entire protein structure. The general findings from the studies based on specifically N-glycosylation indicate that N-glycosylation does not induce the secondary structure in an unstructured protein permanently, instead, it affects the conformational preferences of the protein backbone, making more compact conformations more probable [136]. These biophysical effects appear to affect only the first few residues of the glycan and are probably driven by steric and hydrophobic/hydrophilic interactions between the core glycan residues and the adjacent amino acid side chains [155]. Another recent study showed that N-glycosylation does not cause major changes in protein structure, but it reduces protein dynamics, which likely enhances protein's thermodynamic stability [41]. Their findings indicate that glycosylation plays a common role in proteins and that proper glycosylation is necessary for some proteins to achieve their inherent dynamic properties. Thus, N-glycans might function as molecular glue, holding residues around glycosylation sites through favorable interactions and leading to reduced protein dynamics [41]. However, the study also highlighted exceptions to this general trend, particularly the reduced thermodynamic stability of tyrosinase and related proteins upon glycosylation [156].

This paragraph leads us to a question of this thesis: "How do N-glycans specifically influence protein dynamics, and why do their effects vary among different proteins? Furthermore, do these effects vary for identical proteins when mutated, and if so, why or why not?"

To address these questions, detailed MD simulations can be used. By utilizing an appropriate force field, water model, and ensuring well-equilibrated sampling, MD simulations can offer accurate insights into the underlying molecular mechanisms of glycoprotein flexibility such as rotameric states of glycosidic linkages. Thus, in the current study, MD simulations are used to investigate the rotameric states of the glycans and their impact on protein's conformational dynamics in the selected glycoprotein AGP and its mutants, using carb-RAMA plots. The detailed analysis and results will be discussed in the following chapters.

Chapter 2

Research questions

Proteins are dynamic entities characterized by inherent motions and conformational changes that enable their functional diversity. The conventional understanding of protein behaviour has been based on amino acid sequences and static protein folds. However, many proteins exhibit highly dynamic or structurally ambiguous features that cannot be adequately represented by fixed static coordinates. Capturing these behaviours requires protein dynamics data, typically obtained from NMR spectroscopy and MD simulations, including

- **correlating multiple protein conformations with physiological contexts, such as PTMs like glycosylation**: In the context of impact of PTMs on protein conformational dynamics and flexibility, the inherent flexibility of glycans in glycoprotein complexes often results in insufficient or missing experimental data. To address this, 3-D modeling of glycan-protein complexes combined with MD simulations presents a promising solution, although it is computationally expensive.

- **interpreting their conformational states and flexibility on a large scale**: Generally, while the release of AlphaFold2 has democratized access to large-scale static 3-D structures of proteins, analyzing and predicting protein dynamics remains a major challenge—even without considering the effects of glycosylation.

Consequently, understanding proteins' biophysical behaviour and flexibility remains a 'black box,'as existing approaches fall short in capturing the conformational behaviour of proteins—both in relation to glycosylation and, and, more generally, at a large scale, highlighting the need for integrative computational and experimental approaches to investigate protein flexibility. Along these lines, the primary objective of this Ph.D. thesis is **to uncover the protein conformational dynamics and flexibility** from two distinct perspectives: a detailed, "**close-up**" perspective and a broader, "**panoramic**" perspective.

The close-up perspective involves investigating,

1. ***How mutations and glycosylation, individually and in combination, affect the conformational dynamics and flexibility of a (glyco)protein α-1 acid glycoprotein (AGP)?***

Thus, the first key research question will be referred to as **AGP dynamics**. In contrast, the panoramic perspective involves investigating,

2. ***what is the relationship between protein flexibility predicted by computational methods and observed through experimental techniques on a large scale?***

The second key research question will be referred to as **Gradations of protein dynamics**. This thesis will detail and address these two key research questions, each focused on different contexts of protein flexibility.

2.1 AGP dynamics

In the above context, the close-up perspective provides a detailed examination of AGP and its mutants, before and after glycosylation, focusing on both local and global conformational dynamics of the protein backbone and flexibility, solvent accessibility, and glycan dynamics. In line with this key research question, the project includes the following objectives:

1. **To assess how glycosylation alters AGP's backbone flexibility**. Using the X-ray crystallography structure of AGP as a starting point, which lacked glycans, five AGP-specific glycans were sourced from open-access mass spectrometry glycan databases and subsequently modeled using open-source modeling tools. MD simulations were then conducted on both glycosylated and unglycosylated AGP structures to evaluate changes in local backbone flexibility induced by glycosylation using RMSF. At the local level, the dynamics within three key regions of the protein conformation were examined: the central binding cavity, the entrance to this cavity, and the closure of the cavity, which forms a potential protein-protein interaction site.

2. **To explore how selected cancer-associated mutations in AGP affect flexibility, both before and after glycosylation**. To achieve this, the amino acid sequences of relevant cancer mutations in AGP were obtained from open-source cancer mutation databases, and both their glycosylated and unglycosylated structures were modeled. MD simulations were then performed on all the mutant structures to analyze the global changes in backbone flexibility.

3. **To analyze the similarities and differences in the conformational dynamics among all glycosylated and unglycosylated systems of AGP, including both the wild-type and mutant structures**. To achieve this, Principal Component Analysis (PCA) was applied to each system using their MD trajectories, providing a comparative view of their backbone dynamics by Root-mean-square-inner-product (RMSIP).

4. **To examine the effect of glycosylation and mutations on the local solvent accessibility of AGP and its mutants**. For this analysis, three probes with different radii, representing a water molecule (0.14 nm), a small molecule (0.5 nm), and a peptide (1.0 nm), were used to assess the solvent accessibility of all AGP systems. This analysis was conducted to evaluate how glycosylation influences the accessibility of these molecules to AGP's three key regions (outlined in Objective 1), while also accounting for the backbone dynamics of the protein revealed by the previous objectives.

5. **To investigate how mutations disrupt conformational dynamics of glycans**. For this analysis, distributions of Phi and Psi torsion angles of all glycan chains for glycosylated AGP and its systems were plotted on a Carb-Ramachandran (carb-RAMA) plot. The goal of this analysis was to provide a complete picture of glycan dynamics in all five glycan chains in all glycosylated systems of AGP, and the individual saccharides in each glycan chain.

2.2 Gradations of protein dynamics

Given the high computational cost of MD simulations and the challenges of inferring dynamics from multidimensional NMR data, the focus is on integrating protein dynamics data from both experimental and computational approaches. This involves leveraging NMA, a coarse-grained and computationally efficient method, alongside sequence-based predictive approaches for backbone dynamics, to serve as a unifying framework for advancing our understanding of protein dynamics on a large scale. Thus, the central objective is to establish **a link between pLDDT values of AlphaFold2 structures and protein flexibility metrics derived from various experimental and computational methods**, including NMR order parameters, MD simulations, and NMA.

This research question was addressed in collaboration with a fellow Ph.D. student Jose Gavalda Garcia as a shared first author and with other contributors from the Bio2Byte lab. The primary data collection, dataset curation and analysis of the project was carried out by Jose, including AlphaFold2 structures dataset, MD dataset, and ShiftCrypt dataset. The remainder of the work, including the analysis of the flexibility of AlphaFold2 structures and their corresponding NMR structures using NMA, is addressed in the current Ph.D. thesis. In line with the individual contributions, the research objectives are as follows:

1. **To assess the link between pLDDT and backbone flexibility of AlphaFold2 structures as predicted by chemical shift derived S^2_{RCI} values, experimental S^2 order parameters** as well as the differences in secondary structures between AlphaFold2 models and their corresponding NMR models.

2. **To compare the correlation between RMSF profiles derived from NMA and S^2_{RCI} values for both AlphaFold2**

models and their corresponding NMR ensembles. This comparison aimed to assess how well AlphaFold2 models capture protein dynamics relative to the NMR data. For this analysis, NMA was carried out on AlphaFold2 models and their associated NMR ensembles. From the resulting eigenvalues and eigenvectors from NMA, RMSF profiles were computed for each model. In the next step, the correlation coefficients were computed between the RMSF profiles and S^2_{RCI} values for each AlphaFold2 model and its associated NMR model within the NMR ensemble.

Chapter 3

Methodological background

*This chapter provides a detailed overview of the computational and experimental metrics utilized in this thesis to study protein flexibility, as outlined in the key research questions, with a focus on **AGP dynamics** and the **Gradations of protein dynamics**. The chapter begins with MD simulations, covering their underlying algorithms, force fields, statistical mechanics principles, and methodological framework. This is followed by the discussion of flexibility metrics derived from MD simulations such as root-mean-square fluctuations (RMSF), NMR chemical shifts as S^2_{RCI}, DynaMine, and ShiftCrypt. The chapter concludes with a discussion of Normal Mode Analysis (NMA) and NMA-derived RMSF. The goal is to offer detailed guidance and additional context that supplements the information presented in the key research questions.*

3.1 MD Simulations

As introduced in Chapter 1, all-atom MD simulations are a powerful and widely used computational approach for generating the conformational ensembles of proteins. This section will delve into the principles of MD simulations and outline the specific simulation parameters utilized in **AGP dynamics** specifically for a glycoprotein.

In the late 1950s, Alder and Wainwright pioneered simulations of liquids [157]. They explained that to follow the dynamics of a many-particle system with any interaction potential, one could calculate the force on each particle at any given moment by considering the influence of its neighbours. The particle trajectories could then be traced by allowing them to move under this constant force for a short time interval, followed by recalculating the force for each subsequent interval, and continuing this process iteratively. A decade later, MD became more popular amongst computational chemists and biologists to study and investigate the structure and dynamics of proteins, nucleic acids, and other macromolecules, and remains one of the widely used techniques. MD methods can be categorized into two main families: classical and quantum mechanics, distinguished by their models and the mathematical toolkits used to represent a physical system [158, 159]. In this chapter, we will focus on classical MD simulations.

3.1.1 The algorithm behind MD

Frenkel and Smit defined the algorithm of MD in several key steps, which we will discuss in detail [160]. For a system consisting of N particles, the steps are as following:

1. Initialize the system by selecting initial positions and velocities of all particles in the system.

2. Compute forces acting on all particles in the system.

3. Integrate Newton's equations of motion to compute the positions and velocities of all particles at each discrete time step Δt, continuing until the system evolves for the desired length of time.

4. After completion, compute and visualize averages of measured parameters.

We can now think of a movie of interacting atoms over time, generated by MD simulations. This movie is created by solving Newton's equations of motion eq. 3.1.

$$F_i = m_i \frac{d^2 r_i\,(t)}{dt^2} \tag{3.1}$$

where $r_i(t)$ is the position vector with coordinates $x_i(t)$, $y_i(t)$, $z_i(t)$ of the i^{th} particle, and F_i represents the force acting upon the i^{th} particle at time t, while m_i is the mass of the particle. The force $F_i = -\nabla_k U(r_1, r_2, ..., r_N)$ where the potential energy U is a sum of contributions from bonded and non-bonded interactions (discussed in next section). Assuming that all the particles and molecules in the movie are purely classical with a fixed electron density, and constant inter-atomic force laws in time, the dynamics in this case are Newtonian and deterministic [14]. Once the atoms start moving, the biophysical behaviour of the system is determined by initial positions $r_i(0)$ and velocities $v_i(0)$. Using numerical integrators, the positions and velocities are propagated with a finite time interval [161]. An example of numerical integrator is the Verlet algorithm which uses Taylor series expansion. The integrator is required to figure out the positions $r_i(t + \Delta t)$ and velocities $r_i(t + \Delta t)$ after a short time-step, Δt from the already known positions at time t. The Δt is determined by the fastest motions of the system. Newton's equations conserve the total energy, so the numerical solutions should ideally preserve this energy. To ensure stability and accuracy in energy conservation, the time step should generally be an order of magnitude smaller than the fastest time scale in the system [162]. A larger time step can limit the length of the MD trajectory and lead to instability, causing energy to increase rapidly over time. For example, the vibration of the O-H bonds occurs in around 10 fs scale, therefore, to ensure stable integration, the Δt should generally be at least two times smaller than the period of the fastest vibration [163]. The typical choice of Δt in MD for proteins is 2 fs [164] which is also used in **AGP dynamics**.

3.1.2 The force field

To obtain the dynamics of a complex system such as a protein, glycan and/or a glycoprotein molecule, an interaction potential is required to compute the forces on each amino acid or saccharide due to its neighbouring amino acids and/or saccharides at any instant. The interaction potentials between specific chemical bonds within a physical system (biomolecules in our case) are thus described by atomic force field models. Based on Meller et al.[161], the force field model and derivation is adapted for the discussion below. In a force field model, the interactions within a system are specified by the potential $U(r_1, ..., r_N)$, which denotes the potential energy of N atoms that are interacting as a function of their positions $r_i = (x_i, y_i, z_i)$. The force acting upon the i^{th} atom is estimated by the gradient as shown in eq. 3.2

$$F_i = -\nabla_{r_i} U(r_1, ..., r_N) = -\left(\frac{\partial U}{\partial x_i}, \frac{\partial U}{\partial y_i}, \frac{\partial U}{\partial z_i}\right) \tag{3.2}$$

Unlike quantum methods, classical MD uses the Born-Oppenheimer approximation to simplify molecular systems by treating the movements of nuclei and electrons separately, based on their mass differences because the nuclei are much heavier and move much more slowly than the electrons. This leads to the potential energy surface (PES) which describes the dynamics of nuclei, while ignoring the detailed behaviour of electrons. Therefore, the force fields are a set parametrized functions that approximate the PES by modeling atomic interactions through empirical and ab initio potentials. These potentials approximate the influence of the electrons on the nuclei, while focusing on the motion of the nuclei as if they were interacting particles. Thus, force fields consist of mathematical functions that estimate the system's energy as a sum of various contributions, including bonding interactions and non-bonding interactions. In MD simulations, force field parameters are optimized to match empirical (experimental or ab initio) data, ensuring accurate system behaviour, with a typical force field used in simulations of biomolecules as shown in eq. 3.7

$$
\begin{aligned}
U\left(r_i,...,r_N\right) = & \sum_{\text{bonds}} \frac{a_i}{2}(l_i - l_{i0})^2 && (3.3)\\
& + \sum_{\text{angles}} \frac{b_i}{2}(\theta_i - \theta_{i0})^2 && (3.4)\\
& + \sum_{\text{torsions}} \frac{c_i}{2}[1+\cos(n\omega_i - \gamma_i)] && (3.5)\\
& + \sum_{\text{atom pairs}} \left[4\epsilon_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] && (3.6)\\
& + \sum_{\text{atom pairs}} k\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} && (3.7)
\end{aligned}
$$

are then used in simulations to approximate these interactions, making it easier to study complex systems like biological molecules. In eq. 3.7, the summation in the first three terms is taken over all the bonds, angles, and torsion angles which are defined by the covalent bonding in the system. The remaining two terms are non-bonding terms in which the summation is taken over all the atom pairs including point charges (q_i), which are separated by distances, $r_{ij} = |r_i - r_j|$ (fig. 3.1). The terms corresponding to bonds and angles is used to describe deformation energies of the bond lengths l_i, and bond angles θ_i from their equilibrium positions, l_{i0} and θ_{i0} respectively. These terms with their corresponding force constants a_i and b_i ensures that there is no breaking of the chemical bonds. The next term is used to describe the rotations around the chemical bond including periodicity n, phase angle γ_i in a cosine series, and rotational barrier height c_i. The next term is used to describe the vdW interactions as Lennard-Jones (LJ) potential, followed by the Coulomb electrostatic potential in the last term (fig. 3.1). Different force field models may include additional bonding and non-bonding terms.

In **AGP dynamics**, CHARMM36m force field is applied to account for both protein and the attached N-glycans. In CHARMM36m, additional bonding terms including Urey-Bradley angle, and improper dihedral angle are added. The Urey-Bradley term is defined as a spring connecting the outer atoms of a bonded triplet (fig. 3.1). An improper torsion angle is the dihedral angle, defined using three atoms that are all bonded to a single 'central' atom (fig. 3.1). In CHARMM36m, H-bonding interactions are not treated with a separate term; instead, they are accounted for in the parametrization through a combination of LJ and Coulombic interactions [145]. The CHARMM36m has the same potential energy function for both proteins and glycans [145].

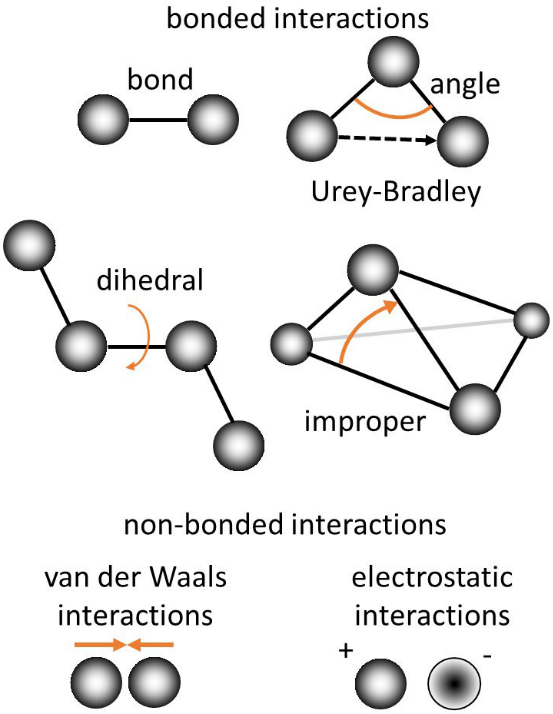


Figure 3.1: Bonded and non-bonded terms in a force field illustrated by a diatomic molecule including bond, angle, dihedral angle, Urey-Bradley angle, and improper torsion, a vdW interactions, and electrostatic interactions.

3.1.3 Periodic boundary conditions

Given the general form of the force field, at each time step, the forces are recalculated by updating the positions and velocities of the atoms. Due to these recalculation steps especially long-range interactions, the computational cost of MD simulations typically scales with the number of atoms, making it computationally expensive, especially for large systems. Also, detailed all-atom representation of solvent is added to the system, to realistically represent the aqueous environment of the biomolecule (protein), which also increases the computational load. To manage these limitations, only a finite portion of an otherwise infinite system can be explicitly modeled on a computer.

This finite model often results in many atoms being positioned near the edges of the simulation box, a situation that is not ideal for deriving the bulk properties of the system because most atoms are influenced by the boundaries [165]. This issue, known as the boundary effect, must be carefully managed in simulations. To accurately simulate bulk systems, Periodic Boundary Conditions (PBC) are typically implemented [165]. The key idea behind PBC is that we can simulate an infinite system by replicating our finite system multiple times, effectively creating a seamless, repeating environment. This allows us to study a single molecule over an extended time scale, instead of having to simulate a large number of molecules, based on the concept of ergodicity from statistical mechanics. The principle of ergodicity asserts that the average properties of an infinite number of identical systems at a specific moment are equivalent to the average obtained by observing a single system over an infinite period [166]. Thus, PBC approach allows for more realistic and accurate simulations of bulk properties in biomolecular systems. Under PBC, long-range interactions can be managed in one of two ways: (i) by truncating pairwise Coulomb interactions at a specified cutoff distance, or (ii) by using lattice summation methods, both of which rely on certain approximations which may introduce errors in the calculated interactions [167]. To address the limitations of these conventional approaches, optimized lattice summation methods such as Particle-Mesh Ewald (PME) methods are now commonly employed for more accurate treatment of PBC for explicit-solvent simulation of proteins in solution [167].

3.1.4 Statistical mechanics in MD

Referring to the previous example of statistical ensemble from Anfinsen's test tube from Chapter 1, a statistical ensemble represents the collection of all possible microstates a system can occupy and an associated probability distribution over the collection, sharing common macroscopic properties such as volume (V), energy (E), or the number of particles (N). For instance, if Anfinsen's test tube contained N particles moving in three dimensions, $6N$ real numbers would be needed to describe the system's physical state at any given moment as shown in eq. 3.8. Three of these numbers would specify each particle's position (x, y, z), while the remaining three would represent each particle's momentum (p_x, p_y, p_z).

$$(x_1, y_1, z_1, p_{x_1}, p_{y_1}, p_{z_1}, \ldots, x_N, y_N, z_N, p_{x_N}, p_{y_N}, p_{z_N}) \in R^{6N} \quad (3.8)$$

If Anfinsen were to know the exact microstate of the proteins in his test tube, assuming classical physics governs the system, he could predict the past states and future states of the protein. However, in reality, it is impossible to determine the positions and momenta of every protein molecule in the test tube. Instead, measuring macroscopic properties is more feasible. A macrostate encompasses all the microstates corresponding to macroscopic properties.

An ensemble and macrostate are related but not the same. An ensemble is a theoretical construct that represents a collection of possible microstates of a system, while a macrostate is a description of the system's overall state based on macroscopic properties. Thus, if Anfinsen knows the macrostate of his protein, he understands that the protein molecules in his test tube are equally likely to truly be in any of the microstates they contain. A system is therefore most likely to reside in a macrostate that corresponds to the greatest number of microstates [168]. Here, a macrostate would describe, for example, 50 µg/ml of lysozyme protein in Anfinsen's test tube, at pH 7.5, with a salt concentration of 50 mM KCl, a temperature of 300 K, a pressure of 1 bar, and a volume of 100 ml. In this macrostate, a vast number of microstates would then describe the positions and velocities of all atoms in the system, including proteins, water molecules, and ions. For instance, the system might contain approximately 10^{17} protein molecules, 10^{23} water molecules, and 10^{21} ions, each contributing to the overall configuration of microstates. In practice, rather than tracking the deterministic motions of individual particles using Newtonian equations of motion for infinite sampling time, significant insights about a system can be gained by focusing on its probabilistic configurations near equilibrium. This statistical mechanics approach links the microstates of a system to macroscopic thermodynamic properties like temperature T, pressure P, and energy E [169]. The Boltzmann factor $e^{\frac{-E}{k_{\mathrm{B}}T}}$ quantifies the relationship between energy and probability by assigning probabilities to different microstates based on their energy E. A one-dimensional relationship describing the Boltzmann factor is shown between probability and potential energy in eq. 3.9. In combination with temperature, T, potential energy U governs key aspects of the system, including its probability distribution [170]. This relationship allows us to predict protein dynamics statistically.

$$\mathrm{pdf}(x) \equiv p(x) \propto e^{\frac{-U(x)}{k_{\mathrm{B}}T}} \tag{3.9}$$

where p(x) or pdf(x) is the probability density for the system to be in a particular microstate x. Here, x describes a specific conformation or a microstate of the system such as the set of atomic positions in the protein, $U(x)$ is the potential energy of the conformation x, k_{B} is the universal Boltzmann constant, and T is the absolute temperature. This probabilistic framework is vital for understanding the statistical behaviour of proteins and their fluctuations in various states. In standard unit system, k_{B} is equal to 1.380649×10^{-23} J/K.

According to this factor, the probability of a microstate decreases exponentially with higher energy, making high-energy states rare. These high-energy points, known as barriers in the energy landscape, must be crossed for the system to transition between states. As a result, the system fluctuates around a minimum energy configuration but occasionally transitions to another state. Thus, to replicate experimental conditions in MD simulations, it is not enough to simply explore the energy landscape using Newton's equations. It is necessary to maintain constant temperature and pressure during simulations. Depending on which state variables such as thermodynamic variables such as N, V, E, T, and P are held constant, different statistical ensembles or more specifically thermodynamic ensembles can be generated [169]. From these ensembles, various structural, energetic, and dynamic properties can be calculated by analyzing the averages or fluctuations of the relevant quantities.

The various ensembles in MD are constant-energy constant-volume ensemble (NVE), constant-energy constant-volume (NVT), and constant-energy constant-pressure (NPT) (Fig. 3.2, A). In MD simulations, when we talk about using a certain thermodynamic ensemble (e.g., NVT, NPT), we are specifically referring to the kind of constraints imposed on the simulation to mimic real physical conditions. These ensembles are subsets of statistical ensembles that obey particular thermodynamic principles. The NVE ensemble, or microcanonical ensemble, is generated by solving Newton's equations of motion without regulating temperature or pressure, ensuring that the total energy remains conserved [169]. In contrast, the NVT ensemble, also known as the canonical ensemble, maintains a constant temperature by applying direct temperature scaling during initialization and coupling to a thermal bath [169]. The NPT ensemble provides control over both temperature and pressure throughout the simulation [169]. Prior to running MD simulations including NVT and NPT runs, it is essential to prevent the simulation from collapsing due to high energy arising from steric clashes or inappropriate geometry. To address this, energy minimization is performed to eliminate the modeling and experimental artifacts by optimizing the system's geometry toward a lower or minimum energy state. For proteins, the computational cost of searching the entire energy landscape and the number of local minima make it unfeasible to find the global energy minimum [171]. Due to this reason, it is more practical to search a local minimum near the X-ray structure, if available. One commonly used energy minimization algorithm is the steepest-descent method [171]. In this approach, at each step, a displacement opposite to the potential energy gradient is applied to the atomic coordinates. The step size is adjusted based on whether a lower energy is achieved, with the step size being increased if energy decreases and reduced if not. Although this method can be slow to converge, its gentle positional shifts make it effective for small adjustments, such as removing steric clashes. In the MD simulations related to AGP dynamics, following the energy minimization, the NVT ensemble is first applied to bring the system to the desired temperature, allowing the solvent and ions to adjust to the temperature based on the kinetic energies, thereby establishing the proper orientation around the solute (the protein). This is followed by the NPT ensemble, where pressure is applied to ensure that the system reaches the correct density. The NVT and NPT equilibration phases help establish proper solvent orientation and pressure around the protein. After completing these two equilibration steps, the system may be well-equilibrated at the desired temperature and pressure, with proper solvent and ion distribution [172]. However, since equilibration is difficult to guarantee, it is generally verified by checking that properties, such as temperature, pressure, and energy, stabilize and remain constant over time. Once the system has equilibrated, the position restraints applied during equilibration can be released, and the system is ready for the production phase [172]. In this phase, unrestrained

MD simulations are run to collect data for further analysis, with the system now stable and equilibrated. These general steps in MD simulations are outlined in (Fig. 3.2, B).

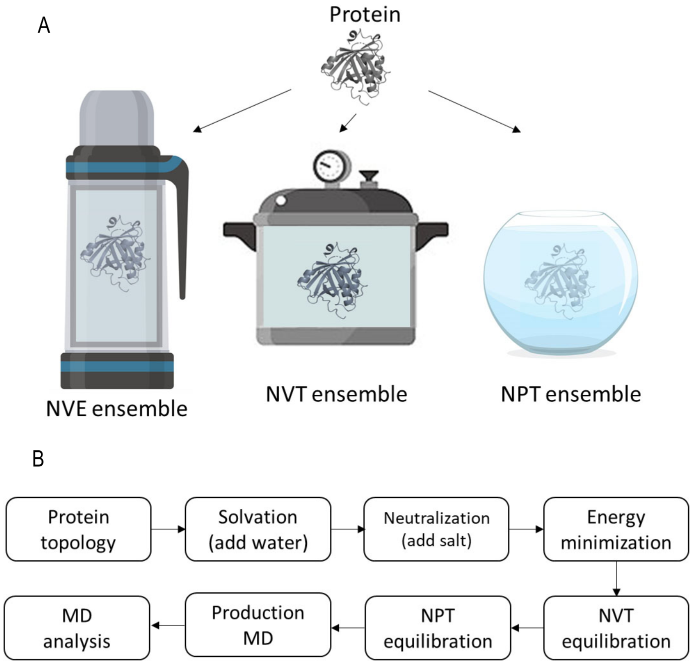


Figure 3.2: Statistical ensembles in MD simulations (A) Thermodynamic ensembles in MD simulatons representing a protein structure in an NVE, NVT, and NPT ensemble. (B) An overview of the general steps involved in a MD simulation.

3.2 Computational and experimental metrics of flexibility

Given the high computational cost of molecular dynamics (MD) simulations and their reliance on a fully resolved 3D structure to assess protein flexibility, MD simulations is often impractical for large-scale studies. In such cases, sequence-based approaches offer a more feasible alternative, providing a reliable approximation of protein flexibility without the need for computationally intensive simulations. For large-scale analysis of proteins' biophysical behaviour, two sequence-based tools from b2bTools https://bio2byte.be/b2btools/, DynaMine and ShiftCrypt. For **AGP dynamics**, DynaMine from b2bTools https://bio2byte.be/b2btools/ was used to predict the backbone dynamics propensity of 59 mutants of AGP, to identify mutants that were most likely to alter the dynamics of AGP. For **Gradations of protein dynamics**, ShiftCrypt and additional computational metrics of flexibility were used including Random Coil Index (RCI) based S^2_{RCI} order parameters from NMR, and NMA derived RMSF. As these flexibility metrics, except for NMA-derived RMSF, rely on NMR-derived chemical shifts (CS) as the foundation for their predictions, the next sections discuss CS, RCI and S^2_{RCI}. This is followed by a brief discussion of computational metrics of backbone flexibility that leverage CS data such as DynaMine and ShiftCrypt. Finally, the section concludes with the underlying theory of NMA and the equations used for calculating RMSF from normal modes.

3.2.1 Fluctuations from MD simulations

Understanding a protein's atomic motions is crucial for gaining insight into its biophysical behaviour and dynamics across various timescales. While obtaining this information experimentally can be challenging, it can be effectively derived from MD simulations [96]. Considering a protein's trajectory obtained from MD simulations, the root-mean-square deviation (RMSD) is often used to evaluate the overall structural deviation of a protein from a reference structure over time (eq. 3.10) [173]. RMSD analysis typically does not include all coordinates in a protein structure, as fluctuations in residue side chains do not reflect overall conformational changes. Therefore, when RMSD is used to study large-scale movements in proteins, the analysis is usually limited to backbone atoms (which form the amide-bond chain) or C_α atoms.

$$RMSD(x, x^{\text{ref}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|x_i - x_{\text{ref},i}\right|^2} \tag{3.10}$$

where x_i represents the position of the i^{th} atom in the current structure from the simulation trajectory after optimal superposition, x^{ref} is the position of the reference structure such as the starting structure, n is the number of atoms. The RMSD metric is commonly used to monitor the convergence of MD simulations, where minimizing RMSD over time suggests that the system has reached thermodynamic equilibrium. The RMSD has the units of length.

Despite its practical use, RMSD has limitations [174]. It depends on the alignment and gives a single value for the entire structure; if the alignment is inaccurate, the RMSD becomes meaningless. In MD analysis, alignment involves superimposing structures by minimizing the RMSD between corresponding atoms or residues to enable accurate comparison of conformations. Also, it fails to distinguish between rigid and flexible regions of a molecule, which can lead to misleading results, as large RMSD values may arise from small, highly flexible regions, even if the overall structure remains mostly unchanged. To address this, a mass-weighted RMSD is used (eq. 3.11), where certain distances have less impact, but this can slow down calculations. Although alternative methods exist, they are often slower and less reliable in finding global optimum alignments. In the case of C_α RMSD, the mass weighting typically does not affect the calculation. RMSD remains the fastest and most widely used metric for structural similarity, especially when dealing with large datasets.

$$RMSD(x, x^{ref}) = \sqrt{\frac{\sum_{i=1}^{n} w_i \left| x_i - x_{ref,i} \right|}{\sum_{i-1}^{n} w_i}} \tag{3.11}$$

where w_i is the mass of the i^{th} atom.

The importance of computing fluctuations becomes evident when considering the limitations of using averages, such as average RMSD over the MD trajectory, to describe a protein's behaviour. While averages provide a broad overview of the protein's overall dynamics, they fail to capture conformational changes reflected in the tails of the distribution [170]. Fluctuations are typically described by variance, which quantifies the spread of the data, and the square root of the variance, or standard deviation, represents the width of the distribution. RMSF is the square root of the time-averaged squared deviations of atomic positions from their average position during the simulation, and is calculated using the following equation,

$$RMSF_i = \sqrt{\langle (x_i - \langle x_i \rangle)^2 \rangle} \tag{3.12}$$

where x_i represents the position of the i^{th} atom and $\langle x_i \rangle$ is the average position of i over the trajectory. In comparison with RMSD, RMSF is used to investigate the regions with high mobility. Regions exhibiting high RMSF values are generally more flexible, whereas regions with low RMSF values tend to be more rigid. In **AGP dynamics**, the RMSF of each heavy atom was computed as the root of the variance of its coordinates across the trajectory for each system, yielding a RMSF profile as a function of the amino acid residue index.

3.2.2 Backbone dynamics from DynaMine and ShiftCrypt

On a large scale, the biophysical features are not readily captured by MD and/or other computational and experimental approaches. DynaMine captures the 'emerging' protein backbone dynamics property, as determined by local interactions between amino acids, while ShiftCrypt reflects the biophysical state of an amino acid residue [99, 101]. Both tools utilize chemical shifts, but only DynaMine includes the RCI. In contrast, ShiftCrypt is unbiased and does not depend on a reference dataset like random coil chemical shifts.

3.2.2.1 Chemical shifts

NMR provides the atomic-level chemical shifts, which are exquisitely sensitive to their environment [101]. These chemical shifts provide an averaged view of local dynamics and are highly sensitive to their environment, making them valuable for studying a wide range of proteins, from fully folded to disordered [101]. NMR works by analyzing the interaction of radio-frequency electromagnetic radiation with nuclei placed in a strong magnetic field [175]. High electron density around a nucleus creates a shielding effect, reducing the net magnetic field the nucleus experiences. This weaker magnetic field causes the nucleus to precess and absorb radiofrequency radiation at a lower frequency to achieve resonance. Each nucleus in a molecule is in a slightly different chemical environment, leading to variations in electron shielding and, therefore, slightly different resonance frequencies. These small differences are what allow NMR to distinguish between different nucleus in a molecule, even though the variations in frequency are minimal. Measuring the exact resonance frequencies with precision is challenging, so instead of determining the precise frequency of each nucleus, a reference compound tetramethylsilane (TMS) (in ^{1}H NMR) is added to the solution of the sample. The resonance frequency of each nucleus in the sample is then measured relative to the resonance frequency of the nucleus in the reference compound.

$$shift\ (Hz) = frequency\ (sample) - frequency\ (TMS) \tag{3.13}$$

When measuring the sample, the nucleus resonances are reported as the shift (in Hz) relative to those of TMS. This shift depends on the strength of the applied magnetic field. For example, in a magnetic field of 1.41 Tesla, a nucleus resonates at around 60 MHz, while in a 2.35 Tesla field, the resonance occurs at about 100 MHz (eq. 3.14). The ratio of the resonance frequencies corresponds directly to the ratio of the magnetic field strengths. For a given nucleus, the chemical shift (in Hz) relative to TMS is $\frac{5}{3}$ parts per million (ppm) greater in the 100 MHz range than in the 60 MHz range.

$$\frac{100\text{ MHz}}{60\text{ MHz}} = \frac{2.35\text{ Tesla}}{1.41\text{ Tesla}} = \frac{5}{3}\text{ppm} \tag{3.14}$$

Due to this difference, a more standard approach is applied by using a parameter independent of field strength, known as chemical shift, (δ) (eq. 3.15). This is achieved by dividing the shift in Hz by the spectrometer frequency in MHz. The chemical shift expresses how much a nucleus's resonance is shifted from TMS in parts per million (ppm) of the spectrometer's base operating frequency. The δ values for a given nucleus remain constant regardless of whether the measurement is made at 100 MHz or 60 MHz. The frequency ratio in eq. 3.15 is multiplied by 10^6 to obtain a mathematically convenient number expressed in ppm.

$$\begin{aligned}\delta &= \frac{shift\ (Hz)}{spectrometer\ frequency\ (MHz)} \\ &= \frac{shift\ (Hz)}{spectrometer\ frequency\ \times 10^6 (Hz)} \times 10^6 ppm\end{aligned} \tag{3.15}$$

For example, in CH3Br, the nucleus shift is 162 Hz at 60 MHz and 270 Hz at 100 MHz, but both correspond to the same δ value of 2.70 ppm:

$$\delta = \frac{162\text{ Hz}}{60\text{ MHz}} = \frac{270\text{ Hz}}{100 \times 10^6\text{ Hz}} = 2.70 \times 10^{-6} \times 10^6\text{ ppm} = 2.70\text{ ppm} \tag{3.16}$$

In proteins, chemical shifts (CS) are typically interpreted as a metric for local structure and can be used to identify torsion angles, H-bonding, secondary structure elements [177]. Thus, the CS values of atoms in an amino acid residue are closely linked to its conformational characteristics within a protein, correlating with the backbone angles the protein can adopt in its folded state, the backbone's flexibility (Fig. 3.4), and its solvent accessibility (Fig. 3.3) [178]. The CS values of two different types of amino acids that experience the same local environment will vary due to the differences in their chemical compositions [178]. In contrast, the same CS values could be observed for atoms in two different amino acids of the same type, regardless

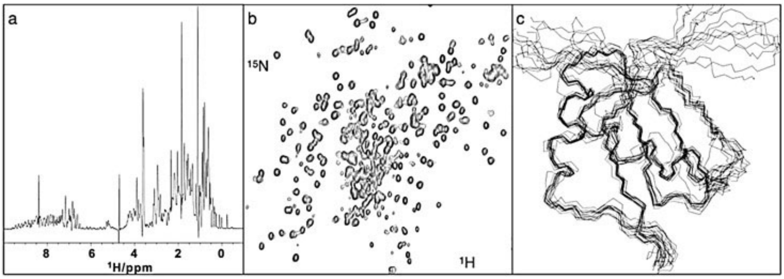


Figure 3.3: NMR chemical shifts NMR data represented as: (a) a one-dimensional spectrum depicting ^{1}H chemical shifts observed in a protein, (b) a 2-D spectrum showing the interaction between ^{1}H and $^{1}5N$ chemical shifts of bonded nuclei within the protein, and (c) an ensemble of protein structures fitted to NMR data and with backbone representation. The image is sourced from Ref.[176].

of their distinct local environments. Additionally, CS data can detect slow protein motions occurring on μs and even longer timescales [179]. CS values are commonly used to estimate biophysical properties, but this can be challenging due to the multidimensional nature of the data. Thus, CS information can be interpreted through 2-D correlations (Fig. 3.3, b), considering the relation between the CS values of two atoms at the same time. Building on this concept, tools like ShiftCrypt, offer a more advanced method for analyzing, comparing, and interpreting CS directly within their native multidimensional space [178]. The ShiftCrypt method utilizes a neural network based auto-encoder architecture to compress per-amino acid chemical shift data into a single, interpretable value that is independent of amino acid type. This value reflects a residue's biophysical state or conformational preferences. The encoded values range from 0 to 1, representing a spectrum of conformational and dynamic states: values near 0 indicate a preference for helices, values near 1 signify a preference for sheet structures, and values around 0.5 represent dynamic behaviour or multiple conformations [178]. In this Ph.D. thesis, the relationship between ShiftCrypt values and pLDDT scores of AlphaFold2 structures was examined in the context of **Gradations of protein dynamics**, with ShiftCrypt values representing the predicted conformational preferences of AlphaFold2 structures in solution.

3.2.2.2 RCI and S^2_{RCI}

To mitigate errors in obtaining protein dynamics parameters derived from CS, Berjanski and colleagues developed a simple metric for assessing protein flexibility known as the RCI [180]. The RCI metric is

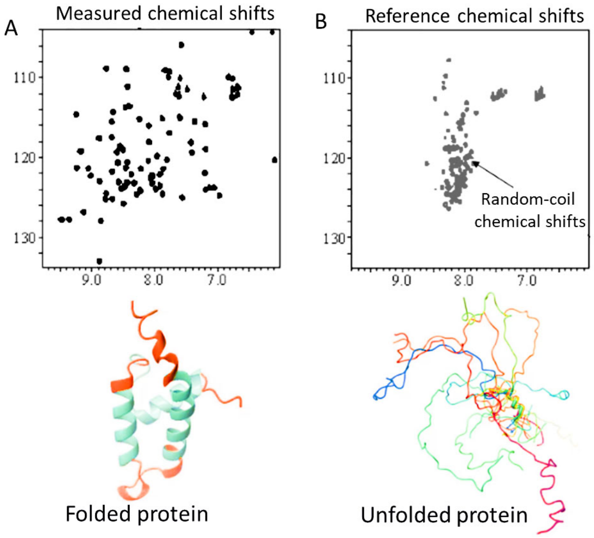


Figure 3.4: Random coil chemical shifts (A) A 2-D plot illustrating the measured chemical shifts in a folded protein compared to (B) a 2-D plot of reference random-coil chemical shifts derived from an ensemble of unfolded proteins. The differences between measured and reference shifts are used to calculate secondary chemical shifts, which are then transformed into the the RCI [180]. The image is sourced and adapted from Ref.[181].

based on an empirically derived relationship between secondary CS values and protein mobility, and can be correlated with standard protein motion metrics such as per-residue RMSF from MD simulations, model-free order parameters S^2, and per-residue root-mean-square deviations (RMSD) from NMR ensembles. Therefore, in addition to identifying the conformational preferences of proteins, CS have been used to predict protein structural disorder by transforming them into RCI (eq. 3.17) [180]. Based on the calculation by Berjanski et al.[180], this transformation was based on their reference data set of ^1H, ^{13}C, and ^{15}N chemical shifts assignments from 28 well-resolved proteins, and the weights for each shift were determined through an optimization process (Fig. 3.4). RCI was therefore computed as the inverse weighted sum of observed secondary chemical shifts for C_α, CO, C_β, N, NH, and H_α, incorporating several transformation steps such as smoothing secondary shifts across adjacent residues, applying neighboring residue corrections, re-referencing chemical shifts, filling gaps, scaling chemical shifts, and making numeric adjustments to avoid divide-by-zero errors. The application of secondary chemical shifts to characterize protein flexibility is based on an assumption that the close proximity of chemical shifts to random coil values is a manifestation of increased protein mobility, while significant differences from random coil values is an indication of a relatively rigid structure. The resulting RCI values were then correlated with the protein backbone mobility, which was assessed using RMSF obtained from MD simulations of the reference data set (4 ns per protein).

$$RCI = \left[\frac{\begin{matrix} A|\Delta\delta_{C\alpha}| + B|\Delta\delta_{CO}| + C|\Delta\delta_{H\beta}| + \\ D|\Delta\delta_{N}| + E|\Delta\delta_{NH}| + F|\Delta\delta_{H\alpha}| \end{matrix}}{n} \right]^{-1} \tag{3.17}$$

where $|\Delta\delta_{C\alpha}|$, $|\Delta\delta_{CO}|$, $|\Delta\delta_{H\beta}|$, $|\Delta\delta_{N}|$, $|\Delta\delta_{NH}|$, $|\Delta\delta_{H\alpha}|$ are the absolute values of the secondary CS (in ppm) of C_α, CO, H_β, N, NH, and H_α, respectively, while nucleus-specific weighting coefficients are given from $A - F$. The number of CS types is denoted as n.

The RCI is a unitless index. Formally, the RCI represents chemical shifts arising from fast conformational exchanges among energy-weighted populations of all theoretically possible conformations of an unfolded polypeptide chain in the absence of long-range inter-residue interactions [182, 183]. Consequently, as the structure and mobility of a protein segment approach a random coil state, the chemical shifts of its atoms converge toward their corresponding random coil values. Many research groups have used the proximity of amino acid chemical shifts to these random coil values to qualitatively assess the degree of protein structural disorder. The RCI score close to zero indicates highly ordered residues and increases to 0.55 for very dynamic residues [184].

Berjanskii et al. [184] also proposed a scaling formula to predict S^2 order parameters directly from the RCI score (denoted as S^2_{RCI}) (eq. 3.18). The correlation between the RCI values and S^2 values derived from short MD simulations was further validated using a dataset of 12 proteins with experimentally determined S^2 order parameters. Generally, S^2 order parameters are the bond vectors computed from NMR relaxation data that represent the amplitude of internal motions on a $ps - ns$ timescale. Higher S^2 values, approaching 1, signify restricted bond vector motion, while lower values, nearing 0, indicate greater motional amplitude and increased flexibility [185]. S^2_{RCI} values range from $0 - 1$, with values close to one indicating high rigidity and values close to zero indicate high flexibility [179]. Thus, the

S^2_{RCI} values and RCI scores exhibit an inverse relationship in their interpretation.

$$S^2_{RCI} = 1 - 0.4 \ln(1 + 17.7RCI) \tag{3.18}$$

Alongwith ShiftCrypt index, S^2_{RCI} values were also computed for AlphaFold2 structures in **Gradations of protein dynamics**, to assess the conformational flexibility of AlphaFold2 structures in solution.

3.2.2.3 DynaMine

The NMR metrics of flexibility mentioned above clearly show that protein disorder is linked to dynamics. However, due to the dynamic nature of proteins, it is challenging to determine whether an amino acid residue can exhibit behaviour in just two distinct states of order and disorder. Protein disorder is 'context-dependent,' with many residues in non-globular proteins showing a 'dual personality,' where their behaviour varies according to environmental conditions [74, 179]. Within disordered regions, there are also nuances: a disordered residue can exist in several conformational states, each with its own frequency of occurrence. Therefore, by predicting protein disorder based on dynamics derived from chemical shifts, the DynaMine tool offers valuable insights into the conformational flexibility and disorder of protein regions [99]. It is based on linear regression approach and predicts per-residue scores on the fast movements of the protein backbone as the backbone dynamics propensity, indicating how amino acid residues behave dynamically, without relying on $3-$D structures. It is trained on carefully curated collection of chemical shifts for $2,015$ proteins ranging from fully folded to disordered ones. Compared to the existing disorder predictors, DynaMine identifies protein disorder without using prior disorder information directly from amino acid sequence, instead depending on the underlying physical dynamics data [99]. DynaMine assigns per-residue scores where values greater than 0.8 suggest rigid conformations, scores above 1 indicate membrane-spanning regions, values below 0.69 reflect flexible regions, and scores between 0.69 and 0.8 are classified as context-dependent, referred to as ambiguous. As a result, DynaMine provides clear insights into the protein disorder based on amino acid sequence. Due to its scalability and efficiency in analyzing large-scale protein sequences, DynaMine was used to predict the impact of mutations on the backbone dynamics of AGP, as discussed in **AGP dynamics**.

3.2.3 Normal mode analysis

Various studies have shown that the low-frequency motions are functionally relevant, as these motions are thought to arise through evolutionary processes rather than random occurrences [81, 186, 187]. These low-frequency collective motions are associated with large-amplitude conformational changes, are commonly studied using NMA, a widely utilized theoretical framework (Fig. 3.5, B) [188]. In theory, by assuming that the system is stabilized by a harmonic potential, NMA offers insights into the equilibrium modes accessible to such systems [81]. A harmonic potential can be visualized as a smooth, bowl-shaped energy landscape—a mathematical approximation where the protein resides at the bottom of the bowl at equilibrium (Fig. 3.5, A). When the protein deviates slightly from this equilibrium position, the energy increases predictably in a parabolic manner. NMA leverages this assumption to model and predict how proteins vibrate or move near their equilibrium state. Under this assumption, the theory of normal modes provides a complete analytical solution to the equations of motion for a molecular system.

3.2.3.1 Steps involved in NMA

Ghysels et al. [189] described the steps underlying NMA for a system, such as a protein structure, as follows:

1. Perform a geometry optimization to identify a stationary energy point on the PES.
2. Calculate the Hessian matrix by determining the second derivatives of the potential energy with respect to the nuclear displacements.
3. Apply mass-weighting to the Hessian matrix and diagonalize it to derive a set of eigenvalues and eigenvectors.
4. Use the eigenvalues to obtain the vibrational frequencies and the eigenvectors to determine the corresponding normal modes.

3.2.3.2 Theory of NMA

As shown earlier in eq. 3.1, solving the equation of motion directly can be computationally demanding. To improve efficiency, NMA approximates the atomic system as small displacements around the equilibrium position R, where the net force on N atoms is zero. The solutions describe oscillations about equilibrium, with characteristic angular frequencies and corresponding displacement patterns as eigenvalues and eigenvectors of a ‘dynamical matrix’ [81]. The complete derivation of solutions described below is adapted from [81]. To compute this matrix, the potential energy U around the equilbrium R can be written as Taylor series expansion and truncated,

$$\begin{aligned} U(r) =& U(r)\Big|_R \\ &+\sum_i \frac{\partial U}{\partial r_i}\Bigg|_R (r_i - R_i) \\ &+\frac{1}{2}\sum_{i,j} \frac{\partial^2 U}{\partial r_i \partial r_j}\Bigg|_R (r_i - R_i)(r_j - R_j) \end{aligned} \tag{3.19}$$

where r_i and r_j are the position vectors of the atoms i and j. In eq. 3.19, the first term is the potential energy evaluated at the equilibrium position, R, while the second term corresponds to the first derivative of the potential with respect to the position r_i, and the third term corresponds to the second derivative or Hessian matrix of the potential energy with respect to the positions r_i and r_j. The first term represents the minimum value of the potential, which can be set to zero. The second term equals zero at any local minimum of the potential. Up to second order, the potential is thus a sum of pairwise interactions (eq. 3.20).

$$\begin{aligned} U(r) =& \frac{1}{2}\sum_{i,j} \frac{\partial^2 U}{\partial r_i \partial r_j}\Bigg|_R (r_i - R_i)(r_j - R_j) \\ =& \frac{1}{2}\sum_{i,j}\Bigg|_R (r_i - R_i) H_{ij} (r_j - R_j) \\ =& \frac{1}{2}\Delta r^T H \Delta r \end{aligned} \tag{3.20}$$

where H is the $3N \times 3N$ real and symmetric Hessian matrix derived from the second derivatives of the potential with respect to the components of r or Δr. The Hessian matrix is the force-constant matrix with components in eq. 3.21

$$H_{ij} = \frac{\partial^2 U}{\partial r_i \partial r_j}\Bigg|_R \tag{3.21}$$

The H matrix is diagonalized by an orthogonal transformation, meaning the transformation allows it to rotate or reflect vectors without changing their length. However, to fully capture the normal modes, both kinetic and potential energy are considered, leading to a slight modification in the form of the matrix to be diagonalized. By treating the system as a group of classical particles, the equation of motion can be expressed as:

$$M\frac{d^2\Delta r}{dt^2} + H\Delta r = 0 \tag{3.22}$$

where M is the diagonal matrix consisting of massess of the particles. For a particle's three Cartesian coordinates, each mass is repeated thrice. Considering the $3N$-dimensional vector as a solution, $u_k(t) = a_k e^{-i\omega_k t}$, where a_k is the complex vector containing both amplitude and phase factor of k^{th} normal mode at time t, ω_k is the frequency of the mode of motion. By substituting this solution into eq. 3.22 the equation of motion takes the general eigenvalue equation form:

$$Hu_k = \omega_k^2 M u_k \tag{3.23}$$

Here, u_k represents the normal mode displacement for each normal mode k, where $k = 1, 2, ..., 3N$, and ω_k^2 is the corresponding eigenvalue (the squared frequency) for the k_{th} mode. The complete set of solutions can be organized into a matrix Q, where each of its column represents one of the normal mode vectors for all $3N$ modes.

$$HQ = MQ\wedge \tag{3.24}$$

where $\wedge$ is a diagonal matrix containing squared frequencies ω_k^2. Eigenvalues that are exactly zero correspond to conformational changes that do not influence the system's internal potential energy. Generally, the matrix H has six zero eigenvalues, which are associated with the rigid-body rotations and translations of the molecule.

Collective or global modes are governed by the protein's overall structure, making them largely insensitive to local interactions or specific force field parameters (Fig. 3.5, B). Tirion et al. [190] demonstrated that even a simplified harmonic force field produces global modes similar to those from detailed nonlinear models. These modes are primarily shaped by the inter-residue contact network, a geometric feature defined by the protein's native topology [81]. For faster calculation of normal modes in protein structures, elastic network models (ENMs) are commonly employed. ENMs are computationally efficient, easily scalable to different levels of coarse-graining, and require minimal parameterization [81]. These models approximate a protein's potential energy as a classical network of masses connected by springs, where each node represents a coarse-grained (CG) site and each edge represents a spring (Fig. 3.5, C). The network topology is defined by the native structure, with edges connecting nodes that fall within a specified cutoff distance.

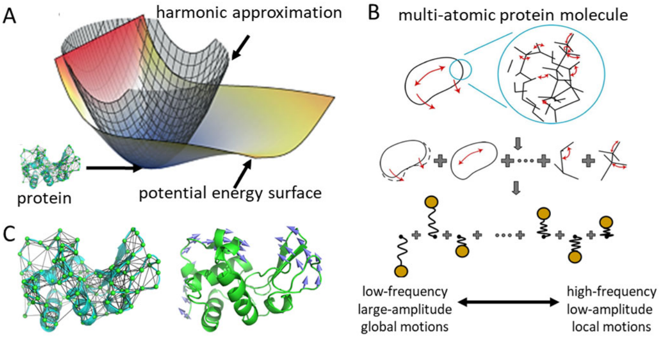


Figure 3.5: Normal mode analysis(A) Comparison of a PES with its harmonic approximation, illustrating a protein at its lowest energy state (adapted from Ref. [191]). (B) A multi-atomic protein with modes of motion (red arrows), categorized into low-frequency global motions and high-frequency local motions (adapted from Ref. [192]). (C) Elastic network model of hen egg-white lysozyme. The lysozyme structure (cyan ribbon) is overlaid with its elastic network, where C$_\alpha$ atoms (green spheres) are connected by black lines (left). The hinge-bending motion is shown by the lowest-frequency non-zero mode, with arrows indicating the direction and magnitude of movement (right) (adapted from Ref. [193]).

3.2.3.3 Computational parameters in the WEBnma tool

Numerous computational tools are available for performing NMA. However, for quick and efficient computation of normal modes, **Gradations of Protein Dynamics** utilizes the WEBnma server [194] to analyze the flexibility of AlphaFold2 structures and their corresponding NMR ensembles. WEBnma uses ENM focuses on a CG representation of proteins containing only C$_\alpha$ and uses a pre-defined elastic potential to simplify computations, making it efficient for large proteins. The ENM relies on the elastic potential

$$U_{ij} = k(\|R\|)(\|R_{ij}\| - \|R_{ij}^0\|)^2 \tag{3.25}$$

where

$$k(r) = \begin{cases} ar - b, & \text{for } r < d \\ cr^{-6}, & \text{for } r \geq d \end{cases} \tag{3.26}$$

Hinsen et al. [195] determined the force constant parameters as follows: $a = 8.6 \times 10^5\,\mathrm{kJ\ mol^{-1}nm^{-3}}$, $b = 2.39 \times 10^5\,\mathrm{kJ\ mol^{-1}nm^{2}}$, $c = 128\,\mathrm{kJ\ mol^{-1}nm^{4}}$, and d is 0.4 nm [195]. Here, R_{ij} is the distance vector between two C_α atoms, and R^0_{ij} is the corresponding distance in the reference configuration.

Since adjacent C_α atoms in proteins typically have distances close to 0.4 nm, these interactions are nearly uniform, while others scale with the inverse sixth power of the equilibrium distance [194]. The parameterization is generally transferable across proteins. Due to computational limitations, WEBnma produces only the first 200 non-trivial modes (excluding the first six zero-frequency modes corresponding to global translation and rotation).

3.2.3.4 Computing RMSF from Normal modes

As mentioned earlier, RMSF from normal modes and eigenfrequencies obtained from WEBnma are calculated on the AlphaFold2 structures and their corresponding NMR ensembles in **Gradations of Protein Dynamics** [194, 196]. The atomic fluctuations were estimated under thermal equilibrium by analyzing the normal modes with the lowest eigenvalues [189]. The squared fluctuation of atom i is given by:

$$\mathrm{fluc}_i^2 = k_\mathrm{B}T \sum_{k=7}^{M} \frac{v_{xi,k}^2 + v_{yi,k}^2 + v_{zi,k}^2}{m_i \omega_k^2} \tag{3.27}$$

Here, k_B is the Boltzmann constant, T is temperature (here 300 K), $M - 6$ the number of contributing eigenvectors, m_i is the mass of the i^th amino acid residue, ω_k^2 is the K^th eigenvalue, $v_{xi,k}$ is the Cartesian x-component for C_α atom i in the corresponding K^th normal mode vector, and similarly for $v_{yi,k}$ and $v_{zi,k}$. The normal mode vectors v_k are mass-weighted and normalized [189]. The sum excludes the 6 zero-frequency modes associated with global translations and rotations, and only the lowest 200 non-trivial normal mode vectors ($M = 206$) are included in the calculation. From these C_α atom fluctuations, the root-mean-square fluctuations $(\mathrm{RMSF})_i$ for each residue i, representing the average fluctuation or displacement of individual atoms from their mean positions, is computed by taking the square root:

$$\mathrm{RMSF}_i = \sqrt{\mathrm{fluc}_i^2} \tag{3.28}$$

The RMSF profile was constructed using the normalized eigenvectors and eigenvalues from WEBnma, with the formulas in Eqs. 3.27-3.28.

3.3 Simulating the AGP protein

3.3.1 Simulated systems of AGP

To study the effect of glycosylation and mutations on AGP dynamics, 18 systems of AGP were simulated using MD simulations with GROMACS software. These included AGP and its 8 mutants in both glycosylated and unglycosylated forms. The detailed methodology for AGP's sequence and 3−D structure, selection of glycosylation, and mutations is detailed in the Research Paper. The X-ray crystal structure of AGP (PDB code: 3kq0) was used as the template for all models. Mutants were generated by introducing single point mutations using CHARMM-GUI's PDB manipulator. Glycosylation was modeled by adding five selected glycan chains at specific sites using CHARMM-GUI's glycan modeler. Each system was solvated with TIP3P water and placed in a cubic periodic box with 10 Å padding, and neutralized with Na^+ and Cl^- ions to reach a 0.15 M concentration.

3.3.2 Computational details

For the 18 systems of AGP, the atomic interactions were modeled using the all-atom CHARMM36m force field for both the protein and glycans. Covalent bonds involving hydrogens were constrained using the LINCS algorithm. The Verlet cut-off scheme was employed with a 12 Å cutoff for non-bonded interactions, and LJ interactions were smoothly switched off between 10 and 12 Å using a force-switch method. Long-range electrostatic interactions were handled using the PME summation. To relax the structures, steepest-descent energy minimization was carried out for 5000 steps, with positional restraints applied to the protein backbone and sidechains at force constants of $400\,\mathrm{kJ/mol/nm^2}$ and $40\,\mathrm{kJ/mol/nm^2}$, respectively, during both energy minimization and NVT equilibration. Following this, an NVT equilibration was performed for 125 ps under constant particle number, volume, and temperature conditions. This was followed by NPT equilibration for each replica under constant pressure for 40 ns. The MD simulations used a 2 fs integration step, with the Nosé-Hoover scheme maintaining the temperature at 310 K using a 1 ps time coupling constant for both the solute and solvent. The Parrinello-Rahman scheme was applied to maintain pressure at 1 bar with a 5 ps time coupling constant. The NPT production run lasted 100 ns, with trajectory snapshots taken every 100 ps, resulting in 1000 frames for each system. To enhance sampling, three replica simulations were performed for each system, totaling 54 MD trajectories across 18 systems. Each "replica" refers to simulations of identical structures with identical parameters, except for initial velocities, which were randomly generated using the Maxwell distribution at 310 K in GROMACS. Each replica underwent 40 ns of NPT equilibration followed by a 100 ns production run, yielding a total of 300 ns of production time per system.

3.3.3 Convergence in glycoprotein MD simulations

MD simulations are a valuable tool for investigating glycoprotein dynamics. However, glycans are significantly more flexible than proteins and require longer timescales, typically on the order of μs, to adequately sample their conformational space. Compared to proteins, the conformational changes in glycans, such as sugar puckering, occur on the order of 100 ns, while the interconversion of primary and secondary OH groups takes place over 1 ns to 100 ps, respectively. In contrast, the rotation around the glycosidic linkages interconverts on the order of 10 ns. Also, transitions between their rotameric states can be relatively rare, often necessitating MD simulation times exceeding 100 ns or the use of enhanced sampling techniques to achieve convergence. Despite this, MD simulations of glycoproteins remain highly informative. Extended simulation times enable convergence and reliable comparisons between simulated and experimental properties. However, even converged results are limited by the accuracy of the underlying methods. Choosing optimal simulation conditions and the right force field is crucial. A simpler model can be as effective as a complex one, as long as it produces a converged result that can be validated. Exploring shorter timescales can often provide sufficient answers to many questions.

In this thesis, all the (glyco)proteins were simulated for 100 ns, and to increase sampling, three replica simulations were carried out per system resulting in a total of 54 MD trajectories for 18 systems. The term ‘replica’ refers to the simulations of identical structures with the same parameters, differing only in their initial velocities. Therefore, the next conclusion relates to the convergence and reliability of these results. As a rule of thumb, in statistical terms, error decreases as $1/\sqrt{N}$, where N is the number of independent samples, and to achieve reasonable precision in MD simulations with around 10% error, at least 10-100 independent samples are needed to adequately capture relevant conformational motions. Achieving exhaustive sampling is often challenging; however, multiple replicas enable the exploration of a broader range of conformational space, providing valuable insights into AGP dynamics and the context-dependent effects of glycosylation and mutations. In summary, while MD simulations can capture the complexity of PTM-induced protein dynamics, their computational cost and sampling challenges may limit convergence. Nevertheless, they remain essential for studying glycoprotein behaviour. It is strongly recommended to run additional replicas of the same system to better capture glycoprotein dynamics.

Chapter 4

Effect of glycosylation and mutation on conformational dynamics of AGP

Conformational dynamics of α-1 acid glycoprotein (AGP) in cancer: A comparative study of glycosylated and unglycosylated AGP.

Bhawna Dixit, Wim Vranken, An Ghysels

Proteins: Structure, Function, and Bioinformatics, (2024), 246-264, 92(2). https://doi.org/10.1002/prot.26607

4.1 Background and methodology

The key research question guiding this study is: **How do mutations and glycosylation, individually and in combination, affect the conformational dynamics and flexibility of AGP?** AGP was chosen as the system of interest due to its multi-functional role as a heavily glycosylated plasma protein in critical physiological processes, including immunomodulation and diverse ligand binding. Its function is significantly influenced by glycosylation patterns and mutations, which are especially relevant in the context of cancer. Understanding how these factors interact to affect AGP's conformational dynamics is crucial for comprehending its broader implications in protein-drug design.

As part of this study, the X-ray diffraction crystal structure of AGP (PDB code: 3kq0) was retrieved from the PDB. Since the 3-D structure of AGP is determined from its unglycosylated form, which lacks N-glycans, relevant glycans that bind to AGP were selected. The AGP structure was then modeled with N-glycans using CHARMM GUI. Following this, a large-scale screening of mutations potentially affecting AGP's backbone dynamics was performed using backbone dynamics propensities computed by DynaMine. The data related to glycans was obtained from the open-source GlyConnect database, while mutation data was sourced from the Catalogue of Somatic Mutations in Cancer (COSMIC), an open-source database of cancer mutations. The cancer mutations, including natural variants, were chosen based on their predicted impact on AGP's backbone dynamics via DynaMine and their proximity to glycosylation sites. The detailed methodology for mutation selection is outlined in the Supporting Information for AGP (Appendices). Next, the 3-D structure of AGP with the selected mutations was modeled, and the resulting mutants were subsequently modeled with the chosen glycans. This resulted in 8 mutants, with a total of 16 3-D models, each representing both glycosylated and unglycosylated forms. MD simulations were performed for AGP and its mutants (both glycosylated and unglycosylated), resulting in a total of 18 AGP systems. To enhance sampling, three replica MD simulations were conducted for each system, yielding a total of 54 MD trajectories across the 18 systems. The term "replica" refers to simulations of identical structures with the same simulation parameters, differing only in their initial velocities. The 54 MD trajectories of the 18 systems were analyzed using various MD observables including RMSD, RMSF, radius of gyration, solvent accessible surface area (SASA), PCA on coordinates, (6) contact analysis, H-bonds, and glycan torsion angles.

4.2 Contributions

My key contributions to this study include the development of a comprehensive Python pipeline for detailed analysis of AGP at both the protein and glycan levels, alongside MD simulations and modeling, and primary data collection using various bioinformatics tools using Python. The pipeline developed includes: 1) large-scale mutation screening using DynaMine, 2) calculation of various MD observables from the trajectories, 3) visualization of these observables, 4) PCA of the MD trajectories, and 5) computation of glycan torsion angles. Additionally, Python and Bash scripts were written to streamline the MD simulation workflow and model construction.

4.3 Concluding remarks

To address the key research question, this study highlighted the complexity of the interplay between glycosylation and amino acid mutations in AGP. Considering the individual effect of glycosylation, it decreases flexibility at the glycosylation site and simultaneously enhances the flexibility of distant regions of the protein. In contrast, mutations in the absence of glycans influence the local flexibility of the protein by inducing long-range conformational effects. The combined effects of mutations and glycosylation on AGP's behavior, or its 'molecular phenotype,' are complex and not yet fully understood. While glycosylated mutants display greater similarity in their backbone dynamics, no specific set of mutants shows consistent biophysical outcomes. However, one particular mutation emerges as structurally significant. Positioned near several glycosylation sites, this mutation plays a critical role in modulating glycan-protein interactions, thereby influencing the dynamics of both the glycans and the protein. The study concludes that mutations control glycan dynamics which modulates the protein's backbone flexibility directly affecting its solvent accessibility.

Conformational dynamics of α-1 acid glycoprotein (AGP) in cancer: A comparative study of glycosylated and unglycosylated AGP

Bhawna Dixit,[abc] *Wim Vranken,*[bc] *An Ghysels*[a,*]

a IBiTech–BioMMeda Group, Ghent University, Belgium

b Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Belgium

c Structural Biology Brussels, Vrije Universiteit Brussel, Belgium

* Corresponding author: an.ghysels@ugent.be

Abstract

α-1 Acid glycoprotein (AGP) is one of the most abundant plasma proteins. It fulfils two important functions: immunomodulation, and binding to various drugs and receptors. These different functions are closely associated and modulated via changes in glycosylation and cancer missense mutations. From a structural point of view, glycans alter the local biophysical properties of the protein leading to a diverse ligand-binding spectrum. However, glycans can typically not be observed in the resolved X-ray crystallography structure of AGP due to their high flexibility and microheterogeneity, so limiting our understanding of AGP's conformational dynamics 70 years after its discovery.

We here investigate how mutations and glycosylation interfere with AGP's conformational dynamics changing its biophysical behaviour, by using Molecular Dynamics (MD) simulations and sequence-based dynamics predictions. The MD trajectories show that glycosylation decreases the local backbone flexibility of AGP and increases the flexibility of distant regions through allosteric effects. We observe that mutations near the glycosylation site affect glycan's conformational preferences. Thus, we conclude that mutations control glycan dynamics which modulates the protein's backbone flexibility directly affecting its accessibility. These findings may assist in the drug design targeting AGP's glycosylation and mutations in cancer.

Keywords: α-1 Acid glycoprotein, Glycosylation, Mutations, Molecular dynamics, Conformational dynamics, Cancer

Introduction

α-1 Acid glycoprotein (AGP), also known as orosomucoid (ORM), is one of the most abundant plasma proteins.[1,2] It is mainly synthesized by the liver and secreted by hepatocytes.[3] It plays a crucial role in binding and transporting various basic and neutral drugs and acts as a biomarker of inflammation in diseases.[4] However, the understanding of its function remains hindered by its capacity to adopt multiple forms through extensive and heterogeneous glycosylation.[5] AGP exists in two primary forms that coexist in human individuals: AGP1 and AGP2.[6] AGP is encoded in three adjacent genes on chromosome 9: the AGP-A gene (ORM1 gene) expresses AGP1, while the AGP-B, AGP-B' genes (ORM2 genes) express AGP2. The ORM1 gene is the most active.[6] It is induced due to acute-phase reactions (APRs),[7] which occur in the human body as a response to injury, infection, inflammation, or cancer.[8,9] The ORM1 gene encodes the protein precursor of AGP1 with a total number of 201 residues, which encompasses a secretory N-terminal signal peptide of 18 residues and the 183 amino acids of mature AGP1. There are three allelic variants of the ORM1 gene, called the F1, F2, and S variants. The ORM2 gene encodes AGP2, which differs in about 20 substituted amino acids from AGP1. As AGP1 is more abundant in plasma than AGP2,[6] this paper will focus on the AGP1 protein and protein precursor, and we will refer to it as AGP hereafter. The sequence numbering includes the 18 signal peptide residues so that the mature AGP runs from amino acid 19 to 201.

Further, the structure of AGP undergoes N-glycosylation, a common and typically conserved post-translational modification. N-glycosylation is estimated to modify more than 50% of all eukaryotic proteins through the attachment of an oligosaccharide moiety (a glycan) to proteins.[10] N-glycans are mostly covalently attached to the side chain of asparagine residues in Asn-X-Thr/Ser (N-X-T/S) motifs, where X can be any amino acid.[11] N-glycosylation may significantly change the conformation and flexibility of a protein.[12] The glycans attached to proteins play a crucial role in molecular recognition and protein surface protection. They are known to behave as "flexible and bulky molecular glue".[13] Their intrinsically flexible behaviour and a lack of a single, well-defined structure render them difficult for structure determination techniques such as X-ray crystallography. AGP is a heavily glycosylated protein, with glycans constituting around 45% of its molecular weight.[14] The *in vitro* 3-dimensional structure of AGP (PDB code: 3kq0) is determined from an unglycosylated form, which does not contain any N-glycans. [7] AGP can *in vivo* have five N-linked glycans attached at N33, N56, N72, N93, and N103 with varying glycan composition and branching, such as e.g. the glycosylation pattern in Figure 1.[14]

Despite the differential glycan heterogeneity and glycoforms observed amongst different experimental studies, AGP shows a common preference for N-glycans with a specific number of antennae at specific sites. According to chromatography experiments of AGP,[15] the glycosylation sites at N33 and N56 showed a high occurrence of bi-antennary glycans, the N33 site strongly disfavored tetra-antennary glycans, and the N93 and N103 sites showed high occurrences of tri-/tetra-antennary glycans.[15,16] However, these glycosylation patterns can change in response to APRs.[9] Previous studies have observed changes in relative proportions of normal glycoforms to abnormal glycoforms along with high plasma concentration of AGP in cancer patients compared to healthy individuals[17]. In addition, the glycans may interfere with drug binding to AGP due to their high flexibility and shielding effect[18], consequently restricting access to binding sites. Besides glycan heterogeneity, the genetic variants of AGP-A which are F1-S variants (ORM1*F1, ORM1*F2, and ORM1*S) demonstrate different binding specificity for drugs[19]. A clinical study showed that ropivacaine administered with dipyridamole in patients with cardiovascular diseases showed a higher binding affinity for the F1 variant than the S variant.[20] However, the binding specificity of drugs amongst F1-S variants is investigated in limited studies as the routine experiments are generally performed on the pooled blood plasma samples with a mixture of AGP F1-S variants. Moreover, there is evidence that AGP is a high-affinity and low capacity drug-binding protein, and due to these properties along with its variable serum levels during APRs, the drug binding efficacy of AGP-binding drugs can be affected.[20–23] The binding specificity of AGP, as well as the binding efficacy of ligands to achieve sufficient therapeutic response targeted to AGP, depend on its binding site accessibility. The accessibility of the binding sites of AGP might be affected by conformational changes caused by glycosylation and/or amino acid mutations.

Within the context of cancer, AGP characteristics such as its variable serum levels and the dynamic changes in glycosylation patterns, along with the altered drug binding specificities exhibited by different AGP genetic variants during APRs in cancer, collectively imply the complexity of AGP's role in protein-drug interactions. AGP mutations may therefore impact AGP's response to (cancer) drugs by altering its conformational dynamics with and/or without glycans, while not directly affecting cancer initiation and progression. To date, there have been no studies linking the effect of aberrant glycosylation and amino acid mutations observed in cancer on AGP's conformational dynamics. Aberrant glycosylation refers to the fundamental changes in the glycosylation patterns of cell surface and secreted glycoproteins that occur during cancer progression.[24] Therefore, it is imperative to investigate the complex conformational dynamics of AGP and its mutants with and without glycans to understand how this might interfere with AGP's biophysical behaviour and function.

In this study, we used Molecular Dynamics (MD) simulations and b2bTools sequence backbone dynamics predictions to investigate the conformational dynamics of AGP. We used the X-ray crystal structure of un-glycosylated AGP (PDB code: 3kq0) to model its glycosylated form, referred to as gly-AGP (Figure 2(a)). To elucidate the effect of oncogenic point mutations on AGP's conformational dynamics, we modelled and simulated the atomic effect of 8 cancer missense mutations from the COSMIC database on the X-ray crystal structure of AGP. To investigate the effect of mutations together with glycans, we in addition modelled the glycosylated mutants (gly-mutants) with glycan chains identical to those of gly-AGP. Finally, we addressed the effect of mutations and glycosylation on AGP's flexibility and solvent accessibility (Figure 2 (b-c)) and the effect of mutations on glycan dynamics. To characterize which regions, undergo conformational changes in AGP due to mutations and glycosylation and to what extent their accessibility is affected, we systematically defined three distinct structural regions of AGP (Figure 2 (c)). These regions were selected based on the lipocalin fold of the protein, consisting of an open-end ligand binding-site entrance (LPE), a central ligand binding site (LBS), and a hypothetical protein-protein interaction site (hPPI) at the closed end of the protein.[25]

We found that mutations perturb the glycan dynamics differently in each system, which in turn variably affect the protein's backbone flexibility and solvent accessible surface area (SASA). Overall, the effect of glycosylation on the protein and glycans is complex and depends upon the structural location and physico-chemical behaviour of mutations.

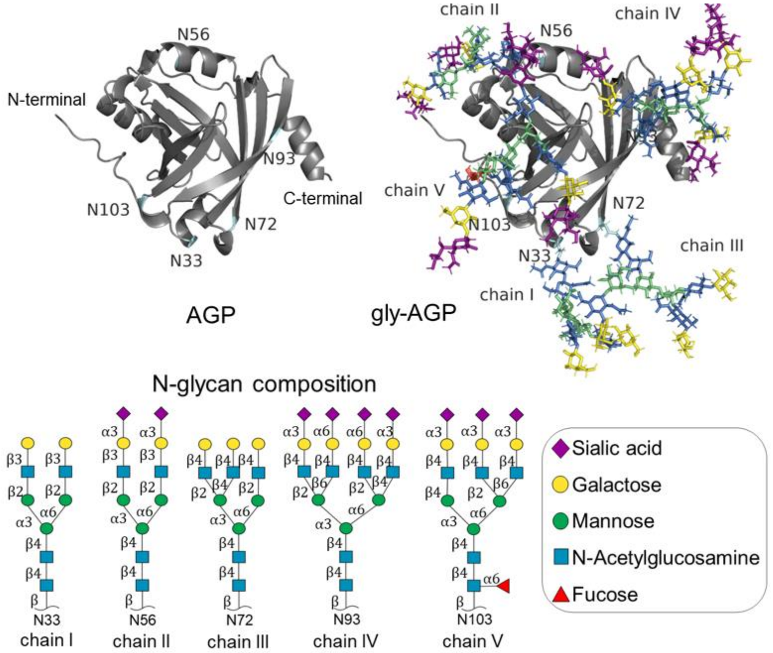


Figure 1. AGP, ribbon representation showing the eight-stranded β-barrel linked by four loops and flanked by an α-helix at the C-terminal. Gly-AGP showing N-glycans, covalently attached at N33, N56, N72, N93, and N103. Heterogeneous glycosylation patterns used in the MD simulations, using symbol nomenclature for glycans (SNFG) and coloring.[26,27]

Methodology

Sequence analysis and selection of the mutations

The AGP expressing ORM1 gene is highly polymorphic due to which there are three genetic variants of ORM1 observed worldwide: ORM1*F1, ORM1*F2 (two subtypes of ORM1*F) and ORM1*S [6]. The three genetic variants differ by a change in a genetic codon either in exon 1 or 5 resulting in three amino acid substitutions (mutations) of AGP. Specifically, the ORM1*F1 codes for AGP sequence with Q38 and V174, ORM1*F2 codes for AGP with Q38 and M174, and ORM1*S codes for AGP with R38 and V174.

In this study, the ORM1*F1 genetic variant of AGP (UniProt accession P02763)[28,29] was chosen as wild-type AGP (referred to as AGP hereafter), since it is more common worldwide than the other two genetic variants [6]. A crystal structure of this variant is available in the PDB database (code 3kq0). The full AGP sequence of 201 residues, including the 18 signal peptide residues of the protein precursor, was analysed with b2bTools for the prediction of backbone dynamics propensity,[30] which served as a reference. The Catalogue of Somatic Mutations in Cancer (COSMIC), was consulted to collect somatic mutations in AGP, mainly originating from carcinoma of the stomach, breast, lung, liver, kidney, ovary, lymphoid and other tissues.[31] Out of 240 somatic mutations observed in 480 samples in the ORM1 gene reported by COSMIC (nonsense, missense, and synonymous), there were 150 samples containing 61 unique missense mutations. Missense mutations change the amino acid residue at the protein level, so often altering the protein structure and function.[32] Moreover, in

COSMIC, out of 150 samples, R38Q (ORM1*S → ORM1*F1) was observed in 65 samples. Since the reference AGP was ORM1*F1, we excluded the R38Q mutation, and included Q38R (ORM1*F1 → ORM1*S) in the list. In addition, V174M (ORM1*F1 → ORM1*F2) was also observed in the missense mutations. Thus, we selected these two special cases of genetic variants for protein modelling.

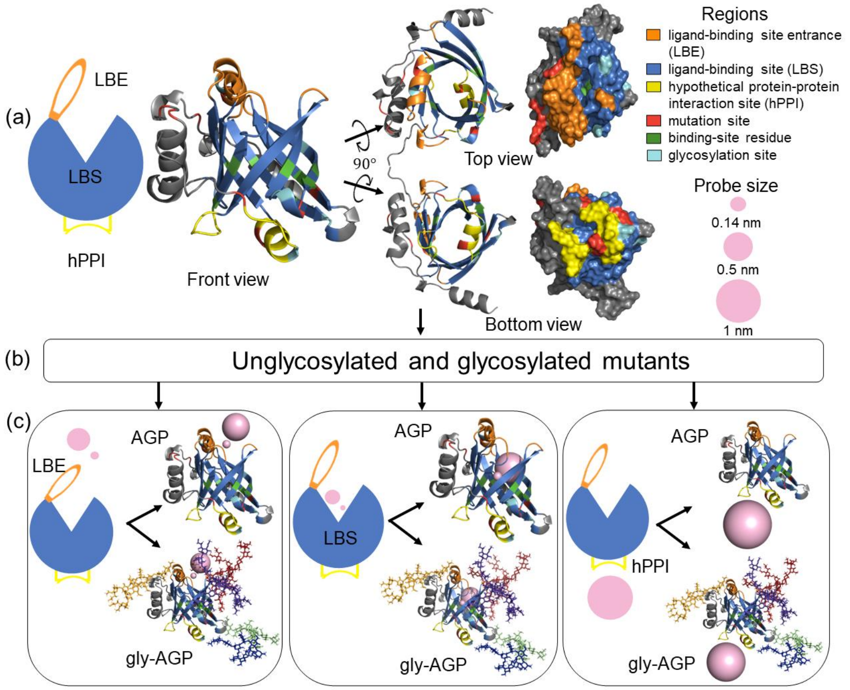


Figure 2. (a) Cartoon of AGP with 3 structural regions relevant to binding: ligand binding-site entrance (LBE, orange), ligand binding-site (LBS, blue), and the hypothetical protein-protein interaction site (hPPI, yellow). Rotated ribbon representations show different viewpoints. The ribbon representation demonstrates LBE composed of a helix flanked with loops and three loops, LBS composed of β-barrel, and hPPI composed of a loop and a helix flanked by loops.[25] (b) AGP and 8 mutants are modelled in MD simulations, without and with the glycosylation pattern of Figure 1. (c) Accessibility of the 3 regions is measured by computing the average SASA in the MD trajectories with a probe (pink). Two small probes (radius 0.14 nm, 0.5 nm) are used to screen LBE and LBS regions, relevant for smaller ligands. A larger probe (radius 1 nm) is used for hPPI.

For the remaining total 59 missense mutations, the backbone dynamics propensity profile was also predicted with b2bTools (Figure S1), and the root-mean-square-error (RMSE) of the backbone dynamics propensity profile was computed (between AGP and each mutant). The mutants were ranked from high to low RMSE (Figure S2), where a high RMSE value indicated that the backbone dynamics propensity of a mutant deviates the most from AGP. The sequence-based analysis with b2bTools thus allowed identifying mutants that were most likely to alter AGP dynamics.[30] Another criterion in our selection of mutants was the structural proximity of the point mutation to the glycosylation sites, as it seemed plausible that close proximity was more likely to interfere with the effect of the glycosylation. Out of the total of 61 mutations, 8 different sequences were finally selected: 6 missense mutations (P28L, Q60L, I78N, R101W, R167C, and P169L) and two genetic variants (Q38R, and V174M).

Selection of glycosylation pattern

We selected glycans from the recent glycoproteomics data curated from mass spectrometry experiments deposited in the GlyConnect database, which is an integrated set of databases containing manually curated information about site-specific glycosylation patterns.[33] For AGP (UniProt accession: P02763), there were a total of 49 N-glycan structures deposited for the five glycosylation sites where each glycan chain was annotated with site-specific composition and branching information, as well as the disease-specificity. To narrow down site-specific glycans, we investigated the site-specific glycan composition and branching preferences of five N-glycosylation sites of AGP based on a previous study as there were multiple structures reported for a single N-glycosylation site. We followed the observations in the experimental work by Higai et al., who reported a preference for bi-antennary glycans at N33 and N56, tri-antennary glycans at N72 and N103, and a sialylated tetra-antennary glycan at N93.[34] Then, out of 49 structures, we filtered the data based on several criteria: 1) complex type composition, 2) reported in disease-causing states such as inflammation or cancer, and 3) a 3D structure available in the GlyTouCan database. In any case, if a 3D structure or glycosidic bond information was missing, we performed a similarity search in GlyTouCan.

The selected complex and heterogeneous N-glycans (Figure 1) were observed in carcinoma, autoimmune disorders, and SARS-CoV-2.[35–40] The accession codes of the GlyTouCan glycan repository for each site-specific glycan are reported, which are linked to the GlyConnect databases. The glycans were mainly composed of β-N-Acetyl-D-glucosamine (β-GlcNAc), β-D-Galactose (β-Gal), α- and β-D-Mannose (α-Man, β-Man), terminal α-N-Acetylneuraminic acid (α-Neu5Ac), and α-L-Fucose (α-Fuc).

First, the N33 glycan (G22140GZ)[36] was observed in pathophysiological conditions such as hepatocellular carcinoma,[41] autoimmune disorders, systemic lupus erythematosus,[37] and SARS-CoV-2.[39] The missing linkage information was taken from structure G64633BH for the β-GlcNac(1→2)α-Man and β-Gal(1→3)β-GlcNac linkages based on glycan composition similarity search on GlyTouCan.[42,43] Second, the N56 glycan (G77252PU) structure was observed in squamous cell carcinoma, colon adenocarcinoma, and acute leukaemia.[44] It contained terminal sialic acids e.g., α-Neu5Ac) on the bi-antennary glycan which was also a characteristic of metastatic cancer cells.[45] Third, the N72 glycan (G36131WL) was observed in hepatocellular carcinoma, and various acute-phase proteins.[36,41,46–48] Fourth, the N93 glycan (G62165AG) was observed in APRs.[38] The missing linkage information for N72 and N93 were added based on a glycan composition similarity search (G85046IS and G55645TD respectively). Fifth, the N103 glycan (G01752VH) was similar to the N72 glycan, except that it was core fucosylated. Core fucosylation is a potential biomarker of cancer.[38,49] Thus, the selected glycan chains represent potential site-specific glycosylation patterns of AGP based on consensus experimental data, however, in-vivo, it may be different in complexity and branching as they are tissue-specific, disease-specific and vary from one individual to another.

Molecular dynamics (MD) simulations

System setup and model building

The X-ray diffraction crystal structure of AGP, variant ORM1*F1[6], was obtained from the Protein Data Bank (PDB code: 3kq0).[7,50] In this AGP structure, the coordinates of the first 18 signal peptide residues, the 19th residue (the starting residue of the mature AGP form), and the last 8 residues located in the disordered C-terminal region were missing. The modelled structure is therefore 174 residues long (residue 20 to residue 193). The PDB coordinates of AGP were used as the template for the construction of all modeled systems. To construct each of the 8 mutants, a single point mutation was introduced in the AGP template with the PDB manipulator of the CHARMM-GUI software.[51] For glycosylated systems, five glycans chains (Figure 1**Error! Reference source not found.**) were added to glycosylation sites with the glycan modeler from CHARMM-GUI. The glycan modeler utilizes average glycosidic torsion angles gathered from the Glycan Fragment Database for optimizing glycan structures.[52] Each protein was solvated with explicit TIP3P water molecules and was put in a cubic periodic box that extended 10 Å beyond the (un)glycosylated protein. As AGP was negatively charged, the systems were neutralized with counterions, placed by a Monte-Carlo method in CHARMM-GUI, reaching a 0.15 M concentration of Na^+ and Cl^- ions. An overview of the atom numbers is given in the supplementary information (Table S 1).

Computational details

The atomic interactions were modeled with the all-atom CHARMM36m force field for both protein and glycans. .[53,54] All covalent bonds with hydrogens were constrained with the LINCS algorithm.[55] The Verlet cut-off scheme was used with a cutoff of 12 Å for the non-bonded interactions. {Citation}The Lennard-Jones interactions were smoothly switched off using a force-switch between 10 to 12 Å. [56] Particle-mesh Ewald summation was applied for long-range electrostatic interactions.[57] The MD simulations were run with the GROMACS software.[58] To equilibrate the structures, steepest-descent energy minimization was performed for 5000 steps. Positional restraints were applied to the protein backbone and sidechains with a force constant of 400 kJ/mol/nm^2 and 40 kJ/mol/nm^2 respectively. Next, an NVT equilibration with a constant number of particles, volume, and the temperature was performed for 125 ps for all systems. In the next step, the NPT equilibration was carried out for each replica with constant pressure for 40 ns. The integration step was 2 fs in the MD. The Nosé-Hoover scheme was applied to maintain the temperature at 310 K with a 1 ps time coupling constant for both solute and solvent. The Parrinello-Rahman scheme was applied to maintain the pressure at 1 bar with a characteristic time coupling constant of 5 ps. The NPT production run was 100 ns. Trajectory snapshots were collected every 100 ps, giving 1000 frames for each system. To increase sampling, three replica simulations were carried out per system resulting in a total of 54 MD trajectories for 18 systems. The term "replica" refers to the simulations of identical structures consisting of identical parameters except their initial velocities that are randomly generated using Maxwell distribution at 310 K using GROMACS.[59] Each replica was then subjected to 40 ns NPT equilibration followed by a 100 ns production run yielding a total of 300 ns production run per system.

Analysis of molecular dynamics trajectories

The MD trajectories of the 18 systems were analysed with in-house Python scripts making use of the MDAnalysis and MDTtraj packages.[60–62] Before analysis, trajectories were aligned with their equilibrated initial structure of the production run to remove the effect of meaningless net translations or rotations. The following list of observables was computed: 1) the root-mean-square deviation (RMSD), (2) the radius of gyration (R_g), (3) root-mean-square fluctuations (RMSF), (4) solvent accessible surface area (SASA), (5) principal component analysis (PCA) on coordinates, (6) contact analysis, (7) H-bonds, and (8) glycan torsion angles. For some properties, the result was a profile as a function of the residue number (or atoms), e.g., for the RMSF or the SASA. To focus on the 3 regions relevant for binding (LBE, LBS, hPPI), contiguous amino acids in these regions were grouped into fragments: 4 fragments for LBE, 8 fragments for LBS, and 2 fragments for hPPI. The per-residue property was then summed over the residues (or atoms) in a fragment, giving 14 fragment-based values. The fragment definitions are provided in Table 1.

Table 1. Three regions of AGP relevant for binding: ligand binding site entrance (LBE), ligand binding site (LBS), and hypothetical protein-protein interaction site (hPPI). Regions are divided into fragments of contiguous amino acid residues. Identified binding site residues are shown as reported by Schönfeld et al.[7] The mutations that occur in the following sequence fragments are mentioned, other mutations R167C, P169L, and V174M occur in the structural vicinity of these regions and not directly in the sequence of the fragments.

Fragment	Residues	Identified binding site residues	Secondary structure	Glycosylation site	Mutation site
LBE					
1	R51-E61		helix	N56	Q60L
2	Q87-D88		loop		
3	G111-G112		loop		
4	N135-W140		helix		
LBS					
1	G41-F50	Y45	β-sheet		
2	Q63-K73	F67, F69	β-sheet	N72	
3	D76-R86	L80	β-sheet		I78N
4	Q89-R101	L97, V99	β-sheet	N93	R101W
5	E102-V110	I106	β-sheet	N103	

6	Q113-I121	Y127	β-sheet		
7	K126-V134	L130	β-sheet		
8	W140-P149		β-sheet		
hPPI					
1	P28-T40	I30, L36	helix	N33	P28L, Q38R
2	L122-K126		loop		

RMSD

To verify the convergence of each simulation, the root-mean-square-deviation (RMSD) of the coordinates of each protein's C_α atoms was computed as a function of time, where the equilibrated initial structure of the production run was used as the reference. To verify structural stability, the X-ray crystal structure of AGP was used as a reference to calculate C_α RMSD of each system. For each system, the RMSD calculations were carried out over 100 ns per replica.

R_g

To measure the compactness of a structure, the radius of gyration (R_g) of each protein was computed based on the protein's C_α atoms for every snapshot. . R_g was calculated at over 100 ns per replica for each system. The average value for each replica is computed with its standard deviation, as well as the average over the three replicas for each system.

RMSF

The RMSF of each heavy atom was computed as the root of the variance of its coordinates over the trajectory over 100 ns for each replica per system. The RMSF calculation gives a profile as a function of the amino acid residue index. The average profile over the three replicas is also computed. In addition to comparing the full profiles, we focused on the regions relevant for binding. To facilitate the comparative analysis of the 18 modeled systems, the RMSF was summed over the residues in the fragments (Table S3),

$$RMSF_{frag} = \sum_{i=1}^{n} RMSF_i \tag{1}$$

where $RMSF_i$ is the RMSF of the i^{th} C_α and n is the number of residues in the fragment. To obtain the RMSF of a complete region, the sum was taken over the fragments *j* in the region:

$$RMSF_{region} = \sum_j RMSF_{frag\ j} \tag{2}$$

SASA

The solvent accessible surface area (SASA) was calculated using the Shrake and Rupley method with the Golden section spiral algorithm[63] as implemented in MDTraj. The algorithm puts a sphere on every atom and computes the area of each sphere that can be touched by the probe. The result for a given coordinate frame is an accessible area for each atom. This quantity fluctuates over time, and thus we reported the average over the 3000 snapshots (300 ns) with its standard deviation.

Mehdipour and Hummer probed the accessibility of ACE2 with and without glycans with different probe sizes.[64] In the present work, three probe sizes were used to represent binding candidates of different sizes (Figure 2(c)). The smallest probe with a radius of 0.14 nm resembles the size of a water molecule. The probe with a radius of 0.5 nm represents a small drug molecule. They were used to compute the SASA of the LBE, LBS, hPPI of all systems. The SASA of the five glycan chains was also computed in glycosylated systems. The largest probe of 1 nm radius represents a binding candidate the size of a protein domain and was used to scan the accessibility of the hPPI. Similarly, to the RMSF, the SASA values of individual atoms were summed over all atoms in the residues belonging to the fragments defined in **Error! Reference source not found.**:

$$SASA_{frag} = \sum_{i=1}^{n} SASA_i \tag{3}$$

where $SASA_i$ is the SASA of the residues and *n* is the number of residues in the fragment. To obtain the SASA of a complete region, the sum was taken over the fragments *j* in the region:

$$SASA_{region} = \sum_j SASA_{frag\ j} \tag{4}$$

As for the glycan chains, the SASA of each glycan in a glycan chain is computed by summing over the relevant atoms.

PCA on coordinates

To identify the principal modes of a protein's movements, principal component analysis was applied to the C_α coordinates of the protein. For each system, a total of 3000 snapshots from their corresponding replicas were first aligned with the X-ray crystal structure of AGP based on the C_α atoms before the covariance matrix was diagonalized. The cartesian coordinates of C_α atoms were projected on the first three principal components (PCs) with highest eigenvalues.

The eigenvectors or PCs corresponding to the highest eigenvalues span a subspace of essential large-amplitude conformational changes. Thus, the subspace of the first three eigenvectors was considered based on the 40-60% variation in the first three PCs. To compare the similar motions between the systems,[65] the root-mean-square-inner-product (RMSIP) between their essential subspaces was calculated,[66] which was calculated from the inner products of the eigenvectors:

$$RMSIP(A,B)= \left(\frac{1}{A}\sum_{a=1}^{A}\sum_{b=1}^{B}(u_a \cdot v_b)^2\right)^{\frac{1}{2}} \tag{5}$$

where A and B are the first 3 PCs, and u_a represents the a^{th} vector in the first subspace and v_b represents the b^{th} vector in the second subspace. The RMSIP score is between 0 to 1, with 0 being no overlap and 1 being maximum overlap.

Contact analysis

The native contacts with X-ray crystal structure of AGP as reference were computed, as well as overall contacts during the simulations (referred to as non-native contact). To measure the number of native contacts, the hard cut method from MDAnalysis with a default 0.45 nm cut-off (minimum distance between two atoms at least as close as reference) to distinguish between native contact and non-native contact was used. To measure the overall number of contacts, the default 0.45 nm cut-off radius (distance between atoms less than 0.45 nm) was used to distinguish contact and non-contact.[67,68] The contacts were computed based on the protein's C_α atoms for every snapshot (3000, from the 3 replicates) for each system.

H-bonds

We calculated the H-bonds using the GROMACS hbond analysis tool over 100 ns for each replica per system.[58]

Glycan torsion angles

The torsion angles of the glycan chains were calculated with MDAnalysis in the 3000 snapshots (1000 snapshots per replica) and their distribution was visualized as carb-Ramachandran plots. The conventional definitions (section 11 in Supp. info.) were used for glycosidic angles. Circular standard deviation (csd) was calculated for all the torsion angles over 300 ns.

Results

The three structural regions LBE, LBS, and hPPI are expected to play an essential role in protein/ligand-protein interactions. To unravel the effect of mutations and glycosylation on these regions, several aspects were investigated. First, the structural conformations and backbone flexibility of the amino acid chains were studied with standard methods, such as RMSF and PCA. Next, the glycan conformations were studied for the series of mutants. Finally, it was studied how glycans might restrict access of small molecule ligands/receptors to the binding regions, by measuring the solvent accessibility surface area in the presence and absence of glycans with varying probe sizes of 0.14 nm, 0.5 nm, and 1 nm.

3.1. Backbone flexibility of AGP and its mutants

To verify the stability of the structures, the C_α RMSD was computed in relation to their initial equilibrated structure (Figure S6, Figure S7). The mean RMSD over the 100 ns trajectory for each replica is reported in Table S2 and overall ranges from 0.13 to 0.33 nm. As all these values are above 0.10 nm, it can be concluded that AGP and its mutants are dynamic, and all systems exhibit an extended range of small conformational changes, which is not surprising given that AGP contains seven flexible loops and two flexible termini. The N-terminal is known to be a disordered region with a lack of crystallographic data, for which we found that Alphafold2 also fails to predict the structure, giving low pLDDT confidence scores.[69,70] The RMSD can vary amongst the three replicas (Table S2), which have different initial equilibrated structures (Figure S6, Figure S7). Most structures showed a gradual increase in RMSD over time (Figure S6, Figure S7), which could be an indication of non-equilibrated MD simulations. However, a more likely cause is the known flexibility of AGP and its mutants, as confirmed by the RMSF discussion (see below). Also, it is expected that replicas of the same systems initialized with random velocities produce different trajectories due to the rugged energy landscape associated with the protein. Even replicas initialized with identical velocities can produce differing trajectories due to the differences in floating-point precision on different machines[59]. On the other hand, the replicas enable a larger sampling of possible conformations of a system.   A study found significant divergence in the structural properties of 9 replicas initiated from 3 different models of P-glycoprotein on a 200 ns timescale[71].

Subsequently, the C_α RMSD to the X-ray crystal structure of AGP was computed (Figure S4, Figure S5). Taking the same reference structure for all systems allows us to gauge how much mutants and glycosylation variants structurally diverge. The average value of mean RMSD to the X-ray crystal structure of all systems for all replicas was approximately 0.24 nm. AGP itself showed a mean RMSD value of 0.24, 0.20, and 0.22 nm for three replicas, which lies well above 0.10 nm, and which means that the MD trajectory at 310 K takes a set of different conformations than its reported crystallographic structure at 100 K. This is in line with the expectations, again given the flexibility of AGP in its loops and termini. The mean RMSD of the other 17 systems, which have point mutations and/or glycans, ranges from 0.19 to 0.35 nm (Table S2). This indicates that mutations and glycans alter the extent to which conformations are sampled with respect to the reference PDB structure. However, this alteration is not drastic. Thus, the analysis with the RMSD indicates that the lipocalin fold of the PDB is retained in all systems, despite high flexibility of the protein.

The compactness of the structure can be assessed with the radius of gyration R_g based on the C_α atoms. The average R_g varies between 1.55 and 1.60 nm among the systems (Table S2). The average R_g (over 300 ns) varies between 1.55 and 1.60 nm among the 18 systems (Table S2). The instantaneous R_g values of each individual system are spread over wider ranges (Figure S8), which is another indication of the conformational flexibility over time of AGP and its mutants. Concerning the mutations, R167C shows lowest average R_g compared to all the other systems.

The average R_g was further affected by glycosylation as most glycosylated systems, except gly-R167C and gly-P28L, show average R_g over 1.57 nm (Figure S8, Table S2). In Figure S8, the inter-replica variation of R_g is much higher for each glycosylated systems than unglycosylated systems, which might be due to 1) high mobility of glycans and their ability to form distinct, rapid transient interactions, and 2) high flexibility of termini and loops. Comparing all unglycosylated and glycosylated systems (Table S2), the average R_g does not present a systematic difference, indicating that glycosylation does not cause a consistent increase nor decrease in protein compactness. This means that the glycosylation does not have a systematic effect on the compactness of the structure. Our results therefore show a complex picture, with the effect of mutations and glycosylation on the compactness of AGP being only minor and highly dependent on the type of mutation and its structural position.

The compact systems with lower R_g presumably have a higher number of intramolecular contacts leading to structural stability. The initial equilibrated structures have 544 to 554 C_α contacts within a 0.45 nm cut-off radius (Table S4, Figure S18). All systems maintained an average of approximately 98% of their initial contacts (~ 538 ± 2) over 300 ns (Table S4). This implies that the simulated structures are all structurally stable, despite the flexibility indicated by the RMSD and R_g. The number of contacts was also computed for the original PDB structure with a 0.45 nm hard cut-off distance, which resulted in so-called native contacts. The fraction of native contacts was determined as a function of time, and on average, all the systems and their corresponding replicas show a fraction of about 57% (~ 320 ± 11) native C_α contacts (Figure S17). This indicates that many of the contacts of the PDB structure, obtained from the crystallised protein at low temperature, are not retained at higher temperature, despite the lipocalin fold itself being retained.

To investigate which regions of AGP contribute the most to the conformational behaviour, the root-mean-square-fluctuations (RMSF) of the C_α positions were computed (averaged over the three replicates, Figure S10). As expected, loops showed the highest RMSF peaks corresponding to flexible regions, and β-sheets showed relatively low RMSF corresponding to rigid C_α backbone. Helices showed context-dependent behaviour based on surrounding residues and were more flexible than β-sheets, but less flexible than loops. The N- and C-termini showed very high RMSF indicating highly flexible behaviour compared to the rest of the protein. The C-terminus (residue 183 to 193) was tightly bound to the side of the β-barrel via a disulfide bond between C90 and C183 and via two H-bonds in AGP's X-ray structure at 100 K.[7] The first H-bond forms between the sidechains of Y68 and E187, and the second H-bond between Y83 and H190 (Figure S22). The MD simulations at 310 K revealed that the two H-bonds only form temporarily, persisting for picoseconds or nanoseconds, with intermittent breaking and reformation (Figure S23). This behaviour of forming and breaking H-bonds occurred in all mutated and glycosylated systems, except for gly-I78N and gly-Q60L (Figure S23). The results were in accordance with the RMSF curves of individual replicas (Figure S9) going up drastically for residues 183 to 193 for all systems. It can therefore be concluded that glycosylation and/or point mutations typically affect the mobility of the C-terminal even when none of the glycosylation or mutation sites is in the C-terminal.

At glycosylation sites, the mutants most often decreased in RMSF (Figure S10), but at site II, IV, and V, some mutants could have an increase in RMSF. Meanwhile some of the mutants occasionally maintained the same RMSF at some other sites. Some examples of an increase in RMSF include AGP at glycosylation site V, and R167C at site II. In contrast, glycosylation caused an increase in the RMSF in the region of residues 120-185. Interestingly, this sequence is sequentially and structurally distant from the glycosylation sites. We can conclude that glycosylation can decrease AGP's backbone flexibility locally, as well as altering the conformational flexibility of a structurally distant region. The latter allosteric effect might be due to reordering of dynamics to compensate for the potential entropy loss due to post-translational modifications.[72] The rapid interactions between protein amino acids and glycans do, in any case, have significant effects on the backbone's flexibility.

To relate the backbone flexibility to the binding process, a comparative RMSF analysis averaged over three replicas per system was done amongst all systems specifically in the LBE, LBS, and hPPI regions. The change in average RMSF in each fragment is computed as

$$\Delta RMSF_{frag} = RMSF_{frag}(g) - RMSF_{frag}(u) \tag{6}$$

where g denotes the glycosylated system and u denotes the unglycosylated system. These summed RMSF values cannot be compared among fragments as their value will depend on the number of residues in a fragment, which is variable. The fragment-based change in RMSF does give information about the change in flexibility upon glycosylation. In Figure 3, a downward arrow implies that glycosylation decreases the flexibility of the fragment, and an upward arrow implies that it increases the flexibility.

For fragment R51-E61, and Q87-D88 in LBE, R167C shows the highest increase in flexibility post-glycosylation, while for fragment G111-G112, R101Wshows the highest decrease post-glycosylation. In N135-W140 of LBE, P169L and I78N, show the highest increase in flexibility post-glycosylation. Thus, in LBE all the fragments except Q87-D88 show high variability in flexibility due to mutations, and within replicas of identical systems. In LBS, 6 out of 8 fragments except G41-F40, and W140-P148 show high variability in $RMSF_{frag}$ within mutants as well as within replicas of identical systems. The reason behind this is that these fragments are linked via highly mobile loops. In hPPI, the high variability in RMSF can be observed amongst all systems of AGP for the fragment P28-T40, while L122-K126 shows similar increase or decrease in RMSF amongst most systems. This means that glycosylation does not cause a general trend in the fragments' flexibility for the wild-type form as well as mutants. There are also mutants that stand out in how they are affected by glycosylation. However, to answer whether their outlier behavior in $\Delta RMSF_{frag}$ is an artefact of the replica simulations or a result of variable protein-glycan interactions, change in glycan dynamics due to mutations, we discuss $\Delta RMSF_{frag}$ of specific fragments and mutants with AGP as a reference.

The results show a high variability in $\Delta RMSF_{frag}$ amongst all mutants and their corresponding replicas, such as in LBE all the fragments (Figure 3). For LBE (Figure 3 (A)), in region R51-E61, R167C shows an outlier with very high $RMSF_{frag}$, implying a significant increase in average flexibility post-glycosylation. In AGP, the glycosylated replicas showed similar $RMSF_{frag}$ while glycosylated replicas showed variable behaviour. In replica 1 of gly-R167C, we observed intermittent H-bonds between the fragment Q87-D88 and R51-E61disrupted due to Neuraminic acid from glycan chain IV around 60 ns, resulting in unfolding of the helix turn at E61 and high flexibility in both fragments (Figure S13). In G111-G112 (Figure 3 (B)), an outlier in gly-AGP occurred due to the disrupted interactions between Q89-R101 and E102-V110 of LBS via β-Gal9 of glycan chain IV. The disruption caused the unfolding of the β-hairpin loop indicating high mobility, while the fold remained intact in the remaining two replicas with lower mobilities in the MD trajectory (Figure S14). Another example of an outlier is gly-I78N which shows very high flexibility of N135-W140 in one of the replicas (Figure 3 (C)). In gly-I78N, the N135-W140 loop showed two specific structural geometries in all replicas. Due to conformational changes caused by glycans, replica 1 shows a local transition between two geometries, leading to high RMSF of the loop in replica 1 (Figure S15). In E102-V110, a β-sheet linked by a two-residue mobile loop G111-G112 of LBE at one end, AGP and R101W show very high inter-replica variation in $RMSF_{frag}$ (Figure 3 (D)). Interestingly, AGP also shows an upward arrow showing an increase in average flexibility due to glycosylation, while all the other mutants show a downward arrow indicating a decrease in flexibility. Further, in R101W, the unglycosylated replicas showed high variability than glycosylated replicas. Again, in AGP residue R101 interacted with the sidechains of N33 at glycosylation site I, and D76 via intermittent H-bonds lasting several nanoseconds, while W101 in the replica R101W the H-bonds formed for several picoseconds with either N33 or D76 leading to high mobility of Q89-R101W and E102-V110 (Figure S16). The mobility of the loop was reduced in glycosylated R101W as N33 was glycosylated with glycan chain I, and W101 interacted with glycan chain IV.

Therefore, this is another illustration of the diversity in the effect of glycosylation. Even this diversity is not universal among all studied mutants. In summary, comparing all mutants, R101W consistently showed the highest variability in flexibility due to glycosylation, while V174M showed the lowest variability. Thus, in the context of the backbone flexibility of the protein, the effects of glycosylation are complex. On a sequence level, mutations were observed to increase the flexibility of five residues left and right to the mutation site for sequentially distant mutants from glycosylation sites (R167C, P169L, and V174M) (FigureS11). Interestingly, in these cases glycosylation further increased the local effect of mutations (Figure S12). For sequentially close mutations, it was not clear whether glycosylation increased or decreased the local flexibility of mutants. Therefore, the cumulative effect of mutations and glycosylation on the protein's backbone flexibility is ambiguous and highly context dependent. Additionally, the effects of glycosylation may be due to the interactions between protein-glycan, glycan-glycan, and interchain glycan interactions, and mutations may differently perturb these local interactions between glycans and proteins.

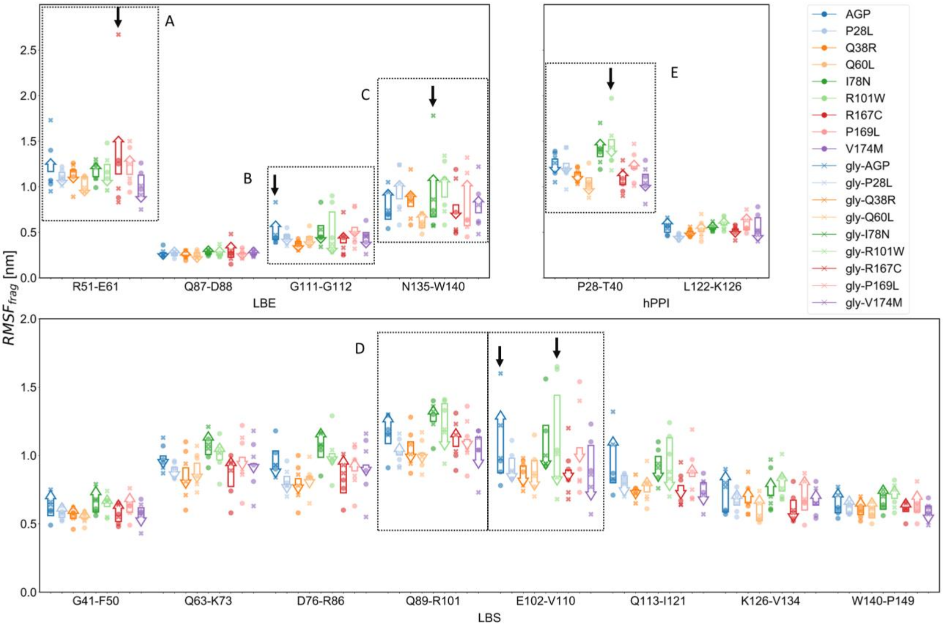


Figure 3 $RMSF_{frag}$ of fragments in the LBE, LBS, and hPPI regions of AGP and its mutants. The colored arrows indicate change in RMSF ($\Delta RMSF_{frag}$) upon glycosylation. The boxes with black arrows are labelled from A to D for the purpose of discussion.

3.2. Mutans span distinct local conformational subspaces

To identify similarities in the motions of all systems, we analysed the MD trajectories using PCA. We found that all systems required the first 10 eigenvectors to capture 80% of its protein motions apart from R167C and P28L which required the first 5 and first15 eigenvectors respectively (Figure S24). This could be due to the high flexibility of N- and C-termini compared to the other systems, resulting in more local variation being captured in the first 10 eigenvectors compared to other systems, where larger scale motions might dominate. Also, in the projection of the MD trajectories onto the first three PCs (Figure S25), Q38R and P28L demonstrated a compact cluster indicating lower conformational flexibility, while the projections had a wider distribution for other systems. By visualizing the full molecular all-atom MD trajectories, we confirmed that this wider distribution observed in the other systems was indeed due to large amplitude motions mainly involving the C-terminal and loop motions in LBE and hPPI (Figure 4). In addition, we also observed that in the cases of gly-AGP, gly-P28L, gly-I78N, gly-Q60L, and gly-R101W, the first 10 eigenvectors captured more variance than their corresponding unglycosylated systems (Figure S24). Again, we noted that these mutants demonstrated higher RMSF in termini post-glycosylation (Figure S9).

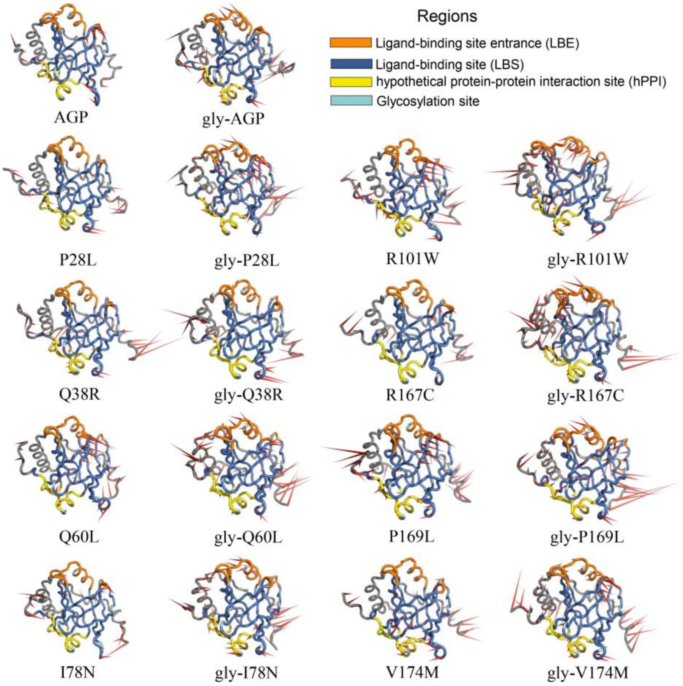


Figure 4 Porcupine plot of first 3 eigenvectors showing dominant motions in AGP and its mutants: unglycosylated or glycosylated. The 3D structure is their equilibrated initial structure of the NPT run. LBE (orange), LBS (blue), and hPPI (yellow) regions are coloured, and glycosylation sites (cyan) are indicated.

Further, to assess the similarity of conformationally variable motions amongst all systems, the similarity between the first three PCs was quantified using RMSIP (see Methods, Figure 5). There was a low overlap between the distribution of RMSIP values of unglycosylated mutants and that of glycosylated mutants (Figure S26), indicating that mutations and glycosylation induced a significant difference in the average. Nevertheless, the RMSIP distribution was wider for the unglycosylated mutants than for the glycosylated mutants (Figure S26). This was caused by the unglycosylated mutants P28L, Q38R, R167C, and Q60L, out of which P28L showed substantially lower RMSIP (below 0.5) with all the other systems, while the remaining mutants often showed low overlap with all the other systems(Figure 5). This means that the lowest three PCs of those unglycosylated mutants span different conformational subspaces. By visualizing the projections of first three PCs on MD trajectory, we observed that compared to all the systems, P28L showed low flexibility of loops and termini. In P28L, the low flexibility of termini and loops was also observed in RMSF plot. It implies that, in distinct cases, a combination of glycosylation with a single point mutation can trigger different sampling of protein conformations. Such outlier mutants are not present within the group of the glycosylated mutants, leading to a more modest variability of the RMSIP values (Figure S26).

The PCA results confirmed the complex behaviour of mutations and glycosylation on backbone flexibility. Thus, we can infer that the intrinsic motions of AGP remain conserved with slight perturbations along the LBE and hPPI. These perturbations might be enough to retain the required mobility, to enable conformational rearrangements for facilitating diverse interactions. We can also conclude that some distinct combinations of mutation and glycosylation may significantly alter the conformer distributions in mobility and dynamics.

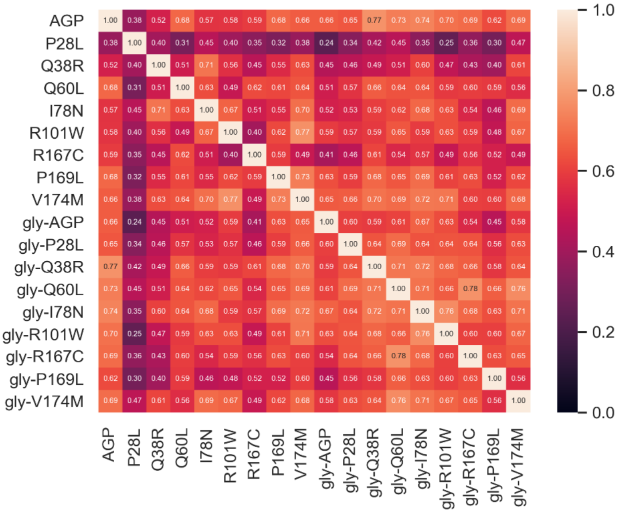


Figure 5. RMSIP scores based on the first 3 principal components between all systems.

3.3. Glycan dynamics

Understanding the conformational and biophysical behaviour of glycans is crucial for understanding their effect on the flexibility and accessibility of a protein and whether their behaviour is further influenced by mutations. We characterized the N-glycan conformers of all 5 glycan chains in gly-AGP and its glycosylated mutants, linked to N33, N56, N72, N93, and N103 via β-GlcNAc (Figure 1), by plotting ϕ and ψ glycosidic torsion angles in a carb-Ramachandran plot (Figure S32). Each of the 9 glycosylated systems has the same glycosylation pattern, and for each system, the ϕ, ψ, and/or ω torsion angles were computed for all 58 glycosidic linkages and 5 N-glycosidic linkages per glycosylated system, resulting in a total of 63 linkages per system. For simplicity, we focus on the ϕ and ψ for the most relevant glycosidic linkages (Figure S33). We observed changes in rotameric conformational states of glycans amongst different systems over the simulation trajectory of 300 ns (Figure 7). Even if transitions are observed, the timescale of our simulations might not be sufficient to observe all rotameric transitions or full convergence of glycan dynamics.[73] These transitions in rotameric states correspond to the changes from one cluster in the ϕ and ψ distribution to another cluster,[74] such as the motion of two glycans in the glycosidic linkage in the opposite directions. This information on rotameric states is valuable for investigating dynamic properties of glycans, such as their flexibility in the context of proteins and (dis)similarity in their essential conformational space due to mutations at the protein level. Thus, we discuss 1) asparagine sidechain conformations calculated by χ angles and their effect on ϕ, ψ and, ω distributions of N-glycosidic linkages, 2) ϕ, ψ, and ω distributions of glycans in site-specific glycan chains amongst gly-AGP and its glycosylated mutants.

First, the asparagine sidechain and the relevant N-glycosidic linkages are discussed. The relative orientations of the N-linked β-GlcNAc and the protein can be influenced by the sidechain conformations of (glycosylated) asparagine and other amino acid residues around glycosylation sites.[12] In unglycosylated mutants, asparagine demonstrated distinct sidechain conformers for N33 and N103 (Figure S27, Figure S31), and multiple populated conformers for N56, N72, and N93 (Figure S28-30). For glycosylated mutants, the sidechain conformations of asparagine were in one stable conformation over 300 ns, apart from fluctuations observed in N93 of all systems except gly-I78N and gly-V174M. N72 and N56 also showed fluctuations in gly-Q38R, gly-R101W, and gly-P169L, and gly-P169L and R167C respectively. N33 and N103 rarely showed fluctuations in the asparagine side-chain conformations. . This is expected, as N-glycosylation is known to reduce the conformational degrees of freedom of the N-linked asparagine sidechain.[75]

To analyse the effect of a mutation on the N-linked β-GlcNAc, we analysed its ϕ, ψ and, ω distributions across the protein variants, with the average mean and standard deviation (std) of their ϕ and ψ reported in Table S6. The ± sign of ϕ and ψ indicated the direction of rotation along torsion angle (clockwise/anti-clockwise). Overall, the five β-GlcNAc(1→)N linkages showed three different outcomes in carb-Ramachandran plot: 1) two clusters in both ±ϕ and ±ψ demonstrating an opposite motion between asparagine and β-GlcNAc indicating a flexible *linkage*, 2) a single cluster in ±ϕ and two clusters in ±ψ, implying rigid β-GlcNAc and a flexible N-glycosidic bond, and 3) a single cluster in both ±ϕ and ±ψ implying a rigid N-glycosidic linkage (Figure 6). We calculated the circular standard deviation (csd as explained in methods) of these angles to compare their spread. In summary, all N-glycosidic linkages demonstrated restricted conformational space, with ϕ in general having a wider-angle distribution than ψ. This observation is supported by a previous MD study of 5 glycosylated and 4 unglycosylated protein pairs obtained from NMR data, in which ϕ also demonstrated this behavior, indicating higher flexibility.[76] Another statistical study of 26 X-ray resolved glycoproteins and 44 glycosylation sites mentioned that the ϕ torsion angles of N-glycosidic linkage show higher amplitude indicating higher flexibility than *ψ. In*terestingly, the rigid N-glycosidic linkages were observed in β-GlcNAc(1→)N33 in gly-Q38R, β-GlcNAc(1→)N56 in gly-P28L, gly-Q38R, and gly-Q60L, β-GlcNAc(1→)N72 in gly-P28L and gly-R167C, β-GlcNAc(1→)N93 in gly-I78N. The β-GlcNAc(1→)N103 linkage shows fluctuations in both ϕ and ω for all glycosylated systems (average csd of 70°). The rigid N-glycosidic linkages might influence subsequent glycans to orient themselves in a specific direction, so restricting their range of conformations until the glycan chain termini, as observed previously.[76] This might also decrease the protein's SASA around the region of weak H-bonds and increase it at the protein surface of protein which was previously occupied. Given the difference in local interactions of N-linked β-GlcNAc among the mutants, the SASA around each glycosylation site may differ from gly-AGP. Thus, glycan chains' SASA around their core might also be affected due to these structural dynamics.

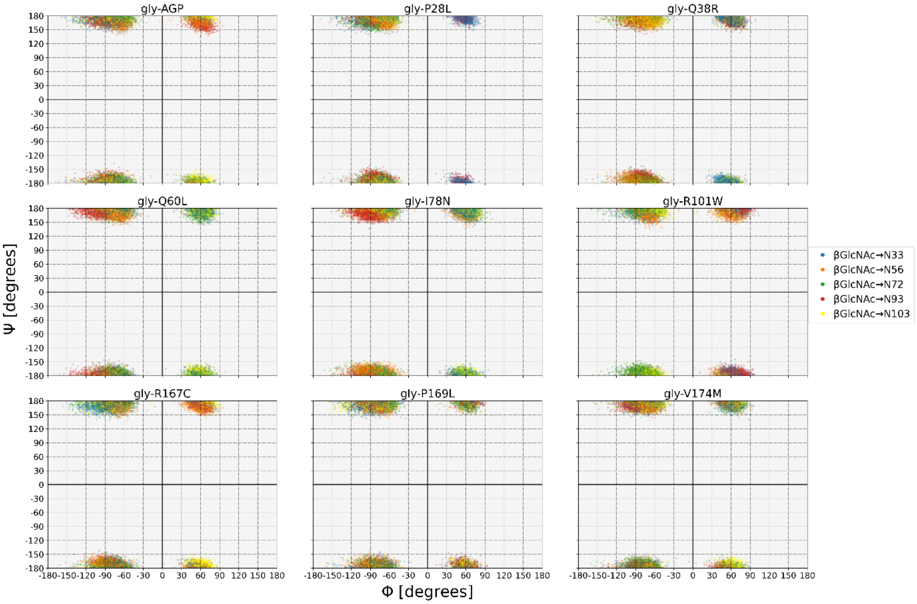


Figure 6 Carb-Ramachandran plot of ϕ and ψ torsion angles of N-glycosidic linkage at five N-glycosylation sites of gly-AGP and its glycosylated mutants for all replicates. Each colour represents a different chain: the β-GlcNAc(1→)N linkage at N33 (blue), N56 (orange), N72 (green), N93 (red), and N103 (yellow).

Secondly, to understand the conformational dynamics of consecutive glycans in a single glycan chain, we discuss the ϕ and ψ in each glycan chain amongst gly-AGP and its mutants. The mean csd for the ϕ and ψ of glycosidic linkages for all systems, as well as ω in the case of 1→6 and 2→6 linkages are reported in Table S7. In glycan chain I, all glycan linkages show similar ϕ and ψ distributions within a well-defined cluster, with some smaller clusters present (Figure S34). The ϕ conformational space is restricted with an mean csd of 14°. The ψ conformational space is more variable, with the highest flexibility (mean csd of 37° in ψ) observed primarily in β-GlcNAc2→β-GlcNAc1. The large csd of β-GlcNAc2→β-GlcNAc1 indicates local transitions in the linkage over time due to transient interactions with nearby glycans from glycan chains III and V. The second highest ψ flexibility is in the α-Man7→β-Man3 linkage in all the systems (mean csd of 34° in ψ and 67° in ω). This is likely due to the presence of an extra C atom in the (1→6) and (2→6) linkages, allowing for more conformational freedom and influencing the flexibility of adjacent glycans. (Table S7). Overall, the glycan chain is more flexible at its core region than its terminal.

Glycan chain II is similar to glycan chain I, except for an additional terminal NeuAc (Figure 1). Chain II shows more variability in the ϕ and ψ distributions than chain I, with some specific interactions present, seemingly driven by the mutations, that cause shifts in glycan chain orientation (see Figure S35) and intermittent formation of hydrogen bonds (Figure S39, Figure S40). Glycan chains III, IV, and V have larger tri-antennary and tetra-antennary structures compared to chains I and II. These larger chains enable interactions with distant glycan chains (Figure S36-38) and result in higher overall flexibility, especially in chain III (β-GlcNAc→α-Man ψ and ϕ with mean csd of 36° and 16°, respectively) (Table S7). The behaviour of the β-GlcNAc→α-Man linkage also affects the dynamics of the consecutive β-Gal→β-GlcNAc (mean csd of 29° in ϕ and 7° in ψ), with different conformer distributions depending on their chain III branch position.

In chain IV, the two terminal α-Neu5Ac show high ϕ variability (mean csd of 25° in ϕ and 16° in ψ) than other two α-Neu5Ac with 2→6 linkages showing higher ψ , and ω flexibility (mean csd of 13° in ϕ, 20° in ψ, and 65° in ω) due to the extra C atom in its (2→6) linkage (Figure S37). Additionally, in chain IV, one of the terminal α-Neu5Ac, and β-Gal formed transient H-bonds with G111 and G112 of LBE, affecting their flexibility (Figure S44). In conclusion, chain IV demonstrated an umbrella conformation with equally positioned antennae around the protein surface, causing a shielding effect (discussed in the next section).[77]

Chain V shows in general very similar ϕ and ψ distributions to other chains, apart from the α-Neu5Ac→β-Gal linkages (mean csd of 24° in ϕ, and 16° in ψ), core α-Fuc (Figure S38), which seems to affect the conformational space of its N-linked β-GlcNAc with a mean csd of 68° in ϕ, 9° in ψ, and 71° in ω (Table S7). Our findings were in line with Fernandes et al., 2015, who investigated the effect of fucosylation on glycosylated AGP and its binding to P-selectin using molecular dynamics modeling.[78] They compared three systems: AGP, gly-AGP, and core-fucosylated gly-AGP. The two glycosylated systems consisted of a tri-antennary glycan at N33, a bi-antennary glycan at N56, and identical, tetra-antennary glycans at N72, N92, and N103. The fucosylated system was the same except that it had a core-fucosylation on N72, N92, and N103. They observed that glycosylation reduces the RMSF of the protein, indicating a reduction in AGP's backbone flexibility, and by analyzing the glycan angle distributions, they concluded that glycan chains affect the backbone dynamics of AGP around its ligand binding site as well as the binding of P-selectin to AGP via glycans attached at N72 and N93.

In summary, glycan chains show very diverse conformational behaviour linked to their branching, linkage type, and composition, which is closely intertwined with the protein's conformational dynamics. Our analysis suggests that especially glycan chain IV may be particularly important due to its presence in the LBE region, where it might restrict access of ligands. Its tetra-antennary structure also allows it to interact with distant glycans. Cancer cells are frequently observed to overexpress aberrant α-Neu5Ac linkages, which might then block access to functionally relevant enzymes and promote metastasis[79]. Based on our analysis, mutations can differentially control the conformational dynamics of glycan chains by disrupting their local interactions via (intermittent) H-bonds. Moreover, the changes in conformational behaviour of the glycan chains due to mutations might also differentially impact the SASA of AGP, potentially leading to changes in its binding specificities to ligands.

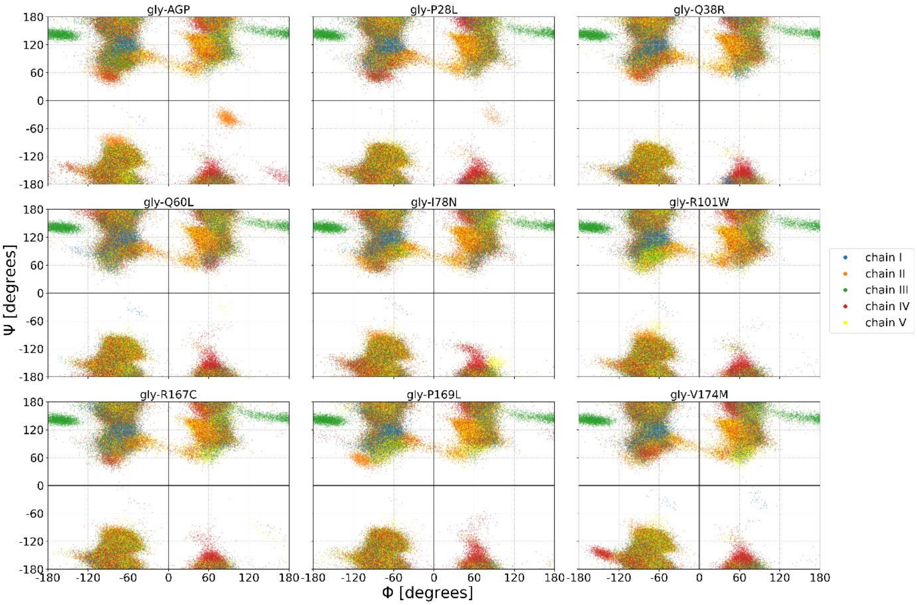


Figure 7 Carb-Ramachandran plot for ϕ, ψ torsion angles of all 58 glycosidic linkages. Each color represents a different chain, namely chain I, II, III, IV, and V.

3.4. SASA

In the lipocalin family of proteins, the loop scaffolds that constitute the LBE and hPPI regions participate in protein-protein interactions.[25,80,81] The variability in the composition, length, and conformation of loops can lead to high affinity and selectivity of ligands which might be affected by mutations.[25]. However, since AGP is a glycoprotein, it is also crucial to consider the role of glycosylation along with mutations, as it is known to significantly influence the solvent accessibility of the amino acid surface of a protein, thereby affecting its biological function.[82] Given that mutations interrupted glycan-protein interactions and influence the conformational preferences of glycans, we investigated the changes in SASA for both protein and glycans using 3 probe sizes (radius 0.14 nm, 0.50 nm, and 1 nm, see methods).

Firstly, we investigated the effect of glycosylation on protein's SASA, by calculating the difference in SASA between glycosylated and unglycosylated proteins ($\Delta SASA_{frag}$) (computed similarly to Eq. (6) for the difference in RMSF). Our findings show that overall the $SASA_{frag}$ decreases with increasing probe size (Figure 8) as shown on y-axis, indicating that the amino acid surface of the protein becomes especially less accessible upon glycosylation for small-molecule drug and protein domain interactions.

In the three defined regions considering all probe sizes (**Error! Reference source not found.**), the strongest reduction in SASA occurred in the fragments R51-E61, Q89-R101, E102-V110, and P28-T40, which contain the glycosylation sites. In addition, N135-W140 in LBE also showed a decrease in SASA as probe size increased from 0.14 nm to 0.5 nm and 1 nm. Consistent with so-called glycan shielding as discussed in the next section, the accessibility drops significantly in these fragments upon glycosylation. Comparing mutations, all mutants showed distinct accessibility varying from fragment to fragment. Mutant P28L showed the strongest decrease in accessibility in fragment N135-W140 of LBE for 0.5 nm and 1 nm probes. Meanwhile, for the same fragment, it showed the weakest decrease in accessibility due to glycans for a 0.14 nm probe. In another example, for 0.14 nm probe, mutant V174M showed strongest decrease in accessibility in N135-W140 of LBE, weakest decrease in Q63-K73 or even an increase in accessibility in K126-V134 of LBS, and weakest decrease in accessibility in P28-T40. However, when the probe size is increased, V174M no longer showed the strongest or weakest change in accessibility except for fragment Q63-K73. This indicates that a general trend cannot be formulated concerning the strength of the glycans' and mutations' impact on SASA for individual mutants: the decrease in SASA can be strong in one fragment but weak (or even an increase) in another fragment of the protein. This pattern further varies across different probe sizes.

Next, focusing on the effect of glycosylation specifically on LBE, the region crucial to facilitate the entrance of binding partners of AGP to LBS, we discuss the average $SASA_{region}$ (as explained in methods). In unglycosylated systems, the average $SASA_{region}$ was higher for all three probe sizes compared to glycosylated systems, indicating a substantial reduction in LBE's SASA due to glycosylation (Table S5). Additionally, glycosylated systems showed higher standard deviation in $SASA_{region}$ for all probe sizes (Table S5). These observations imply that variation in mean $SASA_{region}$ values amongst glycosylated systems is more pronounced compared to unglycosylated systems, as evident from their higher average std. This variability is mainly due to the dynamic behaviour of glycan chain IV at N93, which transiently interacts with the residues G111 and G112 of LBE, exhibiting a potential shielding effect by restricting LBE access.

Thus, to gain further insight into glycan shielding, we computed the number of glycan chain IV atoms within a cut-off distance of 2.6 nm around LBE atoms. Lee and workers employed a 2.6 nm cut-off to measure glycan coverage from surface protein residues by analysing the longest length of the glycan in HIV-1 Env protein trimer averaged over 500 ns MD simulations[83]. In addition, another study observed that glycan residues that were approximately 2.0 nm from N-glycosylation sites often interacted with protein surface[13]. On the other hand, an experimental study based on Hydrogen-Deuterium (H-D) exchange showed that N-glycan affected the H-D exchange rate of protein residues within 3.0 nm from the N-glycosylation site[84]. Thus, using the 2.6 nm cut-off as compromise, we observed that on an average 420 to 451 glycan chain IV atoms were proximal to LBE, out of a total of 454 atoms (Figure S20). In addition, amongst the glycosylated systems gly-I78N showed the lowest number of average glycan atoms around LBE (420±16) and highest mean $SASA_{region}$ for all three probe sizes indicating low glycan shielding and high solvent accessibility (Table S 5). The conformational impact of the I78N mutation led to substantial increase in transient interactions between glycan III and chain IV, consequently moving the glycan chain IV away from LBE increasing its solvent accessibility (Figure S20). These results show that glycosylation has a more significant impact on altering the SASA of LBE compared to single amino acid mutations. Thus, we can infer that the effect of glycans on solvent accessibility is far more complex as their conformational dynamics also play an important role. With their dynamics altered due to mutations (see previous section), it is reasonable to assume that solvent accessibility can be disrupted unexpectedly due to mutations. Such changes in biophysical behaviour and solvent accessibility should influence the interactions with diverse binding partners of AGP and could lead to disruptions of its functional behaviour.

Next, to further investigate the effect of mutations on SASA of AGP's glycans, we calculated the SASA per glycan chain for all three probe sizes as these may also affect binding interactions of relevant glycosylation processing enzymes with AGP. Specifically, from the perspective of cancer, where complex glycan branching and elongation are observed in AGP, the access of glycosyltransferases and glycosidases to the glycan core is essential for the processing of N-glycans[85]. The results show that glycan chain I and III were comparatively less exposed than the other glycan chains with β-Gal showing low SASA than other terminal glycans α-Neu5Ac10 (Figure S21). The glycan core showed the lowest SASA for all glycan chains. Further, amongst all glycan chains for all systems, the terminal α-Neu5Ac of chain III, IV, and V showed the highest and most variable SASA apart from α-Neu5Ac10 (Figure S21). Due to intermittent H-bonds between glycan chain IV (via α-Neu5Ac10, β-Gal9, and β-GlcNAc8) and amino acid residues of LBE, which lasted for several nanoseconds before breaking, α-Neu5Ac10 showed relatively low SASA compared to other α-Neu5Ac. Thus, the terminal α-Neu5Ac10 of tetra-antennary glycan chain IV might play a critical role in regulating the access of ligands to

LBE.

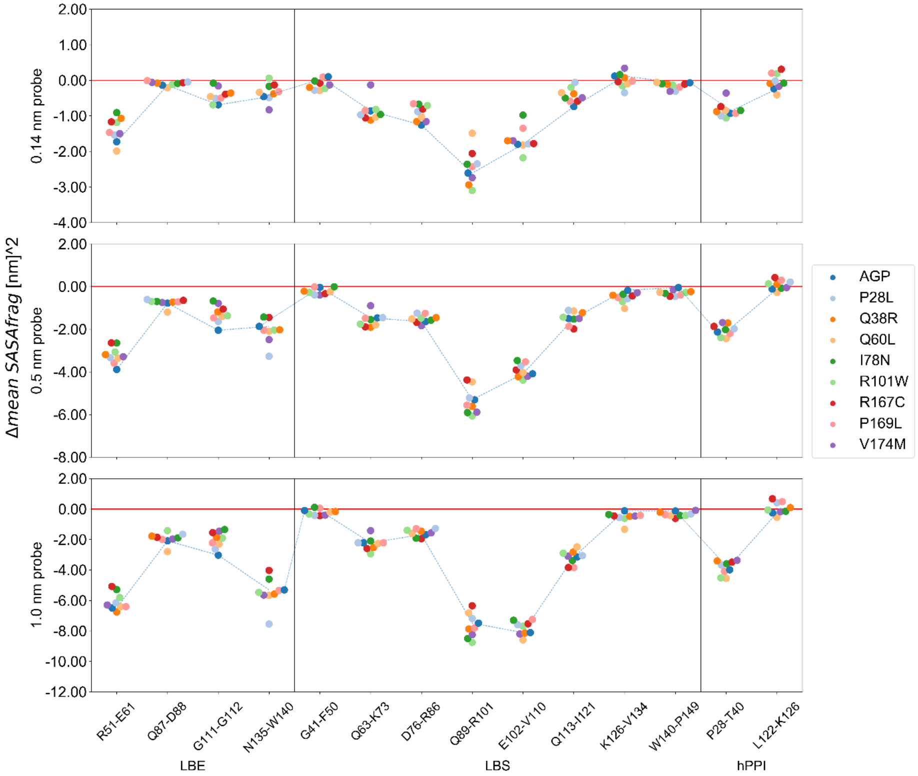


Figure 8 $\Delta SASA_{frag}$ of fragments per region of LBE, LBS, and hPPI of AGP and its mutants. The red line indicates no change in SASA upon glycosylation. The blue dashed line connects the values for (gly-)AGP to guide the eye. SASA with three different probe sizes with radius 0.14 nm, 0.50 nm, and 1 nm are shown in the three panels.

Conclusion

We carried out a total of 54 MD simulations with 3 replicas of 100 ns each for 18 systems of AGP. Our findings, illustrate the complexity of the interplay between the effects of glycosylation and amino acid mutations. In unglycosylated mutants, point mutations alter the local flexibility of the protein due to their long-range conformational effect. In glycosylated systems of AGP, we found that glycosylation decreases the local flexibility at the site of glycosylation and increases the flexibility of structurally distant regions. Considering the associated effect of mutation and glycosylation at the protein structure level in glycosylated systems, we found that point mutations reshuffle glycan dynamics by changing their rotameric states, disrupting the transient H-bond interactions between protein-glycan, intra-glycan, and inter-glycan. Amongst all glycosidic linkages, we found that β-GlcNAc→α-Man and α-Neu5Ac→β-Gal show the highest flexibility and influence the flexibility of adjacent glycans. Amongst the five glycan chains, we found that glycan chain IV is conformationally versatile and restricts the access of LBE, while interacting with other chains due to its tetra-antennary branching and complex composition. Due to mutation-induced flexibility changes in glycans, we found that the solvent accessibility differs significantly for different glycosylated mutants for distinct probe sizes. These general conclusions illustrate that the combined effect of mutations and glycosylation on the behavior of AGP (its 'molecular phenotype') is highly complex and ambiguous.

When we consider the combined impact of mutations and glycosylation, glycosylated mutants exhibit greater similarity in their backbone dynamics. However, no specific set of mutants can be identified with specific similar biophysical effects. Amongst all the investigated mutants, our findings indicate that R101W substitution exhibits the most substantial impact on AGP's conformational dynamics as shown by RMSF and SASA. In addition, the mutation W101 is located structurally proximal and equidistant to three glycosylation sites: site I at N33, site V at N103, and site III at N72. This precise location of this mutation emerges as structurally critical, as it substantially affects the interactions among these glycan chains which in turn affects the overall glycans' and protein's dynamics. Thus, our research strongly indicates that this distinctive position enables the mutant with the ability to disrupt interactions more effectively when compared to other mutants. Consequently, this mutation-induced alteration in glycan dynamics might exert a substantial immunomodulatory effect within the cancer-associated cellular mechanisms. These conclusions also highlight the importance of considering the *in vivo* context of the protein into account in relation to cancer, such as glycosylation patterns, which as shown here can significantly alter the impact of mutations. Indeed, at the glycan level, due to the microheterogeneity of glycans and their individual- and disease-specific characteristics, more studies are required to investigate varying glycoforms of AGP. Generally, in contrast to proteins, glycosylation is not template driven. Although N-glycans share a similar core, they differ widely amongst individuals in terms of branching, linkages, and composition, depending on cellular environments, APRs, and pathophysiological conditions, complicating this picture even further.[8]

The changes in conformational behaviour of AGP and its mutants before and after glycosylation could, however, help explain why specific binding partners of AGP might be unable to access its key binding site in different cancer mutants. At the same time, the distinct conformational changes might enable new binding and recognition specificities leading to altered functional mechanisms for different mutants. Thus, our findings help build a biophysical description of the importance of cancer-associated glycans to AGP behavior and their influence on AGP's function, also in relation to cancer mutants. These findings might extend to the lipocalin family, but this will require further experimental and computational studies. On the other hand, our findings of the expected accessibility of key binding sites might assist in designing and selecting the efficient binding partners of AGP in cancer patients. We hope that the availability of our molecular dynamics data will help for further studies on the topic, thus contributing to a deeper and more complex understanding of AGP.

Acknowledgements

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Vrije Universiteit Brussel, Ghent University, FWO and the Flemish Government – department EWI. We acknowledge funding to the BOF of Ghent University and to the FWO (projects G002520N, G094023N, G.032816N and G.028821N).

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.

References

1. Schmid K. PREPARATION AND PROPERTIES OF AN ACID GLYCOPROTEIN PREPARED FROM HUMAN PLASMA. *J Am Chem Soc*. 1950;72(6):2816-2816. doi:10.1021/ja01162a553

2. Weimer HE, Mehl JW, Winzler RJ. STUDIES ON THE MUCOPROTEINS OF HUMAN PLASMA. *Journal of Biological Chemistry*. 1950;185(2):561-568. doi:10.1016/S0021-9258(18)56341-9

3. Lehman-McKeeman LD. Absorption, Distribution, and Excretion of Toxicants. In: *Casarett and Doull's Toxicology: The Basic Science of Poisons, 8e*. McGraw-Hill Education; 2012. Accessed December 16, 2021. accesspharmacy.mhmedical.com/content.aspx?aid=1100085093

4. Fischer K, Kettunen J, Würtz P, et al. Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality: An Observational Study of 17,345 Persons. *PLOS Medicine*. 2014;11(2):e1001606. doi:10.1371/journal.pmed.1001606

5. Ruiz M. Into the Labyrinth of the Lipocalin α1-Acid Glycoprotein. *Frontiers in Physiology*. 2021;12:847. doi:10.3389/fphys.2021.686251

6. Yuasa I, Umetsu K, Vogt U, et al. Human orosomucoid polymorphism: molecular basis of the three common ORM1 alleles, ORM1*F1, ORM1*F2, and ORM1*S. *Hum Genet*. 1997;99(3):393-398. doi:10.1007/s004390050378

7. Schönfeld DL, Ravelli RBG, Mueller U, Skerra A. The 1.8-Å Crystal Structure of α1-Acid Glycoprotein (Orosomucoid) Solved by UV RIP Reveals the Broad Drug-Binding Activity of This Human Plasma Lipocalin. *Journal of Molecular Biology*. 2008;384(2):393-405. doi:10.1016/j.jmb.2008.09.020

8. Ceciliani F, Pocacqua V. The Acute Phase Protein &#945;1-Acid Glycoprotein: A Model for Altered Glycosylation During Diseases. *Current Protein & Peptide Science*. 2007;8(1):91-108. doi:10.2174/138920307779941497

9. Smith SA, Waters NJ. Pharmacokinetic and Pharmacodynamic Considerations for Drugs Binding to Alpha-1-Acid Glycoprotein. *Pharmaceutical Research*. 2019;36(2). doi:10.1007/s11095-018-2551-x

10. Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database11Dedicated to Prof. Akira Kobata and Prof. Harry Schachter on the occasion of their 65th birthdays. *Biochimica et Biophysica Acta (BBA) - General Subjects*. 1999;1473(1):4-8. doi:10.1016/S0304-4165(99)00165-8

11. Varki A, Cummings RD, Esko JD, et al., eds. *Essentials of Glycobiology*. 2nd ed. Cold Spring Harbor Laboratory Press; 2009. Accessed December 5, 2021. http://www.ncbi.nlm.nih.gov/books/NBK1908/

12. Jo S, Lee HS, Skolnick J, Im W. Restricted N-glycan Conformational Space in the PDB and Its Implication in Glycan Structure Modeling. *PLOS Computational Biology*. 2013;9(3):e1002946. doi:10.1371/journal.pcbi.1002946

13. Nagae M, Yamaguchi Y. Function and 3D Structure of the N-Glycans on Glycoproteins. *Int J Mol Sci*. 2012;13(7):8398-8429. doi:10.3390/ijms13078398

14. Fournier T, Medjoubi-N N, Porquet D. Alpha-1-acid glycoprotein. *Biochim Biophys Acta*. 2000;1482(1-2):157-171. doi:10.1016/s0167-4838(00)00153-9

15. Treuheit MJ, Costello CE, Halsall HB. Analysis of the five glycosylation sites of human alpha 1-acid glycoprotein. *Biochem J*. 1992;283(Pt 1):105-112.

16. Perkins SJ, Kerckaert JP, Loucheux-Lefebvre MH. The shapes of biantennary and tri/tetaantennary α1 acid glycoprotein by small-angle neutron and X-ray scattering. *European Journal of Biochemistry*. 1985;147(3):525-531. doi:10.1111/j.0014-2956.1985.00525.x

17. Johnson DA, Smith KD. The efficacy of certain anti-tuberculosis drugs is affected by binding to α-1-acid glycoprotein. *Biomedical Chromatography*. 2006;20(6-7):551-560. doi:10.1002/bmc.641

18. Casalino L, Gaieb Z, Goldsmith JA, et al. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent Sci*. 2020;6(10):1722-1734. doi:10.1021/acscentsci.0c01056

19. Herve F, Gomas E, Duche JC, Tillement JP. Evidence for differences in the binding of drugs to the two main genetic variants of human alpha 1-acid glycoprotein. *Br J Clin Pharmacol*. 1993;36(3):241-249.

20. Yokogawa K, Shimomura S, Ishizaki J, et al. Involvement of α1-acid glycoprotein in inter-individual variation of disposition kinetics of ropivacaine following epidural infusion in off-pump coronary artery bypass grafting. *Journal of Pharmacy and Pharmacology*. 2007;59(1):67-73. doi:10.1211/jpp.59.1.0009

21. Smith KD, Paterson S. Binding of alpha-1-acid glycoprotein to imatinib following increased dosage of drug. *Haematologica*. 2005;90 Suppl:ELT01.

22. Ascenzi P, Fanali G, Fasano M, Pallottini V, Trezza V. Clinical relevance of drug binding to plasma proteins. *Journal of Molecular Structure*. 2014;1077:4-13. doi:10.1016/j.molstruc.2013.09.053

23. Huang Z, Ung T. Effect of alpha-1-acid glycoprotein binding on pharmacokinetics and pharmacodynamics. *Curr Drug Metab*. 2013;14(2):226-238.

24. Munkley J, Elliott DJ. Hallmarks of glycosylation in cancer. *Oncotarget*. 2016;7(23):35478-35489. doi:10.18632/oncotarget.8155

25. Flower DR. The lipocalin protein family: structure and function. *Biochem J*. 1996;318(Pt 1):1-14.

26. Varki A, Cummings RD, Aebi M, et al. Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology*. 2015;25(12):1323-1324. doi:10.1093/glycob/cwv091

27. Neelamegham S, Aoki-Kinoshita K, Bolton E, et al. Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology*. 2019;29(9):620-624. doi:10.1093/glycob/cwz045

28. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. 2021;49(D1):D480-D489. doi:10.1093/nar/gkaa1100

29. Wang Y, Wang Q, Huang H, et al. A crowdsourcing open platform for literature curation in UniProt. *PLOS Biology*. 2021;19(12):e3001464. doi:10.1371/journal.pbio.3001464

30. Kagami LP, Orlando G, Raimondi D, et al. b2bTools: online predictions for protein biophysical features and their conservation. *Nucleic Acids Research*. 2021;49(W1):W52-W59. doi:10.1093/nar/gkab425

31. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*. 2019;47(D1):D941-D947. doi:10.1093/nar/gky1015

32. Zhang Z, Miteva MA, Wang L, Alexov E. Analyzing Effects of Naturally Occurring Missense Mutations. *Comput Math Methods Med*. 2012;2012:805827. doi:10.1155/2012/805827

33. Alocci D, Mariethoz J, Gastaldello A, et al. GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. *Journal of Proteome Research*. 2019;18(2). doi:10.1021/acs.jproteome.8b00766

34. Higai K, Aoki Y, Azuma Y, Matsumoto K. Glycosylation of site-specific glycans of α1-acid glycoprotein and alterations in acute and chronic inflammation. *Biochimica et Biophysica Acta (BBA) - General Subjects*. 2005;1725(1). doi:10.1016/j.bbagen.2005.03.012

35. Lauc G, Huffman JE, Pučić M, et al. Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet*. 2013;9(1):e1003225. doi:10.1371/journal.pgen.1003225

36. Nilsson J, Rüetschi U, Halim A, et al. Enrichment of glycopeptides for glycan structure and attachment site identification. *Nat Methods*. 2009;6(11):809-811. doi:10.1038/nmeth.1392

37. Perdivara I, Peddada SD, Miller FW, Tomer KB, Deterding LJ. Mass spectrometric determination of IgG subclass-specific glycosylation profiles in siblings discordant for myositis syndromes. *J Proteome Res*. 2011;10(7):2969-2978. doi:10.1021/pr200397h

38. Sturiale L, Nassogne MC, Palmigiano A, et al. Aberrant sialylation in a patient with a HNF1α variant and liver adenomatosis. *iScience*. 2021;24(4):102323. doi:10.1016/j.isci.2021.102323

39. Zhang Y, Zhao W, Mao Y, et al. Site-specific N-glycosylation Characterization of Recombinant SARS-CoV-2 Spike Proteins. *Mol Cell Proteomics*. Published online October 19, 2020:100058. doi:10.1074/mcp.RA120.002295

40. Shajahan A, Archer-Hartmann S, Supekar NT, Gleinich AS, Heiss C, Azadi P. Comprehensive characterization of N- and O- glycosylation of SARS-CoV-2 human receptor angiotensin converting enzyme 2. *Glycobiology*. 2021;31(4):410-424. doi:10.1093/glycob/cwaa101

41. Hwang H, Lee JY, Lee HK, et al. In-depth analysis of site-specific N-glycosylation in vitronectin from human plasma by tandem mass spectrometry with immunoprecipitation. *Anal Bioanal Chem*. 2014;406(30):7999-8011. doi:10.1007/s00216-014-8226-5

42. Yoshima H, Matsumoto A, Mizuochi T, Kawasaki T, Kobata A. Comparative study of the carbohydrate moieties of rat and human plasma alpha 1-acid glycoproteins. *J Biol Chem*. 1981;256(16):8476-8484.

43. Mechref Y, Zidek L, Ma W, Novotny MV. Glycosylated major urinary protein of the house mouse: characterization of its N-linked oligosaccharides. *Glycobiology*. 2000;10(3):231-235. doi:10.1093/glycob/10.3.231

44. Oliveira T, Zhang M, Joo EJ, et al. Glycoproteome remodeling in MLL-rearranged B-cell precursor acute lymphoblastic leukemia. *Theranostics*. 2021;11(19):9519-9537. doi:10.7150/thno.65398

45. Pearce OMT, Läubli H. Sialic acids in cancer biology and immunity. *Glycobiology*. 2016;26(2):111-128. doi:10.1093/glycob/cwv097

46. Chandler KB, Brnakova Z, Sanda M, et al. Site-specific glycan microheterogeneity of inter-alpha-trypsin inhibitor heavy chain H4. *J Proteome Res*. 2014;13(7):3314-3329. doi:10.1021/pr500394z

47. Pompach P, Ashline DJ, Brnakova Z, Benicky J, Sanda M, Goldman R. Protein and site specificity of fucosylation in liver-secreted glycoproteins. *J Proteome Res*. 2014;13(12):5561-5569. doi:10.1021/pr5005482

48. Oortwijn BD, Roos A, Royle L, et al. Differential glycosylation of polymeric and monomeric IgA: a possible role in glomerular inflammation in IgA nephropathy. *J Am Soc Nephrol*. 2006;17(12):3529-3539. doi:10.1681/ASN.2006040388

49. Miyoshi E, Moriwaki K, Terao N, et al. Fucosylation Is a Promising Target for Cancer Diagnosis and Therapy. *Biomolecules*. 2012;2(1):34-45. doi:10.3390/biom2010034

50. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. *Protein Science*. 2018;27(1):129-134. doi:10.1002/pro.3289

51. Park SJ, Lee J, Patel DS, et al. Glycan Reader is improved to recognize most sugar types and chemical modifications in the Protein Data Bank. *Bioinformatics*. 2017;33(19):3051-3057. doi:10.1093/bioinformatics/btx358

52. Jo S, Im W. Glycan fragment database: a database of PDB-based glycan 3D structures. *Nucleic Acids Research*. 2013;41(D1):D470-D474. doi:10.1093/nar/gks987

53. Huang J, Rauscher S, Nawrocki G, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods*. 2017;14(1):71-73. doi:10.1038/nmeth.4067

54. Guvench O, Hatcher ER, Venable RM, Pastor RW, MacKerell AD. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J Chem Theory Comput*. 2009;5(9):2353-2370. doi:10.1021/ct900242e

55. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*. 1997;18(12):1463-1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

56. Darden T, Perera L, Li L, Pedersen L. New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*. 1999;7(3):R55-60. doi:10.1016/s0969-2126(99)80033-1

57. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*. 1995;103(19):8577-8593. doi:10.1063/1.470117

58. Abraham MJ, Murtola T, Schulz R, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015;1-2:19-25. doi:10.1016/j.softx.2015.06.001

59. Knapp B, Ospina L, Deane CM. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *J Chem Theory Comput*. 2018;14(12):6127-6138. doi:10.1021/acs.jctc.8b00391

60. Gowers RJ, Linke M, Barnoud J, et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference*. Published online 2016:98-105. doi:10.25080/Majora-629e541a-00e

61. Gowers RJ, Linke M, Barnoud J, et al. *MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations*. Los Alamos National Lab. (LANL), Los Alamos, NM (United States); 2019. doi:10.25080/Majora-629e541a-00e

62. McGibbon RT, Beauchamp KA, Harrigan MP, et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J*. 2015;109(8):1528-1532. doi:10.1016/j.bpj.2015.08.015

63. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol*. 1973;79(2):351-371. doi:10.1016/0022-2836(73)90011-9

64. Mehdipour AR, Hummer G. Dual nature of human ACE2 glycosylation in binding to SARS-CoV-2 spike. *PNAS*. 2021;118(19). doi:10.1073/pnas.2100425118

65. Ghysels A, Miller BT, Pickard IV FC, Brooks BR. Comparing normal modes across different models and scales: Hessian reduction versus coarse-graining. *Journal of Computational Chemistry*. 2012;33(28):2250-2275. doi:10.1002/jcc.23076

66. Amadei A, Ceruso MA, Di Nola A. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*. 1999;36(4):419-424. doi:10.1002/(SICI)1097-0134(19990901)36:4<419::AID-PROT5>3.0.CO;2-U

67. Best RB, Hummer G, Eaton WA. Native contacts determine protein folding mechanisms in atomistic simulations. *Proceedings of the National Academy of Sciences*. 2013;110(44):17874-17879. doi:10.1073/pnas.1311599110

68. Franklin J, Koehl P, Doniach S, Delarue M. MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. *Nucleic Acids Research*. 2007;35(suppl_2):W477-W482. doi:10.1093/nar/gkm342

69. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*. 2022;50(D1):D439-D444. doi:10.1093/nar/gkab1061

70. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

71. Condic-Jurkic K, Subramanian N, Mark AE, O'Mara ML. The reliability of molecular dynamics simulations of the multidrug transporter P-glycoprotein in a membrane environment. *PLOS ONE*. 2018;13(1):e0191882. doi:10.1371/journal.pone.0191882

72. Lee HS, Qi Y, Im W. Effects of N-glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study. *Sci Rep*. 2015;5(1):8926. doi:10.1038/srep08926

73. Woods RJ, Tessier MB. Computational glycoscience: characterizing the spatial and temporal properties of glycans and glycan–protein complexes. *Current Opinion in Structural Biology*. 2010;20(5):575-583. doi:10.1016/j.sbi.2010.07.005

74. Gonzalez-Outeiriño J, Kirschner KN, Thobhani S, Woods RJ. Reconciling solvent effects on rotamer populations in carbohydrates — A joint MD and NMR analysis. *Can J Chem*. 2006;84(4):569-579. doi:10.1139/v06-036

75. Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology*. 2004;14(2):103-114. doi:10.1093/glycob/cwh008

76. Pol-Fachin L, Fernandes CL, Verli H. GROMOS96 43a1 performance on the characterization of glycoprotein conformational ensembles through molecular dynamics simulations. *Carbohydrate Research*. 2009;344(4). doi:10.1016/j.carres.2008.12.025

77. Montreuil J. Spatial conformation of glycans and glycoproteins. *Biology of the Cell*. 1984;51(2):115-131. doi:10.1111/j.1768-322X.1984.tb00291.x

78. Fernandes CL, Ligabue-Braun R, Verli H. Structural glycobiology of human α1-acid glycoprotein and its implications for pharmacokinetics and inflammation. *Glycobiology*. 2015;25(10):1125-1133. doi:10.1093/glycob/cwv041

79. Dobie C, Skropeta D. Insights into the role of sialylation in cancer progression and metastasis. *Br J Cancer*. 2021;124(1):76-90. doi:10.1038/s41416-020-01126-7

80. Matsumoto K, Nishi K, Kikuchi M, et al. Receptor-Mediated Uptake of Human α1-Acid Glycoprotein into Liver Parenchymal Cells in Mice. *Drug Metabolism and Pharmacokinetics*. 2010;25(1):101-107. doi:10.2133/dmpk.25.101

81. Qin Z, Wan JJ, Sun Y, et al. ORM Promotes Skeletal Muscle Glycogen Accumulation via CCR5-Activated AMPK Pathway in Mice. *Frontiers in Pharmacology*. 2016;7. Accessed November 8, 2022. https://www.frontiersin.org/articles/10.3389/fphar.2016.00302

82. Solá RJ, Rodríguez-Martínez JA, Griebenow K. Modulation of protein biophysical properties by chemical glycosylation: biochemical insights and biomedical implications. *Cell Mol Life Sci*. 2007;64(16):2133-2152. doi:10.1007/s00018-007-6551-y

83. Lee M, Changela A, Gorman J, et al. Extended antibody-framework-to-antigen distance observed exclusively with broad HIV-1-neutralizing antibodies recognizing glycan-dense surfaces. *Nat Commun*. 2021;12:6470. doi:10.1038/s41467-021-26579-z

84. Joao HC, Scragg IG, Dwek RA. Effects of glycosylation on protein conformation and amide proton exchange rates in RNase B. *FEBS Letters*. 1992;307(3):343-346. doi:10.1016/0014-5793(92)80709-P

85. Lau KS, Dennis JW. N-Glycans in cancer progression. *Glycobiology*. 2008;18(10):750-760. doi:10.1093/glycob/cwn071

Chapter 5

Gradations of protein dynamics: AlphaFold2 vs NMR

Gradations in protein dynamics captured by experimental NMR are not well represented by AlphaFold2 models and other computational metrics. Jose Gavalda-Garcia, **Bhawna Dixit**, Adrián Díaz, An Ghysels, and Wim Vranken Journal of Molecular Biology (Accepted)

5.1 Background and methodology

The key research question guiding this study is: **what is the relationship between protein flexibility predicted by computational methods and observed through experimental techniques on a large scale?**

Considering the computational expense of MD simulations and the advancements in AlphaFold2-based modeling, large-scale modeling is becoming increasingly practical. However, AlphaFold2 struggles with proteins that are not well-folded and its ability to predict multiple conformations remains limited, with scarce experimental data to validate such predictions. In addition, functionally important smaller conformational changes are challenging to capture experimentally and often require specialized NMR techniques. Computational methods offer promise, but large-scale comparisons between predicted dynamics and experimental data are still needed. AlphaFold2 partially addresses prediction accuracy through its pLDDT metric. Therefore, to address the second key research question, large-scale NMA was performed on AlphaFold2 models and their corresponding NMR ensembles. This approach aimed to explore the correlation between flexibility metrics derived from NMA, MD simulations, and experimental data, including ShiftCrypt, S^2_{RCI}, and experimental S^2 order parameters. Additionally, the correlation between the confidence of AlphaFold2 predictions indicated by the pLDDT metric and the flexibility metrics was evaluated, to address whether the prediction confidence provided by AlphaFold2 aligns with experimentally and computationally derived flexibility metrics.

For this study, three distinct datasets of AlphaFold2 models were curated, encompassing per-residue flexibility data from ShiftCrypt, S^2_{RCI}, S^2, and RMSF derived from MD simulations, along with secondary structure annotations and pLDDT scores for each model. The datasets include:

1. S^2_{RCI} dataset
2. MD dataset from Constava
3. S^2 dataset

For each of the three datasets, NMA was carried out on the AlphaFold2 models using WEBnma. For the S^2_{RCI} dataset NMA was carried out on both of the AlphaFold2 and their NMR ensembles, and the secondary structure annotations were also computed for each NMR model within the NMR ensemble. From the obtained normal modes with lowest eigenvalues, RMSF was computed for all the datasets. To investigate the relationship between the flexibility metrics, Pearson correlation coefficients were computed and analyzed at both the residue and protein levels.

5.2 Contributions

In this work, my colleague Jose Gavalda-Garcia generated three datasets by integrating multiple data sources: 1) NMR-derived metrics retrieved by Prof. Dr. Wim Vranken using an in-house pipeline, 2) MD simulations from his previous contributions to Constava, 3) AlphaFold2 models, and 4) AlphaFold3 models obtained from the AlphaFold server. Additionally, he created a Docker container to facilitate the reproducibility of both the dataset generation and analysis processes. Lastly, in collaboration with Adrián Díaz, Jose Gavalda-Garcia contributed to the development of a web service for interactive, per-protein analysis, enabling further exploration of the findings.

My key contributions include development of an open-source Python pipeline for the NMA related analysis of the AlphaFold2 models and NMR ensembles. The Python pipeline includes carrying out NMA using WEBnma tool, computing RMSF from normal modes, computing RMSF from normal modes, and analyzing and visualizing the results. The Python code for visualization related to NMA results was combined with Jose's docker pipeline. In addition, my contributions include generating some of the AlphaFold2 structures from the dataset using high-performance computing (HPC) resources.

5.3 Concluding remarks

This study explored, on a large scale, the relationships between the AlphaFold2 pLDDT metric, observed dynamics from NMR metrics, interpreted MD simulations, and computed dynamics from NMA for single AlphaFold2 models and NMR ensembles. The study found that the flexibility metrics align well for rigid residues that adopt a single, well-defined conformation, distinguishing them from residues that exhibit dynamic behavior and adopt multiple conformations. This distinction between order and disorder is reflected in the correlations between the parameters but becomes less clear when focusing on likely dynamic residues. Consequently, the gradations of dynamics observed by NMR in flexible regions of proteins are not fully captured by these computational methods.

Gradations in protein dynamics captured by experimental NMR are not well represented by AlphaFold2 models and other computational metrics.

Jose Gavalda-Garcia [1,2,†], Bhawna Dixit [1,2,3,†], Adrián Díaz [1,2], An Ghysels [3], and Wim Vranken [1,2,4,5,6,*]

[1]Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Brussels, Belgium
[2]Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium
[3]IBiTech - BioMMedA group, Ghent University, Belgium
[4]AI lab, Vrije Universiteit Brussel, Brussels, Belgium
[5]Chemistry department, Vrije Universiteit Brussel, Brussels, Belgium
[6]Biomedical sciences, Vrije Universiteit Brussel, Brussels, Belgium

[†]Authors contributed equally to this work.
[*]To whom correspondence should be addressed: wim.vranken@vub.be

February 10, 2025

Keywords: Protein dynamics, Nuclear Magnetic Resonance, AlphaFold2, Structural biology, Large-scale analysis

Abstract

The advent of accurate methods to predict the fold of proteins initiated by AlphaFold2 is rapidly changing our understanding of proteins and helping their design. However, these methods are mainly trained on protein structures determined with X-ray diffraction, where the protein is packed in crystals at often cryogenic temperatures. They can therefore only reliably cover well-folded parts of proteins that experience few, if any, conformational changes. Experimentally, solution nuclear magnetic resonance (NMR) is the experimental method of choice to gain insight into protein dynamics at near physiological conditions. Computationally, methods such as molecular dynamics (MD) and Normal Mode Analysis (NMA) allow the estimation of a protein's intrinsic flexibility based on a single protein structure. This work addresses, on a large scale, the relationships for proteins between the AlphaFold2 pLDDT metric, the observed dynamics in solution from NMR metrics, interpreted MD simulations, and the computed dynamics with NMA from single AlphaFold2 models and NMR ensembles. We observe that these metrics agree well for rigid residues that adopt a single well-defined conformation, which are clearly distinct from residues that exhibit dynamic behavior and adopt multiple conformations. This direct order/disorder categorisation is reflected in the correlations observed between the parameters, but becomes very limited when considering only the likely dynamic residues. The gradations of dynamics observed by NMR in flexible protein regions are therefore not represented by these computational approaches. Our results are interactively available for each protein from `https://bio2byte.be/af_nmr_nma/`.

1 Introduction

The advent of accurate methods to predict the fold of proteins initiated by AlphaFold2 [1] is rapidly changing our understanding of proteins [2, 3] and helping their design [4–6]. However, these methods only reliably cover well-folded parts of proteins that experience few, if any, conformational changes. Methods are now being developed that predict multiple conformations, for example by modifying the multiple sequence alignment input that captures evolutionary information [7] or by adapting the underlying deep learning model [8]. Unfortunately, reliable data on such multiple conformations that enables validation of these methods remains scarce. For example, it is estimated that up to 4% percent of proteins can change their fold, while we only have experimentally detected very few of such changes [9]. In addition, smaller conformational changes, which can be functionally important, are difficult to detect and often only captured by specific experimental NMR relaxation measurements[10]. Finally, protein dynamics, which encompass the timescales of interconversion between multiple conformations, are functionally highly relevant but rarely well-defined experimentally, with computational developments essential [11]. While computational successes are reported on individual protein cases, an extensive large scale comparison for proteins between predicted models and their computed dynamics to experimental observations of dynamics and the gradations thereof is still missing.

AlphaFold2 captures the local accuracy of the structure predictions by the predicted local distance difference test (pLDDT) [1, 12]. The pLDDT value for a residue is an estimation of the resemblance between the prediction and an experimentally determined protein structure [13]. Given that the dataset employed to train AlphaFold2 exclusively contains static protein structures [1] obtained with experimental X-ray diffraction and cryo-electron microscopy (cryo-EM), high pLDDT values therefore indicate that a protein residue is likely in a well-folded rigid state that can be represented by fixed coordinates [14]. Meanwhile, longer loops are often missing in the X-ray diffraction structures. This is illustrated by AlphaFold2's capacity to predict "disorder" versus "order" for protein residues from pLDDT values [15]. However, such a "disorder" versus "order" distinction does not capture gradations in dynamics, only its presence or absence. Indeed, cryo-EM and, typically, X-ray diffraction study proteins at cryogenic temperatures, with proteins in crystalline form in diffraction studies. These measurements do thus not represent the native dynamics and multiple conformations that proteins experience in solution at temperatures enabling life [16]. Such dynamic residues cannot be described by a static set of coordinates and often will be poorly resolved, or might even prevent the formation of crystals altogether [17]. Instead, proteins with multiple conformations, for example as observed for intrinsically disordered regions (IDRs) and proteins (IDPs), require ensemble representations of structures [18] and other descriptors for their dynamics at different timescales.

The experimental method of choice to gain insight into protein dynamics is solution nuclear magnetic resonance (NMR), which can study protein dynamics and allows structural space at near physiological conditions [17]. Dedicated NMR experiments can accurately determine protein dynamics at different timescales, often captured by the S^2 order parameter [19], but such measurements remain scarce. More readily available are NMR chemical shift values [20] stored at the BioMagResBank (BMRB). Chemical shifts can be interpreted to gain less accurate, but still useful, information on dynamics and conformation at the per-residue level. Notably, the random coil index (RCI) estimates the backbone dynamics at the residue level from a simple model that interprets experimental chemical shift values. It also provides an approximation of the backbone S^2 order parameter as the (S^2_{RCI}) value [21, 22]. Additionally, ShiftCrypt [23] presents a machine-learning based alternative by deriving a single encoded per-residue value from NMR chemical shifts, which captures a combination of conformational preference and dynamics at the protein residue level. These residue-level experimental metrics therefore allow a direct comparison of experimentally determined protein dynamics, of movements averaged when faster than ms timescales, with structure-related per-residue values such as pLDDT. The use of RCI in this context was previously illustrated by an investigation of predicted AlphaFold2 models versus NMR models calculated from experimental data. The ANSURR method (Accuracy of NMR Structures Using random coil index (RCI) and Rigidity) used a dataset of 904 human proteins [24] to compare their scaled RCI value to a local structure-based rigidity measure [25]. The NMR models with highest-scoring ANSURR scores in each ensemble, indicating a good match between in-solution observations and structure models, showed accuracy comparable with AlphaFold2, with AlphaFold2 performing significantly better in 30% of cases, particularly in relation to regions with extensive hydrogen-bond (H-bond) networks. Only in 2% of cases, the ANSURR score of the NMR structure ensemble was higher, primarily in dynamic regions [24], indicating that AlphaFold2 struggles with these regions.

Computationally, simulations of a 3D representation of the protein can also provide valuable insights on their dynamics. With molecular dynamics (MD) simulations, the protein's flexibility and conformational states can be investigated. Besides MD, Normal Mode Analysis (NMA) allows the estimation of a protein's intrinsic flexibility based on a single protein structure [26]. NMA provides information on the low-energetic motions that are accessible to the system at finite temperature. These large-amplitude motions are often related to biological function. Usually the elastic network model (ENM) approximation is applied, which models the protein as a set of C_α beads in 3D space interacting with each other within a certain cutoff range [27]. The ENM therefore successfully captures the connectivity of the protein, which is a primary factor for the protein's flexibility. Moreover, NMA is computationally orders of magnitude cheaper than a standard MD simulation, and hence it is a valuable tool to swiftly assess possible dynamics for an extended dataset of proteins, especially since the low-frequency normal modes generated by NMA have been shown to effectively capture the collective motions of proteins observed in both NMR experiments and MD simulations [28–31]. This correlation between NMA predictions and experimental observations highlights the significant influence of the backbone in describing collective dynamics, and NMA can therefore in principle be used to assess the flexibility of static AlphaFold2 models.

This work addresses, on a large scale, the relationships for proteins between the AlphaFold2 pLDDT metric, the observed dynamics in solution from NMR metrics, interpreted MD simulations and the computed dynamics with NMA from single AlphaFold2 models and NMR ensembles. We observe that these metrics agree well for rigid residues that adopt a single well-defined conformation, which are clearly distinct from residues that exhibit dynamic behavior and adopt multiple conformations. This direct order/disorder categorisation is reflected in the correlations observed between the parameters, but becomes very limited when considering only dynamic residues. Gradations of dynamics, as observed by NMR in flexible protein regions, are therefore not represented by current computational approaches to predict protein folds. Our results are interactively available for each protein from `https://bio2byte.be/af_nmr_nma/`.

2 Methods

2.1 Overview of information in three datasets

Three datasets were constructed: the S^2_{RCI} dataset, the S^2 dataset, and the Molecular Dynamics (MD) dataset (Table 1). Besides common elements, such as the the protein sequence information, the datasets differ by per-residue metrics such as S^2_{RCI} values. Per dataset, all the per-residue information for a given entry was collected in integrated Pandas data frames [32, 33] that are made available on `https://zenodo.org/doi/10.5281/zenodo.10977724`.

2.1.1 Structure models and NMR data

AlphaFold2 structures were downloaded from AlphaFold2's EBI database [13] (S^2_{RCI} dataset), when available, with the in-house software package AlphaFetcher (`https://pypi.org/project/AlphaFetcher/`). All other AlphaFold2 structures were calculated on the Vlaams Supercomputer Centrum (VSC) infrastructure (S^2 and MD datasets), with the cut-off date for the structures employed as templates the 15th of February 2021, for uniformity with the S^2_{RCI} dataset. AlphaFold3 C_α pLDDT values were calculated on the publicly available AlphaFold3 server `https://golgi.sandbox.google.com/`. Due to job number limitations on this server the S^2_{RCI} dataset was excluded. The per-residue AlphaFold2 and/or AlphaFold3 pLDDT values are integrated in the Pandas data frame [32, 33].

Chemical shift data was collected from the Biological Magnetic Resonance Data Bank (BMRB) [34], using previously described criteria [35, 36]. Briefly, only entries reported with pH between $5-7$, temperature between $293-313$K, for which chemical shift data was available for ^{1}H, ^{13}C and ^{15}N for at least half of the residues in the sequence, were selected. Entries with samples containing agents that strongly influence protein behaviour (Supplementary Table 1 in [35]) were excluded, and chemical shift re-referencing was performed with VASCO [37].

NMR structure ensembles corresponding to the BMRB sequences were collected from the Protein Data Bank (PDB) [38] only if there was a 100% sequence identity match between the BMRB and PDB sequence. Only proteins for which predicted models were available from the AlphaFold database were retained[13], with the BMRB and AlphaFold2 sequences then locally aligned. Only entries where the longest uninterrupted aligned fragment covered at least 40% of the BMRB sequence were finally retained, resulting in a final S^2_{RCI} dataset of 762 proteins with a corresponding AlphaFold2 model and an NMR ensemble. The total number of models in all 762 NMR ensembles is 14,334.

Experimental S^2 values were collected from the BMRB for a set of 52 proteins. Overlap from multiple BMRB IDs mapping to the same Uniprot accession code (Table 1) was removed with the BMRB entry with the highest number of S^2 data points retained, resulting 42 proteins in the S^2 dataset. The experimental S^2 order parameters of a protein are per-residue values and were integrated in the Pandas data frame for each of the 42 entries [32, 33].

2.1.2 Information derived from models and NMR data

S^2_{RCI} order parameters were calculated for the S^2_{RCI} dataset from the available chemical shift values using the RCI software [22] and integrated in the Pandas data frame for each entry as per-residue values [32, 33].

Secondary structure assignments for all residues using were obtained with STRIDE [39, 40] for all AlphaFold2 models and for all models in all NMR ensembles. For proteins with multiple models in the NMR ensemble where the STRIDE secondary structure of a residue was the same in all models, a "STRIDE unique" value was assigned. The majority secondary structure assignment of a residue across all models was then labelled as "STRIDE consensus". The STRIDE secondary structure assignments are per-residue values and were integrated in the Pandas data frame [32, 33].

MD simulations of single chain entries for a 100 proteins from the Constava dataset [36] were analysed using the **conformational state variability (Constava)** method [36]. Trajectories were sampled using window size 3 sampling, as recommended for balanced variability computation for time-series data (such as the MD simulations of this dataset). All calculations were performed utilizing the Constava PyPI package using the grid interpolation model `https://pypi.org/project/constava/`. The Constava values of a protein are per-residue values and were integrated in the Pandas data frame for each entry [32, 33].

2.1.3 Normal Mode Analysis

Root mean square fluctuations (RMSF) from NMA were calculated on the AlphaFold2 structures for all datasets and on the NMR ensembles for the S^2_{RCI} dataset.

The normal modes and eigenfrequencies of all AlphaFold2 and NMR ensemble models were computed with the open-source WEBnma webserver [27, 41]. Based on the normal modes with lowest eigenvalues, the atoms' fluctuations were estimated under thermal equilibrium [42]. The squared fluctuation of atom i is

$$\mathrm{fluc}_i^2 = k_{\mathrm{B}}T\sum_{k=7}^{M}\frac{v_{xi,k}^2+v_{yi,k}^2+v_{zi,k}^2}{m_i\omega_k^2} \tag{1}$$

Here, k_{B} is the Boltzmann constant, T is temperature (here 300 K), $M-6$ the number of contributing eigenvectors, m_i is the mass of the i^{th} amino acid residue, ω_k^2 is the k^{th} eigenvalue, $v_{xi,k}$ is the Cartesian x-component for C_α atom i in the corresponding k^{th} normal mode vector, and similarly for $v_{yi,k}$ and $v_{zi,k}$. The normal mode vectors v_k are mass-weighted and normalized [42]. The sum skips the 6 zero-frequency modes corresponding to global translation and rotations and only the lowest 200 non-trivial normal mode vectors ($M = 206$) were included in the sum. Based on these C_α atom fluctuations, the root mean square fluctuations $(\mathrm{RMSF})_i$ of each residue i, which is a measure of the average fluctuation or displacement of individual atoms from their mean positions, was obtained by taking the square root,

$$\mathrm{RMSF}_i = \sqrt{\mathrm{fluc}_i^2} \tag{2}$$

The normalized eigenvectors and eigenvalues from WEBnma were used to construct the RMSF profile with Eqs. 1-2.

In unfolded parts of the protein, often at the termini, the standard WEBnma calculations led to several unrealistically large RMSF values, which were removed using a pre-processing step as detailed in the Supplementary Information (Section Truncation criterion). In summary, N- and/or the C-terminal residues were truncated if these residues made fewer than 13 C_α-C_α contacts below a distance of 10 Å (computed using MDAnalysis), so removing all residues until the first residue with at least 13 contacts (Supplementary Fig. 14). Unfolded terminal residues are thus removed, while highly-connected terminal residues are not (Supplementary Fig. 15). WEBnma was re-run on these truncated structures and the per-residue RMSF computed with Eqs. 1-2 based on the C_α atoms.

NMA fluctuations were also computed with WEBnma for the 762 AlphaFold2 models in the S^2_{RCI} dataset, of which 755 were truncated, and on each model of the 762 NMR ensembles. Some models in the NMR ensembles had to be discarded, due to: limited overlap between a truncated AlphaFold model and the NMR models (10 discarded); PDB file could not be split into models (2 discarded); WEBnma error due to invalid distances (<0.278 Å) between C_α atoms for cis peptide bonds (all 174 models for 13 proteins discarded, as well as 51 models of various other proteins). The NMA based RMSF was thus calculated for all AlphaFold2 models for 762 proteins, as well as for a subset of 746 proteins of 14,069 NMR models. The WEBnma-based RMSF was successfully computed for the S^2 and MD datasets for respectively all 42 (of which 41 truncated) and 100 (of which 30 truncated) AlphaFold2 structures.

The RMSF are per-residue values and were integrated in the Pandas data frame for each protein entry [32, 33]. As a protein can have multiple NMR models, the entry may contain multiple RMSF profiles corresponding to different models in its NMR ensemble.

2.1.4 Normalisation of pLDDT ranges

The propensity of each pLDDT range towards each conformational state was normalised using the following process:

- pLDDT ranges are categorized into low (<60), mid (60-80), and high (>80).
- S is the assigned secondary structure of a residue.
- $N_{S,\mathrm{low}}$, $N_{S,\mathrm{mid}}$, and $N_{S,\mathrm{high}}$ are the number of residues with secondary structure S within the specified pLDDT range.
- $T_S = N_{S,\mathrm{low}} + N_{S,\mathrm{mid}} + N_{S,\mathrm{high}}$ is the total number of residues with secondary structure S across all pLDDT ranges.

The normalized counts for residues in secondary structure S across different pLDDT ranges (low, mid and high) are calculated as follows:

$$\text{Normalized count }_{\mathrm{S,range}} = \frac{N_{S,\mathrm{range}}}{T_S} \tag{3}$$

Following this, the relative normalized fractions are calculated using the formula:

$$\text{Relative normalized fraction }_{\mathrm{S,range}} = \frac{\text{Normalized count }_{\mathrm{S,range}}}{\sum_{\mathrm{all\ S}} \text{Normalized count }_{\mathrm{S,range}}} \tag{4}$$

which ensures the sum over fractions equals 1 within each pLDDT range. These two steps convert the raw counts into proportional measures that reflect the relative abundance of each secondary structure S within each pLDDT range.

2.2 Data analysis and plotting

A docker file containing all necessary code for the analysis and plots in this work can be found in `https://hub.docker.com/repository/docker/jgavalda/alphafold_analysis_pipeline/`. The code (available on `https://bitbucket.org/bio2byte/af2_analysis_datagen_and_plots/src/main/`) contains all necessary tools to download AlphaFold2 structures and integrate them with the different data sources. The generated AlphaFold2 and AlphaFold3, as well as the NMA calculations are not included in this pipeline to alleviate the required computing resources. The pipeline's documentation includes a GitHub link to the NMA pipeline code. The complete pipeline can be ran in approximately 2 hours with a standard desktop computer.

3 Results

This study investigates to which extent AlphaFold2 models and the associated pLDDT metric can capture protein dynamics information via three datasets: the NMR chemical shift derived S^2_{RCI}, the NMR directly measured S^2 order parameter and MD simulations providing traditional MD metrics and the conformational state variability (described in Methods 2.1 and Table 1).

3.1 Secondary structure comparison between AlphaFold2 models and NMR ensembles

NMR structure ensembles are calculated from available experimental NMR information such as NOEs from NOESY spectra, which are converted to distances used in the NMR structure calculations. NMR structure ensembles therefore capture experimental NMR information, but only to a certain degree as the data-to- model process is highly complex [43]. A comparison between predicted AlphaFold2 models and experimentally derived NMR structure ensembles is therefore a relevant first step to detect discrepancies. We focused on the STRIDE [39, 40] secondary structure assignment for AlphaFold2 models and the corresponding NMR ensembles in the S^2_{RCI} dataset (see Methods 2.1.1). STRIDE assigns secondary structures in AlphaFold2 models by evaluating H-bonds and backbone torsions, and if the predicted H-bond networks are energetically unfavorable based on backbone torsions, it does not classify them as valid hydrogen bonds, minimizing bias in secondary structure assignments [39]. For the AlphaFold2 structures, the STRIDE assignments were subdivided by pLDDT range, as high (> 80), mid ($80 > \mathrm{pLDDT} > 60$) and low (< 60) and their abundance normalized (see Methods 2.1.2) to reflect the secondary structure tendencies for each pLDDT range.

The resulting comparison between the AlphaFold2 and NMR secondary structures (Fig. 1 & Supplementary Fig. 2) shows that residues with low pLDDT values, which are not confidently predicted, have a higher tendency for coil than in the corresponding NMR ensembles, while β-sheet and helix are very similar in content and turn is underrepresented. This indicates that AlphaFold2 is not able to capture transient turn conformations that are still present in the NMR ensembles, where sufficient experimental data, such as distances or dihedral angles, must be included in the structure calculation process to detect such conformations. For the mid pLDDT values, helix is overrepresented in AlphaFold2 models, while coil is underrepresented in NMR ensembles. This discrepancy likely points towards regions that are in coil conformation but can fold upon binding to an interaction partner, evidenced by the observation of single isolated helices in many AlphaFold2 models [44]. Finally, for high pLDDT values, β-sheet is overrepresented and coil underrepresented in the AlphaFold2 models compared to NMR ensembles. A likely explanation here is that residue interactions in β-sheets are often between residues far apart in the sequence, with information from NMR experiments therefore often sparser and more likely to be insufficient to define such β-sheets with enough precision in the calculated structure models. Secondary structure classification programs such as STRIDE are then less likely to unambiguously assign them, even if they are present in solution.

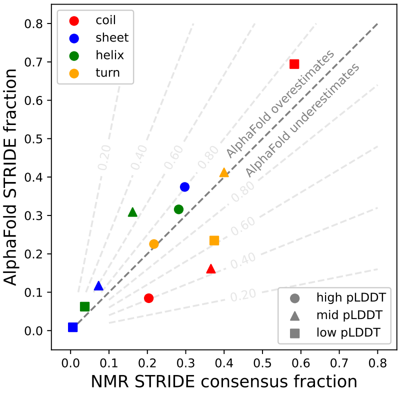


Figure 1: **Comparison of STRIDE secondary structures fractions in AlphaFold2 models and NMR ensembles.** Values are normalised to highlight the differences in secondary structure content (colour coded) captured per pLDDT class (coded by mark). Values above the dotted line indicate a higher presence in AlphaFold2 models, values below the dotted line indicate a higher presence in NMR ensembles. Based on S^2_{RCI} dataset.

Comparing the STRIDE secondary structure assignments of the S^2_{RCI} dataset at the residue level (Fig. 2) reveals that discrepancies in secondary structure assignment increase as the pLDDT decreases. Notably, coil and turn assignments are often swapped at low pLDDT (see also supplementary Fig. 9). The presence or absence of a unique STRIDE assignment in an NMR ensemble is also important, with unique assignments corresponding to higher pLDDT values (Supplementary Figs. 9, 10 and 11). Transient turn conformations captured by NMR ensembles are often not predicted by AlphaFold2, whereas AlphaFold2 predicts turns where the NMR ensembles do not show any (Fig. 2). Turn conformations are essentially locally determined, with often insufficient experimental NMR information available to precisely and consistently define them in an NMR ensemble. The required combination of co-evolutionary and encoded structure information might on the other hand also not be present for a confident turn prediction in AlphaFold2 models.

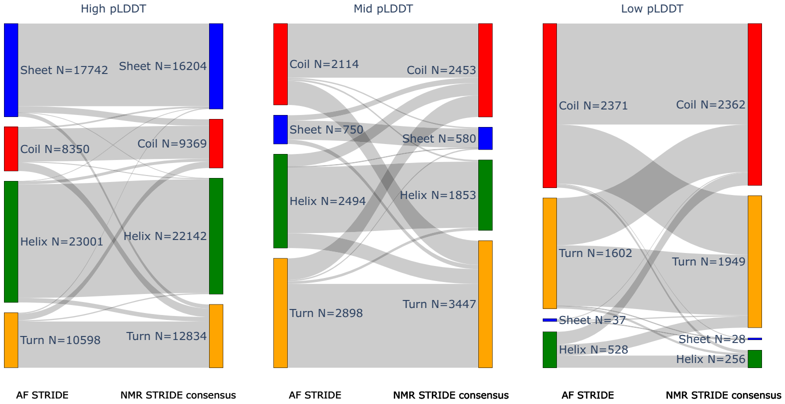


Figure 2: **Per-residue comparison of STRIDE secondary structure assignment in AlphaFold2 models and NMR ensembles.** For each pLDDT class (left, middle, right) the correspondence between the STRIDE secondary structure assignment for the AlphaFold2 (coloured bars on left side of each plot) and NMR ensembles (right side of each plot) is shown. Gray bars indicate the similarities and changes at the residue level between the STRIDE assignments, with no normalisation applied. Based on S^2_{RCI} dataset.

3.2 AlphaFold2 pLDDT vs NMR experimental order parameters

Information on the dynamics of individual residues for proteins in solution can be directly captured by the NMR S^2 order parameter. The RCI method interprets NMR chemical shifts to estimate the S^2_{RCI}, a proxy for the order parameter. The RCI method uses a simple linear model to compare observed chemical shift values to reference 'random coil' chemical shift values, which indicate many highly dynamic conformations. We here compare these measurements, which capture in-solution dynamics, to AlphaFold2 structures, which provide a static snapshot of the protein fold. Nevertheless, AlphaFold2's pLDDT metric is a good predictor for disorder [15], with a clear relationship between the pLDDT and amino acid order/disorder preference (Supplementary Fig. 1). So, we here explore in detail to which extent pLDDT values also capture the degree of in solution dynamics encapsulated by S^2_{RCI} values.

3.2.1 Relation between chemical shift derived S^2_{RCI} backbone dynamics and pLDDT values

A large-scale comparison between pLDDT and S^2_{RCI} values (Fig. 3 A & B) reveals an unambiguous tendency for residues confidently predicted by AlphaFold2 (high pLDDT) to be also highly ordered in solution (high S^2_{RCI}, above 0.80). Conversely, residues not confidently predicted by AlphaFold2 (low pLDDT values) tend to be highly dynamic in solution (low S^2_{RCI}, below 0.80), with a wide spread of S^2_{RCI} values. Mid pLDDT values display ambiguous behaviour, with a peak of higher S^2_{RCI} values complemented by a long tail of lower S^2_{RCI} values that tailor off towards low S^2_{RCI}. These results show that, as expected, well folded protein regions that are rigid in solution are accurately predicted by AlphaFold2, with the secondary structure assignments of experimental models consistent and the matching AlphaFold2 models (see section 3.1). Residues predicted with medium pLDDT cover residues that are rigid to highly flexible in solution, indicating that AlphaFold2 is not able to distinguish the extent of dynamics present. However, the associated conformational ambiguity of such residues is, to some degree, captured by the generally lower pLDDT values. Finally, low pLDDT values unambiguously correspond to residues that are highly dynamic in solution. The Pearson correlation between pLDDT and S^2_{RCI} on the full dataset is relatively high (Table 2, top row), but this is driven by very high pLDDT values with high S^2_{RCI} and very low pLDDT values with low S^2_{RCI}. When considering each of the pLDDT categories, only very limited correlations are observed (Table 2, second grouped rows). This is even more pronounced when looking at S^2_{RCI} stratified classes, where almost no correlation remains (Table 2, fifth grouped rows). The S^2_{RCI} values are here subgrouped as flexible (below 0.70), rigid (above 0.80), and context-dependent ambiguous (0.70-0.80), as defined previously [45]. Overall, these results indicate that while AlphaFold2 picks up on extensive conformational variability, it does not capture the amount of movement between such multiple conformations. Mann-Whitney two-sided U tests confirm significant differences between all subsets with a p-value < 0.001 (Supplementary Table 3).

3.2.2 Relation between experimental S^2 order parameter and pLDDT values

Only few S^2 order parameters are accessible from the BMRB, as opposed to the S^2_{RCI} values, which are calculated from readily available chemical shift data. There are fundamental differences between the S^2_{RCI} and S^2 values, with the former less accurate and capturing movements on longer timescales (up to low ms), while the latter are highly accurate but only capture very fast movements (on the ps-ns timescale). Nevertheless, the relationship between the S^2 order parameters and the pLDDT is similar to the one observed for the S^2_{RCI} (Pearson = 0.51, p-value 4.8×10^{-289}), though the smaller dataset size results in noisier distributions (Fig. 3 C & D). The mid pLDDT value range is more skewed towards S^2 values higher than 0.8. This could be due to bias in this limited dataset, or could indicate that the ambiguous behavior described in section 3.2.1 is more relevant for slower (towards ms) movements, rather than very fast ones, as previously described in an analysis of S^2_{RCI} versus S^2 values [35]. Very equivalent behavior is observed for AlphaFold3's C_α pLDDT (Supplementary Fig. 8). Although there are significant changes in the AlphaFold2-AlphaFold3 pLDDT for individual residues within this dataset, the overall statistics are very similar. Here also Mann-Whitney two-sided U tests confirm significant differences between all subsets with a p-value < 0.001 (Supplementary Table 3)

3.2.3 Relation between chemical-shift derived δ2D secondary structure populations and pLDDT values

The δ2D populations [46] are an estimation of the per-residue secondary structure occupancy, as fraction of 1.0, derived from NMR chemical shift data of the protein in solution. This method works similarly to the RCI method, except that δ2D interprets chemical shift values with a model that relies on chemical shift values typically observed for secondary structure elements. The available δ2D populations in the S^2_{RCI} dataset were compared to the pLDDT values, similar to sections 3.2.1 and 3.2.2 (Supplementary Fig. 3 & Supplementary Table 2). The residues that according to δ2D adopt dominant helix and sheet populations are also typically predicted with high confidence by AlphaFold2. This is expected as both helix and sheet are rigid H-bond stabilized secondary structures. Note that these secondary structures are mutually exclusive, with high helix population corresponding to low sheet population, which accounts for the high proportions of zero occupancy in Supplementary Fig. 3. In contrast, coil and polyproline II (PPII) populations, which feature multiple conformations and variable H-bonding, are predicted with low confidence by AlphaFold2 (low pLDDT). These results are therefore in line with the previous observations from Fig. 2 & Supplementary Fig. 2.

3.2.4 Relation between ShiftCrypt chemical shift interpretation and pLDDT values

ShiftCrypt [23] is a machine learning based method that encodes chemical shift values in values between 0 and 1, encompassing a combination of conformation and dynamics from rigid helix (towards 0.0) and rigid sheet (towards 1.0), with values around 0.5 indicating dynamic behavior and multiple conformations. The method was employed on the chemical shift data in the S^2_{RCI} dataset and compared to pLDDT and stratified ranges (Fig. 4 panels A & B). High pLDDT values are associated with high and low ShiftCrypt values (indicating rigid sheet and helix), while low pLDDT ranges are almost exclusively in the 0.4-0.6 ShiftCrypt range (indicating multiple conformations in dynamic exchange). These results therefore confirm those of the other chemical shift based methods previously discussed.

Previously, the ShiftCrypt values were combined with the STRIDE [39, 40] secondary structure assignment of the NMR ensembles to define 6 conformational states *et al.* [36]: core helix, core sheet, surrounding helix, surrounding sheet, coil, and turn. These conformational states encompass both dynamics and conformation, with the “core” states highly rigid, the “surrounding” states more

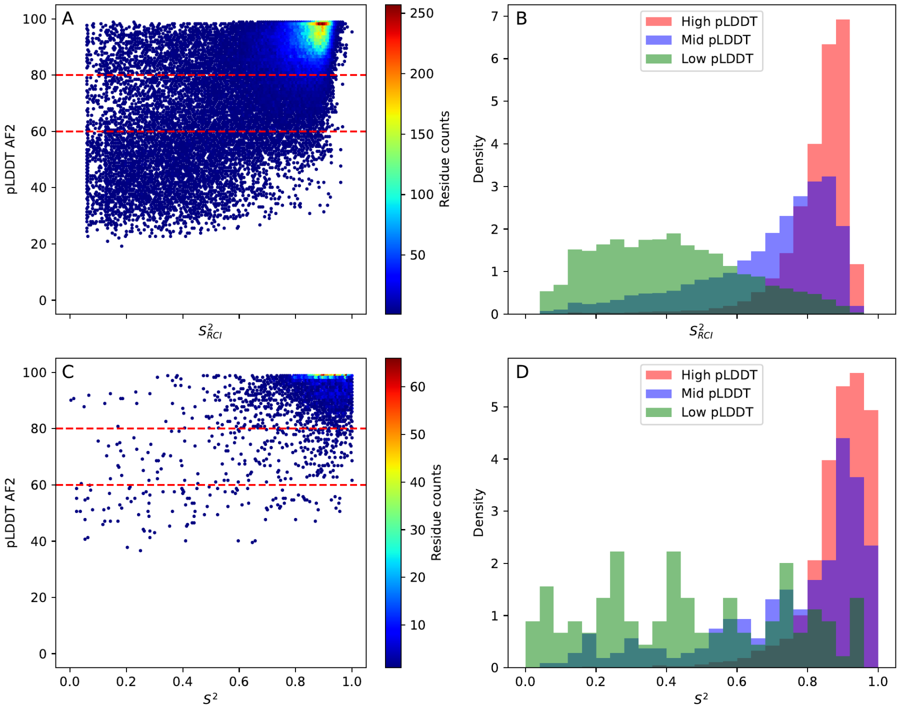


Figure 3: **Comparison of pLDDT vs $S^2_{\mathbf{RCI}}$ and S^2.** (A and C) Per-residue hexagonal binning plot of pLDDT versus order parameter values, with dotted red lines indicating cutoffs for classification of pLDDT values into three groups. The colour-coding indicates the residue count. Three groups for pLDDT-S^2_{RCI} plot (A): high pLDDT (N=62,624), mid pLDDT (N=8,859), low pLDDT (N=5,223). Three groups for pLDDT-S^2 plot (C): high pLDDT (N=4,103), mid pLDDT, (N=267), low pLDDT (N=112). (B and D) Histograms of order parameter value distributions for each pLDDT range.

dynamic, and the "coil" and "turn" states highly dynamic. Essentially, each conformational state represents a low-energy macro-state that a residue can adopt in solution, defined from NMR data as probability density functions in dihedral space. When comparing the conformational state of residues with their corresponding pLDDT values (Fig. 4 C & D), it is again clear that low pLDDT values correspond to turn and especially coil states, with likely multiple conformations in dynamic exchange present. Due to the skew of the overall pLDTT distribution (Fig. 4 E), high pLDDT values dominate for all conformational states, though only the "coil" and "turn" states feature significant densities below pLDDT of 80%.

After normalising the densities of each conformational state to highlight their relative pLDDT tendencies (Fig. 4 F), both sheet states feature the largest relative densities of high pLDDT values, with few values lower than 80 and very few lower than 60. This indicates that if sheet is present in solution, then it is are confidently predicted by AlphaFold2. Since these conformations feature non-local structural stabilisation by hydrogen bonds with other β-strands, this restricts their motions and fixes their position, as well as leading to strong co-evolutionary signals [47–49], so facilitating structural predictions. Helix conformations also typically have high pLDDT values, though shifted towards lower ones compared to sheet. Helix conformations are stabilised by a regular hydrogen bond network defined by local interactions between amino acids, making them easier to predict and design than sheet [50]. The strong local interactions imply that even when helices are more dynamic and partially unfold, they are easier to reform [51–53], with co-evolutionary signals also more difficult to pick up [54]. These factors can here explain the observed differences with sheet. Overall, the "core" helix and sheet correspond to higher pLDDT than the "surrounding" states, highlighting that the

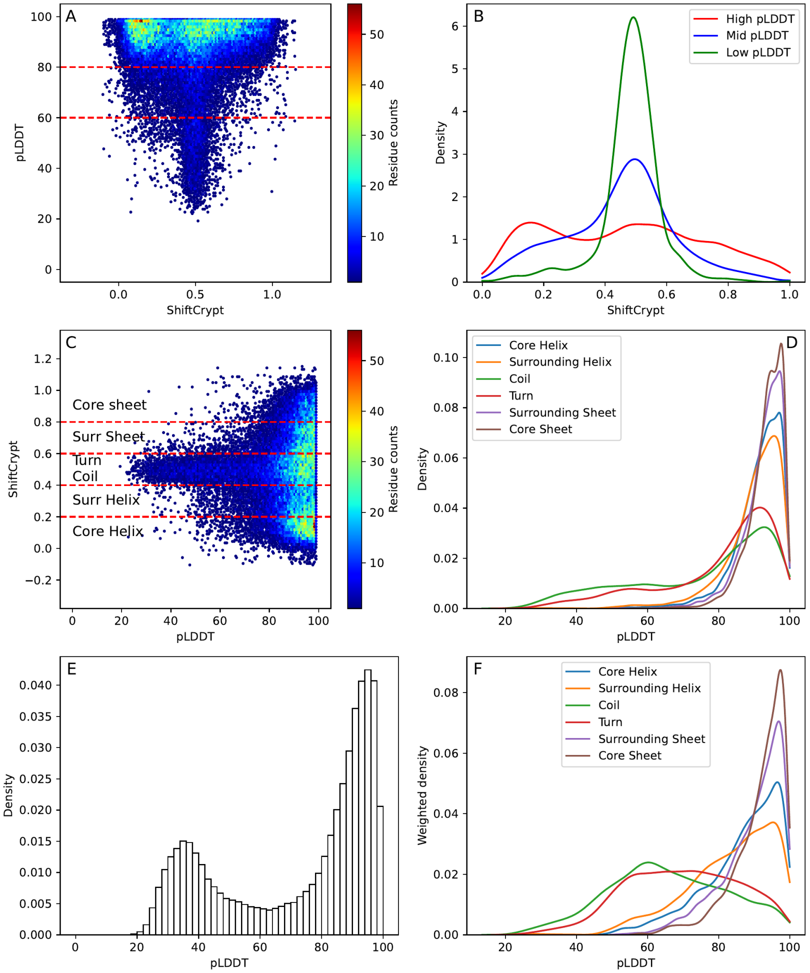


Figure 4: **Analysis of pLDDT and ShiftCrypt values.** A & B: The residues with valid ShiftCrypt values were compared with their corresponding pLDDT and stratified according to their pLDDT values in high (N=41,293), mid (N=5,480) and low (N=2,938). C & D: The residues of the comparison in panels A and B were divided with the same criteria to determine the Constava conformational states [36]: core helix (N=8,122), surrounding helix (N=5,698), coil (N=5,236), turn (N=5,854), surrounding sheet (N=4,947), core sheet (N=3,563). E & F: The distributions in panel D were weighted by the inverse of the overall pLDDT density in the dataset, resulting in a weighted density of pLDDT values for each conformational state.

pLDDT here picks up the increased likelihood of multiple conformations. Finally, the coil and turn states are quite evenly spread over the whole pLDDT range, indicating that these residues, which are likely to adopt multiple conformations, are still sometimes confidently predicted. Possible reasons for this are strong co-evolutionary signals or the presence of similar single defined conformations in the AlphaFold2 training set.

Intrinsically disordered regions that fold upon binding are excellent examples of cases where AlphaFold2 might confidently predict structure, but the ShiftCrypt values indicate multiple conformations. Since the transformer-based model architecture and large training dataset of AlphaFold2 captures intrinsic properties of protein folds and the role of amino acids, such as a residue's location (*e.g.* hydrophobic amino acids would favor protein core location), this consequently affects in the prediction their degrees of freedom and conformation. As all binders were removed in the pre-processing of AlphaFold2's training dataset to obtain a monomer-only set, regions of proteins that fold upon binding are particularly affected, since they only acquire a fold in their bound form [55]. This is then what AlphaFold2 tends to predict [56], so enabling the placement of for example co-factors in AlphaFold2-predicted structures [57]. Such regions should therefore show conflicting, more dynamic behaviours from their NMR-derived metrics than what is indicated by the pLDDT. Indeed, Fig. 4 shows a considerable number of residues with high pLDDT and mid ShiftCrypt values, indicating dynamic resides that were confidentlypredicted by AlphaFold2.

To further investigate these occurrences, all protein regions of at least 15 consecutive residues with pLLDT >80 and $0.4 <$ShiftCrypt <0.6 were identified. As example, we selected the mouse Cytohesin-3 or Grp1 (Uniprot accession code O08967), which features self-inhibition in the absence of a ligand, and is only active in the presence of Ins(1,3,4,5)P_4 or an analogous molecule [58]. Fig. 5 A shows the predicted AlphaFold2 structure with a loop featuring high pLDDT that is unstructured according to ShiftCrypt. Fig. 5 B shows the experimentally determined structure, which illustrates that this loop is located where Ins(1,3,4,5)P_4 binds the PH domain of Grp1. All publicly available experimental structures of this protein or domain include Ins(1,3,4,5)P_4 or a fragment of it in their structure (PDB IDs: 1FGY, 1FGZ, 1FHW, 1FHX, 1U2B, 2R09, 2R0D, 6BBP & 6BBQ), all published before April 2018 and so part of AlphaFold2's training set. In contrast, the ShiftCrypt value was calculated from chemical shift data (BMRB ID 15669), which did not include any ligand, with the loop unbound and accessible to solvent (Supplementary Fig. 5). The dynamical behavior is confirmed by low S^2_{RCI} (data in supplementary data frames `https://zenodo.org/doi/10.5281/zenodo.10977724`) and highlights the limitations of even highly-confident predictions to capture ambiguous, context-dependent behavior of protein regions.

3.3 Low pLDDT values are almost exclusively found in regions with high conformational state variability in molecular dynamics trajectories.

The propensity for each conformational state (*i.e.* the degree of preference to adopt a conformational state) can also be calculated for each residue from the dihedral backbone angles it adopts an ensemble of structures, such as from a MD trajectory[36]. The conformational state variability metric then indicates how likely it is for a residue to adopt multiple conformational states. For example, a fully disordered residue which samples all coil dihedral probabilities during an MD simulation could feature near-0 conformational state variability, as it exclusively adopts a highly dynamic coil state. If it would intermittently adopt helix conformation and switch back to coil, its conformational state variability would in contrast be higher. We calculated this conformational state variability for all residues of all proteins in the MD dataset (Table 1), which were again divided in pLDDT ranges. The distributions of the conformational state variability for each range (Fig. 6 & Supplementary Fig. 4) shows that residues with high pLDDT have, as expected, low conformational state variability as they adopt typically core helix or sheet conformational states. Residues with low pLDDT, on the other hand, have high conformational state variability, confirming that such residues generally have the ability to exist in diverse conformational states and move between them. Mann-Whitney two-sided U test yielded a p-value < 0.001 between all distributions (Supplementary Table 3). The relatively small size of the dataset enabled us to confirm this relationship for the C_α pLDDT of AlphaFold3 models (Supplementary Figs. 6 & 7 & Supplementary Table 1).

3.4 AlphaFold2 pLDDT vs NMA fluctuations of AlphaFold2 models

A common per-residue structure based metric for protein flexibility is the root mean square fluctuations (RMSF). Using 200 normal modes from the NMA as obtained with the WEBnma tool, the RMSF were computed for the AlphaFold2 structures in the S^2_{RCI} dataset based on the C_α positions (see Methods 2.1.2). We explored how well the pLDDT of the AlphaFold2 model captures fluctuations computed from NMA.

Using the standard NMA, as shown in Supplementary Fig. 12, some residues may exhibit extremely high RMSF up to 185 Å, primarily in flexible N- and/or C-termini (Supplementary Table 4), but even in rigid ones. Indeed, when another segment of the protein experiences significant motion, NMA can propagate this motion throughout the entire protein, even if those regions are densely packed and conformationally stable [26]. To avoid this artifical shift of RMSF values to higher values, we truncated such loose termini and recalculated NMA and RMSF (see Methods 2.1.2). These

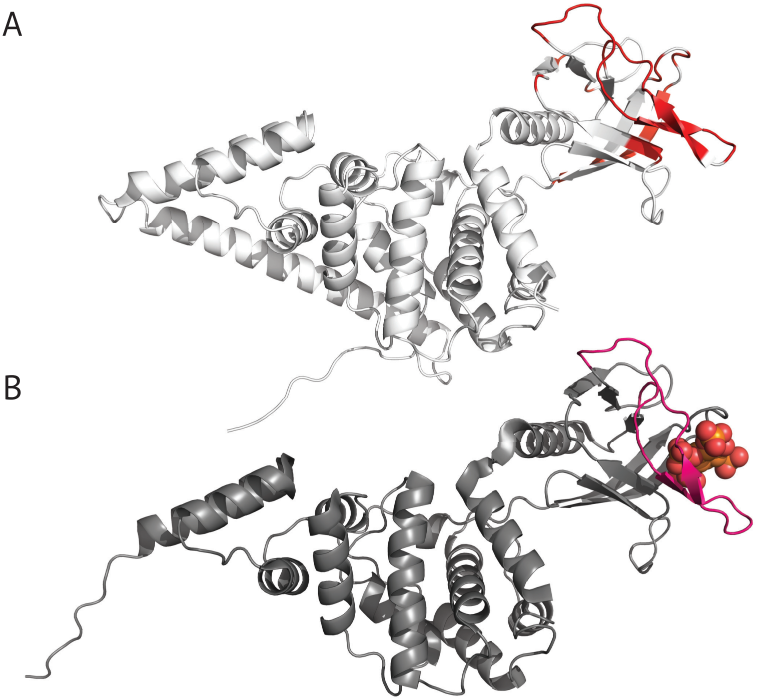


Figure 5: **Example of long region with conflicting pLDDT and ShiftCrypt values.** Above: High pLDDT with mid ShiftCrypt values. Below: Experimental structure (PDB: 2R0D) with Ins(1,3,4,5)P_4 bound. Highlighted is the loop which features a long continuous segment of high pLDDT with mid ShiftCrypt values. All other experimental structures available in PDB also featured the presence of Ins(1,3,4,5)P_4 or a fragment of it (not shown).

truncated models had considerably fewer unrealistically high RMSF values (Supplementary Figs. 12 and 13) and were used for the further analysis in this paper.

To assess the influence of secondary structure elements on RMSF, residues were grouped according to their pLDDT class and STRIDE assignment, i.e. coil, strand, α-helix, turn, 3_{10}-helix, or bridge. The distributions of the RMSF values are reported in Supplementary Table 5, with illustrative two-dimensional histograms of the RMSF and pLDDT for coil and α-helix residues in Fig. 7A and 7B, respectively. The correlations between RMSF and pLDDT per secondary structure element (Supplementary Table 7) are very small but significant (p-value < 0.05) except for strand residues with mid pLDDT, 3_{10}-helix residues with mid pLDDT, and bridge residues in low and mid pLDDT regions.

For residues with low pLDDT, coil residues show highest mean RMSF of 5.65 ± 4.43 Å compared to other secondary structure elements, indicating that coils have the highest flexibility of all secondary structures, as expected [59]. In contrast, the mean RMSF values for the other secondary structures vary, ranging from 1.89 Å for strands and bridges to 2.16 Å for turns (Supplementary Table 5, Supplementary Fig. 16).

For residues with mid pLDDT, coil, α-helix, and 3_{10}-helix residues exhibit a slightly higher mean RMSF of $1.84-1.89$ Å compared to the other secondary structure elements which fall within a similar

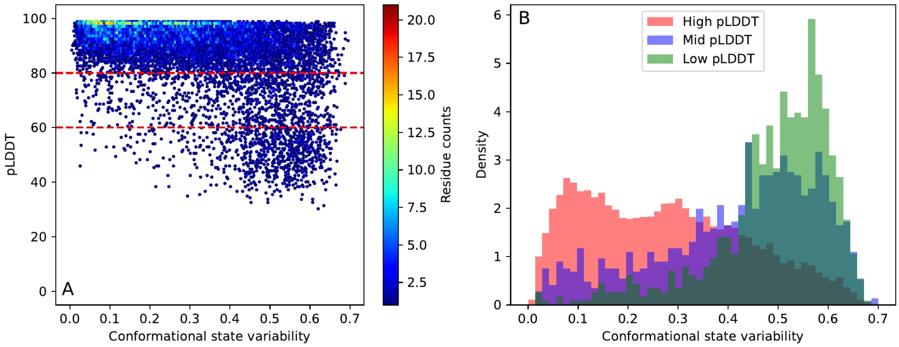


Figure 6: **pLDDT vs conformational state variability.** A) Per-residue hexagonal binning of pLDDT versus the conformational state variability, with high (N=9,523), medium (N=1,038) and low pLDDT (N=809) pLDDT regions indicated by red dotted lines. B) pLDDT-stratified distributions for each of these classes. The distributions for each conformational state propensity can be found in Supplementary Fig. 4 & Supplementary Table 1.

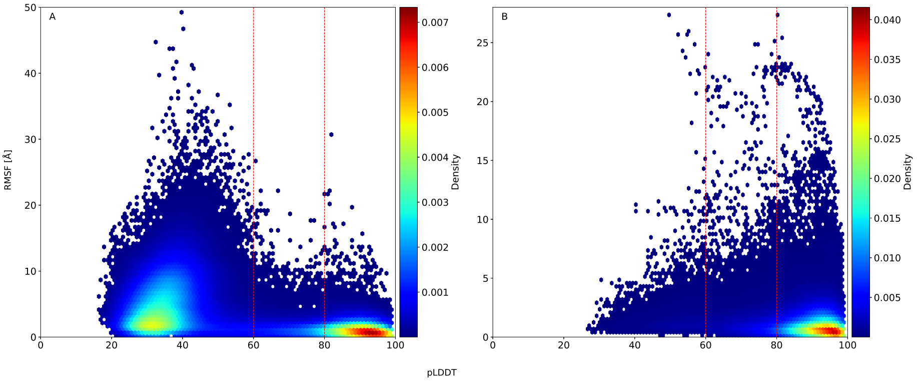


Figure 7: **pLDDT vs RMSF of AlphaFold2 models.** Histogram of pLDDT vs NMA fluctuations of each amino acid, visualized with a Gaussian kernel estimator. Data from S^2_{RCI} dataset. Subplots shown for secondary structures A) coil (N=105,172), B) α-helix (N=109,639), where N represents number of amino acid residues. The red vertical lines divide the dataset into high, mid, and low pLDDT regions.

range, ranging from 1.34 Å to 1.74 Å (Supplementary Table 5). The difference in mean RMSF of coil residues compared to other secondary structure elements is here minimal. As expected, mid pLDDT residues have a reduced mean RMSF compared to low pLDDT residues, independent of the secondary structure element, pointing to reduced flexibility.

Lastly, for residues with high pLDDT, α-helix residues show the highest mean RMSF of 1.38 Å, closely followed by the other secondary structure elements led by coil in the range 1.11-1.33 Å (Supplementary Table 5). Moreover, the mean RMSF is lowest for high pLDDT residues, in line with their higher expected rigidity.

Overall, at lower pLDDT values, coil residues are most flexible, while at higher pLDDT values both coil and α-helical residues are most flexible. The mean RMSF decreases when the pLDDT increases, irrespective of the chosen secondary structural element, with weak negative Pearson correlations (Supplementary Table 7). The overall Pearson correlation coefficient between RMSF and pLDDT of all 338,301 residues, irrespective of secondary structure, is -0.50 (p-value≈ 0) (Table 2). The elastic network model (ENM) underlying NMA models intramolecular interactions as springs between C_α atoms, and therefore lower RMSF will correspond in general to regions with a higher spatial density of C_α atoms, as expected for high pLDDT values, which tend to correspond to well-folded protein regions with tight residue packing [1].

However, this correlation is again driven by most pLDDT values being either high (>80) or low (<40), with correspondingly lower and higher RMSF values. While there is a weak overall negative Pearson correlation, this is not present when considering the different pLDDT or S^2_{RCI} subgroups (Table 2, third and sixth grouped rows). Within secondary structure elements this lack of correlation holds. For example for coils, the mid pLDDT (-0.17, p-value 8.98×10^{-55}) and high pLDDT (-0.16, p-value 1.90×10^{-157}) correlations are slightly negative (Supplementary Table 7), but for low pLDDT regions there is even a weak positive correlation (0.16, p-value ≈ 0.00). Upon visual examination of structures exhibiting these positive correlations, we mostly identified large proteins consisting of long, extended disordered coils connected by small segments of ordered regions. One possible explanation is that such regions might exhibit higher pLDDT than expected, compared to fully disordered regions, through their intermittent association with short folded regions with high pLDDT. Hence, while pLDDT is in general a reliable indicator of the RMSF (with overall Pearson correlation coefficient -0.50 with p-value ≈ 0), this is driven by the difference between very high and low pLDDT values and does not capture the intermediate flexibility gradations within each subgroups of low, mid, or high pLDDT.

3.5 NMA fluctuations vs S^2_{RCI}

The next question is whether the RMSF of the AlphaFold2 models can recapture the experimental S^2_{RCI} values that estimate backbone dynamics (Supplementary Table 6, Supplementary Fig. 17). This would be indicated by a negative correlation between the NMA RMSF and S^2_{RCI} [28, 30], as higher values of RMSF indicate higher flexibility while higher values of S^2_{RCI} (close to 1) indicate rigid residues.

3.5.1 Relationship of S^2_{RCI} with NMA fluctuations on AlphaFold2 models

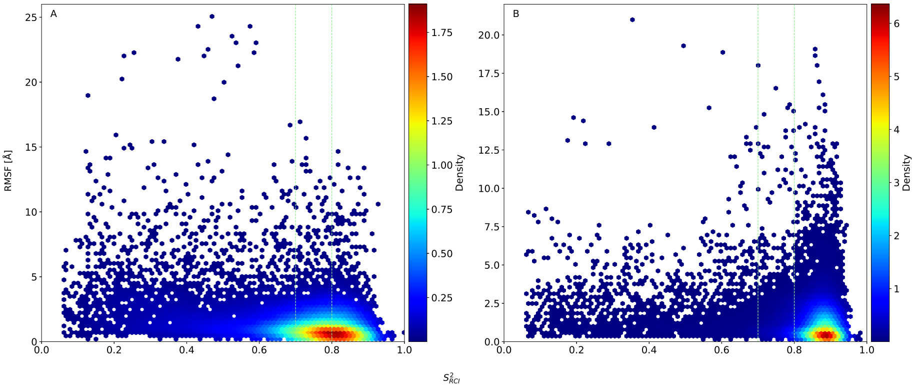


Figure 8: **RMSF vs S^2_{RCI}.** RMSF values versus S^2_{RCI} value of each amino acid, visualized with a Gaussian kernel estimator for truncated S^2_{RCI} dataset. One subplot for each secondary structure element: A) coil (N=11,634), B) α-helix (N=25,861), where N represents number of amino acid residues. The green vertical lines divide the dataset into flexible ($S^2_{\mathrm{RCI}} < 0.70$), ambiguous ($0.70 \leq S^2_{\mathrm{RCI}} < 0.80$), and rigid ($0.80 \leq S^2_{\mathrm{RCI}}$) regions.

The overall Pearson correlation between RMSF and S^2_{RCI} is indeed negative but very weak (Table 2), as illustrated for coil and α-helix residues (Figure 8). The negative correlations increase when considering only coil residues (Pearson = -0.32, p-value = 1.06×10^{-278}, see also Supplementary Table 8. Overall, NMA on AlphaFold2 models does therefore not capture the experimental S^2_{RCI} trends. The NMA fluctuations of the AlphaFold2 models are expected to be dominantly determined by the packing of the protein, which determines the structural flexibility to some extent. However, the S^2_{RCI} values indicate that this is not sufficient to capture the behavior of the protein in solution at higher temperatures. In fact, the pLDDT values are better correlated than the NMA RMSF (Table 2).

3.5.2 Relationship of S^2 with NMA fluctuations on NMR models

The NMA RMSF was also calculated for the models of the NMR structure ensembles in this dataset (Methods 2.1.2), to investigate how the S^2_{RCI} order parameter correlates with the RMSF of these NMR models. These are significantly different from AlphaFold2 structures, generated by a machine learning prediction trained on mostly crystallographic data at lower temperatures, while an NMR structure ensemble is typically generated by successive simulated annealing MD simulations using a force field that incorporates restraints based on experimental NMR information for the protein in solution. AlphaFold2 also predicts a single protein structure state, whereas NMR ensembles are sets of structures that each encompass the experimental NMR information as best as possible. The NMR ensemble of a protein thus likely encompasses more of the protein dynamics.

Two Pearson correlation coefficients were calculated for each protein k; a single ρ_k^{AF2} between the S^2_{RCI} and the RMSF of its AlphaFold2 model, for each protein k individually ($k = 1, \ldots, 746$). and multiple $\rho_{k,m}^{\mathrm{NMR}}$ between the S^2_{RCI} and the RMSF of each model m in its NMR ensemble. For the total S^2_{RCI} dataset, we so obtained 746 Pearson correlation coefficients for AlphFold2 models and 14,069 for individual NMR models (Supplementary Fig. 18). Each $\rho_{k,m}^{\mathrm{NMR}}$ is then paired with the corresponding ρ_k^{AF2}, for each NMR model. The procedure thus results in 14,069 AlphaFold2-NMR pairs of models and their Pearson correlation coefficient between RMSF and S^2_{RCI} order parameters.

A non-parametric Wilcoxon signed-rank test on the distributions of these correlation coefficients indicates a significant difference between the AlphaFold2 and NMR model correlation coefficients (p-value of 0.0001). The NMA fluctuations of the NMR models exhibit a stronger negative correlation with S^2_{RCI} than the AlphaFold2 models. Specifically, among these pairs, 71.22% exhibited a stronger negative correlation for the NMR models compared to the AlphaFold2 model, and only 28.78% showed a weaker negative correlation. This indicates that NMA can better capture the actual dynamics of proteins in solution with multiple input models that better reflect the uncertainty in conformation, compared to NMA on a single AlphaFold2 structure model.

The different correlation coefficients must derive from a difference in the 3D geometry of these models, with the RMSF calculated from NMA on C_α atom positions. For overlapping sequence regions between the AlphaFold2 sequence and NMR model sequence, the STRIDE secondary structure assignments can be: fully identical, fully conflicting, or ambiguous (given that multiple models are present in each NMR ensemble) (Supplementary Fig. 24). The percentage of conflicting residues in AlphaFold2-NMR pairs can then be calculated as the ratio between the number of conflicting STRIDE assignments and the total number of residues that overlap between AlphaFold2 and NMR models. This percentage ranges from 0.86% to 88% for 14,006 out of 14,069 pairs. However, the number of conflicting residues does not influence the correlation between S^2_{RCI} and RMSF (Supplementary Fig. 26). In addition, the size difference between AlphaFold2 and NMR models did not impact the flexibility profiles computed from WEBnma (Supplementary Table 9).

Even if conflicting secondary structure assignments do not overall directly impact the correlation between S^2_{RCI} and RMSF (Supplementary Fig. 25), this may not be the case for individual proteins, where conflicting residues within functionally relevant regions may significantly influence RMSF, especially in flexible regions. Indeed, where AlphaFold2 models show a positive correlation between S^2_{RCI} and RMSF, conflicts in secondary structure predictions significantly impact this relationship. This discrepancy often occurs in regions where NMR identifies coils and turns while AlphaFold2 predicts different secondary structures. Interestingly, high AlphaFold2 confidence levels do not mitigate the effect of these conflicts on RMSF; regions with high, mid, and low pLDDT scores exhibit similar impacts on RMSF and correlation (Supplementary Fig. 21).

An example illustrating this diversity is P0AFW0-2LCL, which showcases significant secondary structure differences between the NMR and AlphaFold2 models (88% of its overlapping AlphaFold2-NMR sequence residues) (Supplementary Fig. 23). The Pearson correlation coefficients are -0.55 for the AlphaFold2 model and -0.74 to -0.86 for the NMR models, with the conflicting residues occurring in low, mid, and high pLDDT regions. Another example is Q922K9-2D8J, where AlphaFold2 showed a stronger negative correlation than the NMR models despite conflicting secondary structure (Supplementary Fig. 20), with closer examination confirming that the effect varies from one protein to another (Supplementary Figs. 19, 21, and 22), with no discernible general trend based solely on secondary structures. Instead, the performance of AlphaFold2 and NMR models in capturing the relationship between S^2_{RCI} and RMSF appears to be very dependent on the specific characteristics and localization of conflicting residues within each protein.

In conclusion, the NMA RMSF for NMR models outperforms AlphaFold2 models in the majority of cases in reproducing S^2_{RCI} values. These results emphasize the importance of 1) validating the potential secondary structure conflicts between AlphaFold2 and NMR models, even in regions where AlphaFold2 shows high confidence, and 2) leveraging the relationship between S^2_{RCI} and RMSF to serve as a guide for refining these conflicts and improving the accuracy of AlphaFold2 models in capturing true solution dynamics.

3.6 Interactive analysis of entries

The entries for all datasets described in this work were processed to build a per-protein interactive analysis and made available on `https://bio2byte.be/af_nmr_nma/`. This resource offers a focused analysis of our entries, notably offering a dynamic mapping of the 3D structures and a selection of biophysical metrics, plotted along the amino acid sequence. Both the structure as well as the source data for any given protein can also be downloaded from their entry view, allowing users to visualize their other biophysical metrics, or to run further analysis on the datasets.

4 Discussion

The large-scale study we present here investigates how much in-solution protein dynamics are captured by the AlphaFold2 pLDDT metric and by NMA interpretation of the models it predicts. In general, residues with high pLDDT values are situated in well-folded, stable, rigid regions of proteins, whereas residues with low pLDDT are likely disordered and highly dynamic, confirming previous studies[15, 60, 61]. However, our results give a more complex and nuanced view of this relationship, linked to two key notions at the amino acid residue level. The first notion is "how many distinct conformations can a residue adopt?", which is primarily thermodynamically determined. This depends on the low energy regions present in the complex energy landscape of the full protein, as well as on how this landscape changes with the overall conformation and the environment of that protein (e.g. folding upon binding due to the presence of a binding partner). The second notion is "how dynamic is a residue?", which is kinetically determined, in other words by the height of the energy barriers between the low energy conformations. This will determine *how often* a residue can move between multiple accessible low energy conformations. Although these concepts are tightly linked, they are distinct from each other; a residue that adopts two distinct low energy conformations could move between them very quickly, very slowly, or anything in between, depending on the energy barrier between them. The observations from NMR that we compare provide a macroscopic view of this behavior as an average of up to low ms timescales movements over the billions of molecular copies of the protein in solution, each with their individual behavior.

With this distinction in mind, the conflicts between the secondary structure of residues as observed in AlphaFold2 models versus NMR ensembles (Figs. 1 and 2) show that residues confidently predicted as helix and sheet are often designated as coil in the NMR ensemble, whereas low confidence coil residues are often observed in turn conformation in the NMR ensembles. Overall, residues that have conflicts in secondary structure assignment tend to have lower pLDDT values, as well being more dynamic in solution, with lower S^2_{RCI} values and ShiftCrypt values between 0.4 and 0.6. (Supplementary Fig. 9). If we assume that such conflicts indicate that such residues have a higher likelihood of adopting multiple conformations, lower AlphaFold2 pLDDT values indeed indicate a higher probability for multiple conformations.

Another indicator of the presence of multiple conformations is available from NMR ensembles, in case a residue is assigned multiple secondary structure states in its models. This indicates that the experimental NMR data used in the structure calculation was incompatible and/or insufficient to uniquely define the residue's conformation. The overall trends are here are similar, with residues with non-unique secondary structures typically having lower pLDDT, lower S^2_{RCI} values and ShiftCrypt values between 0.4-0.6 (Supplementary Fig. 11). However, when in addition distinguishing by AlphaFold2 conflicting residues, a more complex picture emerges. For helix and sheet, unique secondary structure assignments that match the AlphaFold2 model assignment are solidly in the ShiftCrypt rigid helix (values towards 0) or sheet (values towards 1) categories (Supplementary Fig. 9). Even with non-unique secondary structure assignments, a matching AlphaFold2 assignment still tends to reliably capture rigid in-solution helix and sheet. This might explain the observation that AlphaFold2 models often better encompass experimental NMR information than the calculated NMR ensembles themselves [62], as secondary structure elements will be well defined in AlphaFold2 models compared to NMR ensembles, for which the experimental NMR data might be insufficient to accurately form those secondary structures. Conversely, mismatches between the NMR ensemble and AlphaFold2 helix and/or sheet assignments are strongly indicative of dynamic behavior and multiple conformations in solution. Note that our analysis attempts to highlight the differences with AlphaFold2 based on simple observations from NMR ensembles and experimental data, and does not try to define the accuracy of NMR ensembles, for which methods such as ANSURR exist already [25].

The ability of AlphaFold2 to detect the presence of multiple conformations at the **residue level** does not necessarily extend to the full protein level, as it is not capable, when used in default mode, to detect proteins that can switch fold [63]. Likely this is because the overall fold prediction for the full protein is highly dependent on evolutionary information from a multiple sequence alignment, whereas at the residue level the overall atomic interaction information from the PDB that AlphaFold2 has learned is more likely to dominate. This is also reflected by the ShiftCrypt values in Fig. 4 F.

Residues in the core and surrounding sheet regions, which are dependent on long-range interactions between residues, feature higher pLDDT than core and surrounding helix, which are reliant on local interactions between residues. Notably, surrounding sheet, a relaxed conformational state with dynamic character, still has higher pLDDT than core helix, a rigid well-defined conformational state, showing that the prediction confidence of the former is likely driven by evolutionary information for the overall fold of the protein, whereas the prediction for the latter is driven by highly local residue interactions.

The pLDDT values show a binary relationship to experimentally observed dynamics, estimated by the S^2_{RCI} (Fig. 3 A/B) and ShiftCrypt methods (Fig. 4) as well as directly measured S^2 order parameters (Fig. 3 C/D). High pLDDT values correspond to regions that have a stable, single rigid conformation in solution, while residues with mid and lower pLDDT values indicate the presence of multiple conformations and dynamics, but are only very weakly correlated with the degree of dynamics present. Whilst AlphaFold2 is thus capable of detecting a hard order/disorder boundary, as previously reported [15], it does not capture how dynamic a residue might be. This is not surprising given its training data, which mainly constitutes folded proteins organised in crystals and measured at cryogenic temperatures, so not capturing actual in solution dynamics [16]. Residues observed to be dynamic in solution but still predicted with high pLDDT, on the other hand, seem to indicate regions that can fold, but only in the right circumstances. AlphaFold2 tends to predict bound states of proteins if present in its training data, and while disordered regions are typically missing x-ray diffraction PDB structures, they are visible when adopting a single conformation while interacting with another protein or ligand[56].

These conclusions are summarised by the relation of pLDDT to the conformational state variability calculated from MD trajectories (Fig. 6). High pLDDT values correspond to low conformational state variability. Conversely, mid to low pLDDT values correspond to high conformational state variability, but the pLDDT does not pick up the degree of variation. These two trends capture AlphaFold2's capacity for binary order/disorder distinction, but its inability to capture the degree of conformational variability and dynamics.

Finally, this work shows that it is not trivial to estimate actual experimental dynamics from NMA on the AlphaFold2 models, with a complex relationship between these parameters. Rigid, well-folded parts of the protein are easily detected by the NMA RMSF and pLDDT values, but the interpretation of regions with lower pLDDT depends on whether these regions connect folded parts of the protein, with no straightforward relation between NMA RMSF and experimentally determined dynamics (Table 2, Fig. 8). This correlation improves when using NMR ensemble models as input for the NMA (Supplementary Figure 18). This contrasts with the ANSURR results, where rigidity is modelled as geometric constraints in protein structures using graph edges based on H-bonds and other interactions. Flexibility is computed by incrementally removing H-bond edges based on their energy thresholds (converted to Boltzmann population ratios) and observing when the C_α atom becomes flexible [25]. The increased accuracy with regard to this ANSURR rigidity score for AF2 models compared to NMR ensemble models can be primarily attributed to AF2 models having more extensive and correctly placed H-bond networks, making them more rigid. The AlphaFold2 models then perform better in regions with extensive H-bond networks, with NMR structure ensembles rarely better except in dynamic regions [24]. The reliance of ANSURR on specific H-bonding networks, and not C_α atom distances as for NMA, might here be the key in explaining this difference. Whereas the variability between models in an NMR ensemble can interfere with the *in silico* definition of H-bonds at atomic precision, this variability does reflect in solution dynamics to some degree. NMA then seems to be better able to capture this information at the lower resolution of inter-C_α distances.

To conclude, the results in this work show that the AlphaFold2 pLDDT indicates the presence or absence of multiple conformations and the associated protein dynamics. It does not, however, capture the gradations of dynamics nor the number of possible conformations present. The RMSF of NMA on the AlphaFold2 models equally does not capture such information. Experimental data, such as from NMR, and more fine-grained computational approaches, such as MD simulations, therefore remain invaluable to assess the movements and conformational states of proteins. While AlphaFold2 has fast-forwarded the field of structural biology and our understanding of the space of protein folds, its limitations are, as with any machine learning method, determined by its training data, which does not incorporate dynamics. Notwithstanding the applicability of AlphaFold2 on predicting multiple conformations on a selected set of well-studied proteins [7], this study highlights the complexity of the problem when relating AlphaFold2 to large scale experimental data capturing the presence of multiple conformations and the degree of dynamics present. Our ability to predict multiple conformations and dynamics will therefore likely remain limited until the lack of reliable and extensive experimental training data that encompasses multiple conformational states and the dynamics of proteins is resolved.

Tables

Table 1: **Dataset overview.** Content of the S^2_{RCI}, S^2 and Molecular Dynamics (MD) datasets used in this work. The full list of data elements can be found in the supplementary data frames (`https://zenodo.org/doi/10.5281/zenodo.10977724`).

Dataset	Number of proteins	Number of residues	Types of per-residue available data
All	-	-	• AlphaFold2 pLDDT • STRIDE secondary structure annotation from AlphaFold2 model. • NMA fluctuations from AlphaFold2 model. • BMRB, PDB and Uniprot identifiers.
S^2_{RCI}	762	374,358	• Random coil index derived S^2_{RCI} order parameter from NMR ensemble. • STRIDE secondary structure annotation from NMR models. • $\delta 2D$ secondary structure fractions from NMR ensemble. • NMA fluctuations from NMR models.
S^2	42	6,203	• Experimental S^2 order parameter. • AlphaFold3 C_α pLDDT.
MD	100	11,370	• Constava's MD-derived conformational state variability and propensities. • AlphaFold3 C_α pLDDT.

Table 2: **Overview of Pearson correlation for diverse metrics at different pLDDT and S^2_{RCI} ranges.** *Note: p-values marked with * were too low for Scipy to differentiate from 0.*

Metric A	Metric B	Stratification Metric	Stratus range	Pearson's correlation	P-value	n
pLDDT	S^2_{RCI}	None	Unstratified	0.6727	0*	76,711
pLDDT	RMSF	None	Unstratified	-0.5029	0*	338,301
S^2_{RCI}	RMSF	None	Unstratified	-0.2197	0*	73,814
pLDDT	S^2_{RCI}	pLDDT	Low	0.3129	5.21×10^{-119}	5,223
pLDDT	S^2_{RCI}	pLDDT	Mid	0.3123	1.16×10^{-199}	8,859
pLDDT	S^2_{RCI}	pLDDT	High	0.2348	0*	62,629
pLDDT	RMSF	pLDDT	Low	-0.0359	1.93×10^{-26}	87,690
pLDDT	RMSF	pLDDT	Mid	-0.0721	2.99×10^{-47}	40,041
pLDDT	RMSF	pLDDT	High	-0.1272	0*	210,570
S^2_{RCI}	RMSF	pLDDT	Low	-0.2001	1.89×10^{-32}	3,445
S^2_{RCI}	RMSF	pLDDT	Mid	-0.1262	3.34×10^{-30}	8,128
S^2_{RCI}	RMSF	pLDDT	High	-0.1262	3.34×10^{-210}	62,241
pLDDT	S^2_{RCI}	S^2_{RCI}	Flexible	0.4772	0*	13,656
pLDDT	S^2_{RCI}	S^2_{RCI}	Context-dependent	0.1547	2.26×10^{-73}	13,560
pLDDT	S^2_{RCI}	S^2_{RCI}	Rigid	0.1841	0*	49,495
pLDDT	RMSF	S^2_{RCI}	Flexible	-0.2348	5.39×10^{-139}	11,113
pLDDT	RMSF	S^2_{RCI}	Context-dependent	-0.1118	2.80×10^{-38}	13,314
pLDDT	RMSF	S^2_{RCI}	Rigid	-0.1402	2.56×10^{-215}	49,387
S^2_{RCI}	RMSF	S^2_{RCI}	Flexible	-0.1724	6.96×10^{-75}	11,113
S^2_{RCI}	RMSF	S^2_{RCI}	Context-dependent	-0.0542	4.03×10^{-10}	13,314
S^2_{RCI}	RMSF	S^2_{RCI}	Rigid	-0.0627	3.59×10^{-44}	49,387

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement [813239 to J.G.-G. and A.D.]; Research Foundation Flanders (FWO) International Research Infrastructure [I000323N to W.V. and G.0328.16N to A.D.]; COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology). We also acknowledge project funding to A.G. of the FWO (project G002520N and project G094023N), BOF of Ghent University, and the European Union (ERC, 101086145 PASTIME). The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI.

Author Contributions

Conceptualization: WV. Methodology: JGG, BD, WV, AG. Software: JGG, BD. Formal analysis: JGG, BD. Web server and interactive analysis: JGG, AD. Data Curation: JGG, BD, WV, AG. Writing: JGG, BD, WV, AG. Visualisation: JGG, BD. Supervision: WV, AG. Funding acquisition: WV, AG.

Competing Interests

The authors declare no competing interests.

References

[1] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

[2] Bordin, N. *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Communications Biology* **6**, 160 (2023). URL https://www.nature.com/articles/s42003-023-04488-9.

[3] Durairaj, J. *et al.* Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653 (2023). URL https://www.nature.com/articles/s41586-023-06622-3.

[4] Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022). URL https://www.science.org/doi/full/10.1126/science.add2187. Publisher: American Association for the Advancement of Science.

[5] Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023). URL https://www.nature.com/articles/s41586-023-06415-8. Number: 7976 Publisher: Nature Publishing Group.

[6] Sumida, K. H. *et al.* Improving Protein Expression, Stability, and Function with ProteinMPNN. *Journal of the American Chemical Society* **146**, 2054–2061 (2024). URL https://doi.org/10.1021/jacs.3c10941. Publisher: American Chemical Society.

[7] Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* 1–8 (2023). URL https://www.nature.com/articles/s41586-023-06832-9. Publisher: Nature Publishing Group.

[8] Heo, L. & Feig, M. Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins: Structure, Function, and Bioinformatics* **90**, 1873–1885 (2022). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26382. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26382.

[9] Porter, L. L. & Looger, L. L. Extant fold-switching proteins are widespread. *Proceedings of the National Academy of Sciences* **115**, 5968–5973 (2018). URL https://www.pnas.org/doi/abs/10.1073/pnas.1800168115. Publisher: Proceedings of the National Academy of Sciences.

[10] Kovermann, M., Rogne, P. & Wolf-Watz, M. Protein dynamics and function from solution state nmr spectroscopy. *Quarterly Reviews of Biophysics* **49**, e6 (2016).

[11] Noé, F., De Fabritiis, G. & Clementi, C. Machine learning for protein folding and dynamics. *Current Opinion in Structural Biology* **60**, 77–84 (2020). URL https://www.sciencedirect.com/science/article/pii/S0959440X19301447.

[12] Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3799472/.

[13] Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2021). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8728224/.

[14] AlphaFold Protein Structure Database. URL https://alphafold.ebi.ac.uk/faq. Last visited 2023-12-19.

[15] Piovesan, D., Monzon, A. M. & Tosatto, S. C. E. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Science* **31**, e4466 (2022). URL https://onlinelibrary.wiley.com/doi/10.1002/pro.4466.

[16] Fenwick, R. B., van den Bedem, H., Fraser, J. S. & Wright, P. E. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proceedings of the National Academy of Sciences* **111**, E445–E454 (2014). URL https://www.pnas.org/doi/abs/10.1073/pnas.1323440111. Publisher: Proceedings of the National Academy of Sciences.

[17] van den Bedem, H. & Fraser, J. S. Integrative, dynamic structural biology at atomic resolution–it's about time. *Nature methods* **12**, 307–318 (2015).

[18] Varadi, M. *et al.* pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Research* **42**, D326–D335 (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3964940/.

[19] Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W. & Bax, A. Backbone dynamics of calmodulin studied by nitrogen-15 relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. *Biochemistry* **31**, 5269–5278 (1992). URL https://pubs.acs.org/doi/abs/10.1021/bi00138a005.

[20] Romero, P. R. *et al.* BioMagResBank (BMRB) as a Resource for Structural Biology. In Gáspári, Z. (ed.) *Structural Bioinformatics*, vol. 2112, 187–218 (Springer US, New York, NY, 2020). URL http://link.springer.com/10.1007/978-1-0716-0270-6_14. Series Title: Methods in Molecular Biology.

[21] Berjanskii, M. V. & Wishart, D. S. A Simple Method To Predict Protein Flexibility Using Secondary Chemical Shifts. *Journal of the American Chemical Society* **127**, 14970–14971 (2005). URL https://doi.org/10.1021/ja054842f. Publisher: American Chemical Society.

[22] Berjanskii, M. V. & Wishart, D. S. Application of the random coil index to studying protein flexibility. *Journal of Biomolecular NMR* **40**, 31–48 (2008). URL https://doi.org/10.1007/s10858-007-9208-0.

[23] Orlando, G., Raimondi, D., Kagami, L. P. & Vranken, W. F. ShiftCrypt: a web server to understand and biophysically align proteins through their NMR chemical shift values. *Nucleic Acids Research* **48**, W36–W40 (2020). URL https://doi.org/10.1093/nar/gkaa391.

[24] Fowler, N. J. & Williamson, M. P. The accuracy of protein structures in solution determined by AlphaFold and nmr. *Structure* **30**, 925–933 (2022).

[25] Fowler, N. J., Sljoka, A. & Williamson, M. P. A method for validating the accuracy of nmr protein structures. *Nature communications* **11**, 6321 (2020).

[26] Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical reviews* **110**, 1463–1497 (2010).

[27] Tiwari, S. P. *et al.* Webnm@ v2. 0: Web server and services for comparing protein flexibility. *BMC bioinformatics* **15**, 1–12 (2014).

[28] Atilgan, A. R. *et al.* Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal* **80**, 505–515 (2001).

[29] Brueschweiler, R. Normal modes and nmr order parameters in proteins. *Journal of the American Chemical Society* **114**, 5341–5344 (1992).

[30] Doruker, P., Jernigan, R. L. & Bahar, I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. *Journal of computational chemistry* **23**, 119–127 (2002).

[31] Haliloglu, T. & Bahar, I. Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with x-ray diffraction and nmr relaxation data. *Proteins: Structure, Function, and Bioinformatics* **37**, 654–667 (1999).

[32] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (eds.) *Proceedings of the 9th Python in Science Conference*, 56 – 61 (2010).

[33] pandas development team, T. pandas-dev/pandas: Pandas (2020). URL https://doi.org/10.5281/zenodo.3509134.

[34] Hoch, J. C. *et al.* Biological Magnetic Resonance Data Bank. *Nucleic Acids Research* **51**, D368–D376 (2023). URL https://doi.org/10.1093/nar/gkac1050.

[35] Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. From protein sequence to dynamics and disorder with DynaMine. *Nature Communications* **4**, 2741 (2013). URL https://www.nature.com/articles/ncomms3741. Number: 1 Publisher: Nature Publishing Group.

[36] Gavalda-Garcia, J. *et al.* Data-driven probabilistic definition of the low energy conformational states of protein residues. *NAR Genomics and Bioinformatics* **6**, lqae082 (2024). URL https://doi.org/10.1093/nargab/lqae082.

[37] Rieping, W. & Vranken, W. F. Validation of archived chemical shifts through atomic coordinates. *Proteins: Structure, Function, and Bioinformatics* **78**, 2482–2489 (2010). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22756. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.22756.

[38] Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000). URL https://doi.org/10.1093/nar/28.1.235.

[39] Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics* **23**, 566–579 (1995). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340230412. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.340230412.

[40] Lovell, S. C. *et al.* Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics* **50**, 437–450 (2003). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10286. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.10286.

[41] Hollup, S. M., Salensminde, G. & Reuter, N. Webnm@: a web application for normal mode analyses of proteins. *BMC bioinformatics* **6**, 1–8 (2005).

[42] Ghysels, A., Miller, B. T., Pickard IV, F. C. & Brooks, B. R. Comparing normal modes across different models and scales: Hessian reduction versus coarse-graining. *Journal of Computational Chemistry* **33**, 2250–2275 (2012).

[43] Vranken, W. F. NMR structure validation in relation to dynamics and structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy* **82**, 27–38 (2014). URL https://www.sciencedirect.com/science/article/pii/S0079656514000582.

[44] Fossat, M. J., Posey, A. E. & Pappu, R. V. Uncovering the Contributions of Charge Regulation to the Stability of Single Alpha Helices. *ChemPhysChem* **24**, e202200746 (2023). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cphc.202200746. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cphc.202200746.

[45] Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. The dynamine webserver: predicting protein dynamics from sequence. *Nucleic acids research* **42**, W264–W270 (2014).

[46] Camilloni, C., De Simone, A., Vranken, W. F. & Vendruscolo, M. Determination of Secondary Structure Populations in Disordered States of Proteins Using Nuclear Magnetic Resonance Chemical Shifts. *Biochemistry* **51**, 2224–2231 (2012). URL https://doi.org/10.1021/bi3001825. Publisher: American Chemical Society.

[47] Sabzekar, M., Naghibzadeh, M., Eghdami, M. & Aydin, Z. Protein β-sheet prediction using an efficient dynamic programming algorithm. *Computational Biology and Chemistry* **70**, 142–155 (2017). URL https://linkinghub.elsevier.com/retrieve/pii/S1476927117301494.

[48] Jeong, J., Berman, P. & Przytycka, T. M. Improving strand pairing prediction through exploring folding cooperativity. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **5**, 484–491 (2008). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2597093/.

[49] Aydin, Z., Altunbasak, Y. & Erdogan, H. Bayesian Models and Algorithms for Protein β-Sheet Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**, 395–409 (2011). URL http://ieeexplore.ieee.org/document/4745629/.

[50] Bryson, J. W. *et al.* Protein Design: A Hierarchic Approach. *Science* **270**, 935–941 (1995). URL https://www.science.org/doi/10.1126/science.270.5238.935.

[51] Acharyya, A. *et al.* Exposing the Nucleation Site in α-Helix Folding: A Joint Experimental and Simulation Study. *The journal of physical chemistry. B* **123**, 1797–1807 (2019). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6497059/.

[52] Jesus, C. S. H., Cruz, P. F., Arnaut, L. G., Brito, R. M. M. & Serpa, C. One Peptide Reveals the Two Faces of α-Helix Unfolding–Folding Dynamics. *The Journal of Physical Chemistry B* **122**, 3790–3800 (2018). URL https://doi.org/10.1021/acs.jpcb.8b00229. Publisher: American Chemical Society.

[53] Taylor, J. W., Greenfield, N. J., Wu, B. & Privalov, P. L. A calorimetric study of the folding-unfolding of an α-helix with covalently closed N and C-terminal loops 1. *Journal of Molecular Biology* **291**, 965–976 (1999). URL `https://www.sciencedirect.com/science/article/pii/S0022283699930255`.

[54] Abrusán, G. & Marsh, J. A. Alpha Helices Are More Robust to Mutations than Beta Strands. *PLoS Computational Biology* **12**, e1005242 (2016). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5147804/`.

[55] Robustelli, P., Piana, S. & Shaw, D. E. Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein. *Journal of the American Chemical Society* **142**, 11092–11101 (2020). URL `https://doi.org/10.1021/jacs.0c03217`. Publisher: American Chemical Society.

[56] Saldaño, T. *et al.* Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* **38**, 2742–2748 (2022). URL `https://doi.org/10.1093/bioinformatics/btac202`.

[57] Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods* **20**, 205–213 (2023). URL `https://www.nature.com/articles/s41592-022-01685-y`. Number: 2 Publisher: Nature Publishing Group.

[58] DiNitto, J. P. *et al.* Structural basis and mechanism of autoregulation in 3-phosphoinositide-dependent Grp1 family Arf GTPase exchange factors. *Molecular Cell* **28**, 569–583 (2007).

[59] Saldaño, T. *et al.* Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* **38**, 2742–2748 (2022).

[60] Akdel, M. *et al.* A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology* **29**, 1056–1067 (2022).

[61] Bruley, A., Mornon, J.-P., Duprat, E. & Callebaut, I. Digging into the 3d structure predictions of alphafold2 with low confidence: disorder and beyond. *Biomolecules* **12**, 1467 (2022).

[62] Li, E. H. *et al.* Blind assessment of monomeric AlphaFold2 protein structure models with experimental NMR data. *Journal of Magnetic Resonance* **352**, 107481 (2023). URL `https://www.sciencedirect.com/science/article/pii/S1090780723001167`.

[63] Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Science* **31**, e4353 (2022). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4353`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4353.

Chapter 6

Conformation and dynamics of proteins with ambiguous behavior

Challenges in describing the conformation and dynamics of proteins with ambiguous behavior

Joel Roca-Martinez, Tamas Lazar, Jose Gavalda-Garcia, Rita Pancsa, **Bhawna Dixit**, Konstantina Tzavella, Pathmanaban Ramasamy, Maite Sanchez-Fornaris, Isel Grau, and Wim F. Vranken

6.1 Background and methodology

This study addresses the complex challenges of representing the biophysical behavior of proteins and their dynamics, focusing on the interpretation of conformational states and their relationship to function and physiological context such as post-translational modifications. The goal of this study was to categorize protein conformational behavior into three classes: ordered, disordered, and ambiguous. By utilizing three distinct datasets and applying interpretable machine learning techniques,features from AlphaFold2 and sequence-based predictions were analyzed to investigate the similarity and differences between them.

6.2 Contributions

As a co-author, my contribution included writing the manuscript addressing challenges of characterizing protein conformation and dynamics, with a particular focus on PTMs and MD simulations.

6.3 Concluding remarks

In conclusion, this study emphasized the importance of moving beyond the simplistic two-state model of proteins (a single, well-defined static fold or complete disorder) towards a more nuanced, probabilistic perspective, acknowledging that proteins can adopt a range of possible states, reflecting their inherent dynamic behavior.

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 03 August 2022
DOI 10.3389/fmolb.2022.959956

Check for updates

OPEN ACCESS

EDITED BY
Javier Oroz,
Institute of Physical Chemistry
Rocasolano (CSIC), Spain

REVIEWED BY
Vladimir N. Uversky,
University of South Florida,
United States
Adam Liwo,
University of Gdansk, Poland

*CORRESPONDENCE
Wim F. Vranken,
wim.vranken@vub.be

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Protein Folding, Misfolding, and Degradation, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 02 June 2022
ACCEPTED 27 June 2022
PUBLISHED 03 August 2022

CITATION
Roca-Martinez J, Lazar T, Gavalda-Garcia J, Bickel D, Pancsa R, Dixit B, Tzavella K, Ramasamy P, Sanchez-Fornaris M, Grau I and Vranken WF (2022), Challenges in describing the conformation and dynamics of proteins with ambiguous behavior.
Front. Mol. Biosci. 9:959956.
doi: 10.3389/fmolb.2022.959956

Challenges in describing the conformation and dynamics of proteins with ambiguous behavior

Joel Roca-Martinez[1,2†], Tamas Lazar[1,3†], Jose Gavalda-Garcia[1,2], David Bickel[1,2], Rita Pancsa[4], Bhawna Dixit[1,2,5], Konstantina Tzavella[1,2], Pathmanaban Ramasamy[1,2,6], Maite Sanchez-Fornaris[1,2,7], Isel Grau[8] and Wim F. Vranken[1,2*]

[1]Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium, [2]Interuniversity Institute of Bioinformatics in Brussels, VUB/ULB, Brussels, Belgium, [3]VIB-VUB Center for Structural Biology, Brussels, Belgium, [4]Research Centre for Natural Sciences, Institute of Enzymology, Budapest, Hungary, [5]IBiTech-Biommeda, Universiteit Gent, Gent, Belgium, [6]VIB-UGent Center for Medical Biotechnology, Universiteit Gent, Gent, Belgium, [7]Department of Computer Sciences, University of Camagüey, Camagüey, Cuba, [8]Information Systems, Eindhoven University of Technology, Eindhoven, Netherlands

Traditionally, our understanding of how proteins operate and how evolution shapes them is based on two main data sources: the overall protein fold and the protein amino acid sequence. However, a significant part of the proteome shows highly dynamic and/or structurally ambiguous behavior, which cannot be correctly represented by the traditional fixed set of static coordinates. Representing such protein behaviors remains challenging and necessarily involves a complex interpretation of conformational states, including probabilistic descriptions. Relating protein dynamics and multiple conformations to their function as well as their physiological context (e.g., post-translational modifications and subcellular localization), therefore, remains elusive for much of the proteome, with studies to investigate the effect of protein dynamics relying heavily on computational models. We here investigate the possibility of delineating three classes of protein conformational behavior: order, disorder, and ambiguity. These definitions are explored based on three different datasets, using interpretable machine learning from a set of features, from AlphaFold2 to sequence-based predictions, to understand the overlap and differences between these datasets. This forms the basis for a discussion on the current limitations in describing the behavior of dynamic and ambiguous proteins.

KEYWORDS

protein dynamics and conformation, sequence-based prediction, biophysical characteristics, post-translational modification (PTM), deleterious mutation, folding-upon-binding, fold switching

1 Introduction

The importance of protein dynamics for their (mis-)folding (Daggett and Fersht, 2003; Dobson, 2003) and functionality (Karplus and Kuriyan, 2005; Glazer, Radmer, and Altman, 2009) has been long recognized but has been overshadowed by the need to first understand how most proteins fold into well-defined three-dimensional structures (unique conformations) (Hunkapiller, Strickler, and Wilson, 1984; Berman et al., 2007). The recent impressive performance of AlphaFold2 (Jumper et al., 2021) in predicting such unique protein folds from i) protein sequence and evolutionary information curated by UniProt (The UniProt Consortium, 2021) and ii) the carefully assembled protein structure information from the Protein Data Bank over many decades (Berman et al., 2007) indicates that this problem is now largely solved. This also implies that experimental and computational approaches for proteins will now have to necessarily focus beyond their fold, specifically on understanding more about how proteins interact, which alternative conformations they might adopt, and how they move between these conformations. Indeed, many proteins show ambiguous conformational behavior, either in specific regions within folded domains [e.g., loops such as CDRs in antibodies (Armstrong, Piepenbrink, and Baker, 2008) or extracellular loops in GPCRs (Hilger, Masureel, and Kobilka, 2018)], in regions connecting folded domains [e.g., PEVK domain of titin (Hsin et al., 2011)], or the full protein in the case of intrinsically disordered proteins [e.g., Phd antitoxin from Bacteriophage P1 (De Gieter et al., 2014)]. This behavior does not have hard boundaries. For example, systematic studies on ambiguous/disordered proteins have already proved that missing residues in crystal structures do not always correlate with protein disorder. In fact, sometimes they are predicted as highly ordered (Gall et al., 2007). Similarly, residues that are present or missing for the same protein in different X-ray structures are rarely statically disordered and show a partial or conditional disorder under different experimental conditions (DeForte and Uversky, 2016). This different degree of disorder was previously described and categorized into foldable, non-foldable, or semi-foldable regions, where some protein regions undergo a structural rearrangement at a certain point in time, either spontaneously or induced (e.g., after binding with another molecule) (Uversky, 2013). These conformational changes often condition the functions that the proteins perform and break with the classical protein structure-function paradigm (Uversky, 2019), supporting the prevalence and importance of the ambiguous behavior that we are addressing. The move from the traditional paradigm, with the sequence encoding for a single static structure, toward a dynamic paradigm, where the sequence encodes for different possible behaviors, also implies the necessity to approach proteins from a probabilistic viewpoint. This is a reasonable assumption, especially when considering that billions of copies of the same protein exist in cells at thermodynamically high temperatures; all these proteins will have different interactions and (locally) different conformations at any given time point and might have (different) post-translational modifications (Vu, Gevaert and De Smet, 2018). Such a proteomics-based probabilistic *in vivo* view of proteins is in stark contrast to the reductionist and static single-protein view in the traditional paradigm.

There have nevertheless been significant efforts in the experimental investigation of the conformational ambiguity and heterogeneity of protein structures and structural ensembles by various techniques: nuclear magnetic resonance (NMR), circular dichroism (CD) and electron paramagnetic resonance (EPR) spectroscopy, small-angle X-ray and neutron scattering (SAXS/SANS), Förster resonance energy transfer (FRET) measurements, electrospray ionization–ion mobility mass spectrometry (ESI/IM-MS), and hybrid approaches that integrate more than one of the above-mentioned techniques (Dobson, 2019). Although X-ray crystallography and cryo-electron microscopy may both be able to trap more than one protein conformer of globular proteins, solution techniques are undoubtedly preferred for uncovering the dynamics of flexible proteins, with NMR being the approach that initially highlighted these features in proteins using different types of measurements (chemical shifts, R1, R2, J-couplings, NOEs, and RDCs). Lately, there have also been efforts dedicated to studying the dynamics of flexible and intrinsically disordered proteins (IDPs) in the cellular context using *in-cell* NMR and EPR spectroscopy, as a protein's conformational behavior may differ from what is observed in isolation in the test tube (Gerez, Prymaczok, and Riek, 2020; Bonucci et al., 2021). However, due to various experimental challenges, these methods have not become widely used in the community of structural biology. Valid future alternatives for both single proteins (folding) and in-cell determination of protein states might come from mass spectrometry-based methods such as cross-linking (XL-MS) or hydrogen–deuterium exchange (HDX-MS), which are becoming increasingly informative (Britt, Cragnolini, and Thalassinos, 2021).

On the computational side, molecular dynamics (MD) and Monte Carlo (MC) simulations are commonly used to investigate the conformations and/or dynamics of proteins, often in combination with experimental data to either restrain the structure of the protein or reweight a pool of structures generated from the simulation trajectory to obtain a conformational ensemble that complies with the experimental readout (Lindorff-Larsen et al., 2005; Hummer and Köfinger, 2015; Childers and Daggett, 2018; Orioli et al., 2020). Recent advances in force field (FF) development combined with enhanced sampling techniques now enables a more realistic exploration of protein dynamics and flexibility even in the absence of experimental data (Yang et al., 2019; Abriata and Dal Peraro, 2021). Besides the advances achieved in developing FFs that excel on IDPs (e.g., CHARMM36IDPSFF, Amber ffIDPs, and ffIDPSFF) (Huang and MacKerell, 2018; Zapletal et al., 2020; Mu et al., 2021), the major focus nowadays is on those achieving a balanced sampling on both folded and disordered proteins [such as CHARMM36m (Huang et al., 2017), Amber ff19SB (Tian et al., 2020), and DES-Amber (Piana et al., 2020)]. The main advantage of these simulations is their capability to account for context-dependency (e.g., temperature, ionic strength, PTMs, and a partner). However, their disadvantage is their computational cost, which prohibits proteome-wide/large-scale systematic analyses. To this end, various fast and computationally inexpensive sequence-based predictors have been developed, with many focusing on estimating intrinsic disorder. Disorder predictors can be cataloged into three main categories given their underlying prediction model: (1) *ab initio* methods like IUPred (Dosztanyi et al., 2005), which are based on the protein's physicochemical properties; (2) machine learning algorithms trained on experimental annotations like Disomine (Orlando et al., 2022), Disopred (Ward et al., 2004), DisEMBL (Linding et al., 2003), and SPOT-DISORDER2 (Hanson et al., 2019); and (3) the meta-predictors that combine several individual predictors, such as PONDR-FIT (Xue et al., 2010), ESpritz (Walsh et al., 2012), DISOPRED3 (Jones and Cozzetto, 2015), MFDp2 (Mizianty, Uversky, and Kurgan, 2014), and others. Usually, most of these predictors of protein disorder focus on labeling regions of missing electron density as regions of disorder using X-ray crystallography or NMR data, categorizing each residue in only one of two classes, ignoring potentially useful conformational states of the protein. However, there are new predictors that address those kinds of different behaviors, like IUPred2A (Meszaros et al., 2018), ODINPred (Dass, Mulder, and Nielsen, 2020), and DispHred (Santos, Iglesias, and Pintado, et al., 2020), assigning a degree of disorder to each amino acid and other predicted features of the protein indicating the amount or degree of disorder, like NetSurfP-2.0 (Klausen et al., 2019) that outputs solvent accessibility, secondary structure, structural disorder, and backbone dihedral angles for each residue of the input sequences. The intrinsically semi-disordered state has also been studied, with predictors able to identify such behavior often associated with induced folders and aggregation-prone regions (Zhang et al., 2013, Zhang et al., 2017; Katuwawala et al., 2019). In addition, other sequence-based predictors provide useful information, such as backbone dynamics (DynaMine) (Cilia et al., 2013, 2014), fuzziness (FuzPred) (Horvath et al., 2020; Miskei et al., 2020), secondary structure [PSIPRED4 (Jones, 1999), SPOT-1D (Singh et al., 2021)], solvent accessibility [SABLE (Adamczak, Porollo, and Meller, 2004), ACCpro (Magnan and Baldi, 2014), SPOT-1D (Singh et al., 2021)], solubility/aggregation propensity [TANGO (Fernandez-Escamilla et al., 2004), AGMATA (Orlando et al., 2020), PASTA2 (Walsh et al., 2014), CamSol (Sormanni, Aprile, and Vendruscolo, 2015)], liquid-liquid phase separation propensity [catGRANULE (Bolognesi et al., 2016), PScore (Vernon et al., 2018), PSPer (Orlando et al., 2019, p.), Droppler (Raimondi et al., 2021)], and other biophysical features of proteins. As most of these prediction tools only take the sequence as input, with sometimes a few specificities or sensitivity parameters, they remain largely context-independent and cannot take factors such as pH, temperature, or PTMs into account. The exception is a few specific cases, such as (i) oxidation-dependent disorder prediction by IUPred2A (Mészáros et al., 2018, p. 2); (ii) pH-dependent solubility prediction for IDPs by SolupHred (Santos et al., 2020a; 2020b; Pintado et al., 2021); (iii) prediction of molecular recognition features/elements (MoRFs/MoREs) that are interacting regions of IDPs undergoing an increase in the secondary structure propensity upon binding (e.g., α-MoRF-PredII predictors (Oldfield et al., 2005; Cheng et al., 2007), MORFchibi (Malhis, Jacobson and Gsponer, 2016), SPOT-MoRF(Hanson et al., 2020), and fMoRFpred (Yan et al., 2016)); and (iv) experimental condition (pH, temperature, ionic strength, crowding agent, and protein concentration)-dependent prediction of liquid-liquid phase separation by Doppler (Raimondi et al., 2021).

Another significant influence on protein behavior is post-translational modifications (PTMs), which regulate the function, activity, and stability of proteins. Several studies have shown the association of PTMs with various diseases, such as cancer, Alzheimer's, and diabetes (McLaughlin et al., 2016; Song and Luo, 2019; Bai et al., 2021). PTMs alter the biophysical, thermodynamic, and kinetic properties of proteins, leading to a more diverse conformational landscape than dictated by the arrangement of 20 amino acids (Shental-Bechor and Levy, 2008). Therefore, a complete comprehension of a folded protein monomer is useful but insufficient to understand the functioning of a protein in a biological environment. The structural preferences of PTMs are divided into two categories: well-defined secondary structures (N-linked glycosylation, acetylation) and intrinsically disordered regions (phosphorylation, methylation). These PTMs can exist simultaneously in different amino acids (methylation, phosphorylation), or in the same amino acid over time (ubiquitination, phosphorylation), depending on the biological context. The impact of PTMs on protein structures can vary diversely, ranging from local conformational stabilization or destabilization of secondary structure elements to transitions between intrinsically disordered and ordered states (Bah and Forman-Kay, 2016).

In the case of IDPs, the disorder-to-order transitions can be considered "a black box of structural biology." This multifaceted folding/unfolding behavior is widely regulated and modulated by PTMs. The alteration of IDPs' conformational space, dynamics, functionality, cellular expression, and localization caused by PTMs can also be unfavorable and cause protein pathogenicity. This equivocal relationship between PTMs and IDPs significantly enlarges the

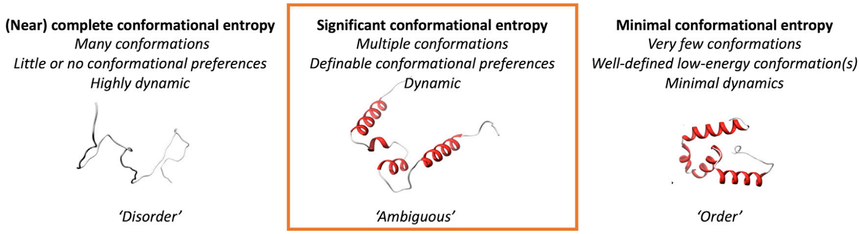

(Near) complete conformational entropy
Many conformations
Little or no conformational preferences
Highly dynamic
'Disorder'
Significant conformational entropy
Multiple conformations
Definable conformational preferences
Dynamic
'Ambiguous'
Minimal conformational entropy
Very few conformations
Well-defined low-energy conformation(s)
Minimal dynamics
'Order'

FIGURE 1
Conceptual definition of the "ambiguous" regions of proteins addressed in this study.

complexity of the black box, which is invisible yet an important attribute of protein folding (Bah and Forman-Kay, 2016). Currently, the change in conformational dynamics of a protein when modified by a PTM can be investigated by MD simulations. However, the systematic force-field parameters required for MD simulations are limited to several PTMs (methylation, phosphorylation, glycosylation) and require optimization and validation, which is computationally expensive. It, therefore, remains a black box since the current tools are deficient in terms of exploring PTMs and the conformational behavior of proteins. On the other hand, the stability of folded regions can also be affected by PTMs. Incorporating information about PTMs into our understanding of *in vivo* protein behavior is, therefore, essential.

We here explore a class of protein regions that are more likely to adopt multiple different conformations and show ambiguous behavior; they can neither be strictly classified as traditional "order," nor as the oppositely defined "disorder" (Figure 1). We focus on three different scenarios of conformational ambiguity: (i) regions that undergo "order-to-disorder" transitions, where a protein (region) that is disordered folds when encountering a binding partner, (ii) regions of folded proteins that can change their conformation, and (iii) regions that have ambiguous behavior in solution based on NMR chemical shift information. Such inherent ambiguous behavior could be relevant for conformational changes in the protein, for example, upon oligomerization, interacting with another molecule or the cell membrane, or when being post-translationally modified. These changes should happen within the context of biologically reasonable environments and protein modifications, for example, in disorder-or-order inducing agents such as TFE, or denaturing agents like urea. We here show, based on two different definitions and their joint one, that ambiguous regions are difficult to define but that combinations of datasets from different sources might help to unravel this complex protein behavior.

2 Materials and methods

2.1 Datasets

2.1.1 DisProt "folding-upon-binding" dataset with CoDNaS dataset (disprot_codnas_set)

DisProt is a large database of manually curated intrinsically disordered protein (IDP) regions (IDRs) (Hatos et al., 2020). Besides the structural state and the function of the region, if available, interaction partners and potential structural transitions (e.g., displaying folding-upon-binding) are also annotated for DisProt entries. For the present study, we downloaded a custom set of human proteins with manually curated disorder-to-order structural transitions, resulting in 138 different proteins with at least one IDR that undergoes ordering. The residues that are classified as undergoing structural ordering were labeled as ambiguous (N = 9,792 residues) and the residues in the IDR flanking regions that are not proven to undergo structural ordering were labeled as disordered (N = 4,232 residues).

CoDNaS (Monzon et al., 2016) stores proteins with multiple X-ray and NMR structures solved under different experimental conditions. The difference between these conformations of "snapshots" varies over a wide range, with rigid globular structures being on one side of the spectrum and disordered structures on the other side. To assemble a set of rigid proteins, we downloaded structural clusters by applying the threshold of a maximum RMSD value of 2Å for each pair of structures available for the same protein region. This way, we obtained a reliable set of 207 human proteins entailing 11,947 residues in ordered segments.

These two datasets were combined into a single dataset, which, therefore, contains highly reliable definitions for ordered residues (O) for which little or no conformational change has been observed in experimental protein structures (from CoDNaS) as well as disorder (D) (from DisProt) and ambiguous behavior folding-upon-binding residues, with a local change in environment (the binding partner) triggering a conformational transition or rearrangement (T) (from DisProt).

2.1.2 MFIB dataset (mfib_set)

MFIB (Fichó et al., 2017) is a database of mutually folded IDPs/IDRs that synergistically fold upon binding, while as monomers, the protein chains are unstructured. A subset of MFIB was manually selected to reduce the redundancy in terms of a sequence-structure relationship. Additional overlap with other datasets has also been filtered out; in total, five protein chains that were part of the DisProt set have been eliminated. The final set of cases includes 17 chains from homo- and 23 chains from heterocomplexes forming various types of folds (including histone-like folds; basic helix-loop-helix; Phe-, Leu-, and Ala-zippers; and ribbon-helix-helix folds), with 1–3 examples selected from each fold category. The complete dataset is available at https://bitbucket.org/bio2byte/protein_ambiguity/.

2.1.3 Metamorphic and fold-switching proteins dataset (foldswitch_set)

The fold switchers dataset is a manually curated list of pairs of experimentally solved structures for the same protein that shows a different topology in some parts of the sequence. This dataset provides experimental proof of residues that can switch from one secondary structure element type to another one (e.g., a residue that in one of the PDB structures is in an α-helix and in the other one is in a β-strand). The original fold switchers list consisted of 94 protein pairs (PDB entries), but we filtered it to keep only the protein sequences that shared the same sequence, as small sequence variations could have an impact on the protein topology and would, therefore, affect our study. A total of 29 structure pairs remained, totaling 8,047 residues. This dataset is available at https://bitbucket.org/bio2byte/protein_ambiguity/ as supplementary material.

The residues were labeled using the DSSP secondary structure annotations (Kabsch and Sander, 1983) extracted from the PDBe API (Mir et al., 2018) for each of the structures in the pair. Residues that stayed in either helix or sheet conformations were labeled as the same (S), while residues that switched from any secondary structure type to another one were labeled as converted (C). We did not use the residues that stayed in the coil for this analysis to avoid including likely disordered regions in either of the two aforementioned categories. A total of 3,751 and 1,341 residues were labeled as S and C, respectively.

2.1.4 Combined dataset (combined_set)

A new dataset merging the disprot_codnas_set and foldswitch_set was generated by combining some of the categories of the previous ones (combined_set). The ordered (O) and same (S) categories from the disprot_codnas_set and foldswitch_set were merged as they were comparably defined. In both cases, the residues that fall into these categories are amino acids that have proved rigid/conformationally stable in several experimental assays. Similarly, the ambiguous folding-upon-binding residues (T) from DisProt and the fold-switching residues (C) also share a particular biophysical behavior, as in both categories the residues undergo conformational rearrangement. The goal is to assess whether this dataset exhibits similar features with respect to the disprot_codnas_set and foldswitch_set or whether it captures different biophysical characteristics. The disordered category (D) remains as defined in the disprot_codnas_set. The total number of residues in this set is 15,698, 10,750, and 4,232 for ordered (O + S), ambiguous (T + C), and disordered (D), respectively.

2.1.5 Post-translational modification dataset (ptm_set)

PTM information was obtained from four different resources: Scop3P (Ramasamy et al., 2020), UniProtKB/Swiss-Prot (The UniProt Consortium, 2021), dbPTM (Huang et al., 2019), and PhosphoSitePlus (PSP) (Hornbeck et al., 2015). Scop3P annotates protein phosphorylation sites by re-processing large-scale public proteomics datasets. dbPTM integrates experimentally validated PTM sites from Swiss-Prot, PhosphoELM, and O-GLYCBASE. UniProtKB includes PTM information that is directly curated from scientific literature and propagates the information to homologues. PSP contains manually curated PTM information obtained from the literature. We downloaded PTM information from all the above-mentioned resources (April 2022). All the obtained PTM sites were checked for correctness in sequence positions with the current UniProtKB/Swiss-Prot human protein sequences. To obtain a reliable set of PTM sites, we only considered sites having at least two different databases of evidence. Multiple sites having more than one PTM type are labeled as "multiple." The final dataset contains 217,082 PTM sites from 15,420 canonical human proteins. The complete data table is available at https://bitbucket.org/bio2byte/protein_ambiguity/.

2.1.6 Alphafold human proteome dataset (af_set)

AlphaFold 2's mmCIF files for the human proteome were downloaded on 2 September 2021, from the AlphaFold protein structure database (Tunyasuvunakool et al., 2021). In this section, we will refer to this dataset as "AF_dataset." According to AF_dataset's description page (https://alphafold.ebi.ac.uk/download), sequences longer than 2,700 residues were split into multiple files. For simplicity, we removed these sequences and kept only the sequences contained in a single file. Then, we extracted the protein ID, sequence, pLDDT, and secondary structure and simplified them to alpha_helix, beta_strand, and all remaining conformations were labeled as the coil.

We also downloaded all human Swiss-Prot entries contained in Uniref90 (Suzek et al., 2007) on 2 September 2021 from UniProt (The UniProt Consortium, 2021). In this section, we will refer to this dataset as "uniref_dataset." From this set, we discarded all proteins shorter than 20 amino acids since some of our predictive tools have this minimum length requirement. Then, we found the sequence intersection between AF_dataset and uniref_dataset and verified that the sequence in both sets was correctly aligned, which resulted in the "selected_human_dataset".

With these sequences, we computed sequence-based predictions with the b2btools predictors, comprising DisoMine (disorder) (Orlando et al., 2022), DynaMine [backbone (Cilia et al., 2013) and side-chain dynamics, conformational propensities (Raimondi et al., 2017)], EFoldMine (early folding propensity) (Raimondi et al., 2017) using a recently developed PyPI package currently in open beta (https://pypi.org/project/b2bTools/3.0.0b16/). We then merged our predictions with the mLDDT and secondary structure predictions that we extracted from the AF_dataset into our selected_human_dataset. Finally, our selected_human_dataset was saved into a NumPy file for later processing and can be found at https://bitbucket.org/bio2byte/protein_ambiguity/.

2.1.7 Deleterious mutant datasets

Even though mutation is a random process, it frequently occurs at highly conserved hotspots of the protein, which represent regions of structural and functional importance (Chang et al., 2018). To explore the definition of ambiguous regions, we downloaded publicly available deleterious somatic mutations from the catalog of somatic mutations in cancer (COSMIC version92_1,121) (Forbes et al., 2008) and Cancer Genome Interpreter (Tamborero et al., 2018) and germline deleterious and benign mutations from ClinVar (Landrum et al., 2018) and UniProtKB/Swiss-Prot (The UniProt Consortium, 2021), respectively. The COSMIC database contains more than 13 million mutations associated with various cancer types. UniProtKB/Swiss-Prot contains variant annotation from literature reports and ClinVar reports on the relationships among human variations and phenotypes, with supporting experimental evidence from the literature.

Two different analyses were performed. For the first one, 9,295 missense mutations were selected and mapped on 1,115 canonical UniProt ids with at least one deleterious and one benign mutation, resulting in 4,690 deleterious and 4,605 benign mutations. The second analysis focused on comparing somatic and germline deleterious missense mutations shared among 173 canonical isoforms, resulting in 2,145 somatic and 1,020 germline mutations. The datasets are available under the names "canonical_mut" and "germline_somatic_deleterious" at https://bitbucket.org/bio2byte/protein_ambiguity/.

2.2 Predictions

2.2.1 Feature generation from sequence

For all protein sequences in the datasets, seven biophysical features were predicted at the residue level using the following methods: backbone dynamics (DynaMine) (Cilia et al., 2013), side-chain dynamics (Raimondi et al., 2017), conformational propensities (helix, sheet, and coil) (Raimondi et al., 2017), early folding propensity (Raimondi et al., 2017), and disorder (DisoMine) (Orlando et al., 2022).

2.2.2 Random forest predictor for folding-upon-binding regions of proteins

The disprot_set describes protein regions that are initially disordered but fold upon binding, with a local change in environment (the binding partner) triggering a conformational rearrangement, while the codnas_set describes residues for which little or no conformational change has been observed in experimental protein structures. The disprot_set was used to define ambiguous/transitioning residues (T) as well as disordered residues (D) and whilst ordered residues (O) were defined from the codnas_set. We used a combination of these datasets (disprot_codnas_set) to train a random forest (RF) predictor, termed folding_upon_binding_RF, with the main aim of creating an interpretable predictor, not necessarily a predictor with the best possible performance. The classification model was trained using seven predicted biophysical features at the residue level (see the previous section). No amino acid codes were used in the training, with all the features computed using a local version of b2BTools from the single input sequences (Kagami et al., 2021). The previously defined residue categories (O, T, and D) were used as labels for the RF training. We used scikit-learn (Pedregosa et al., 2011) version 1.0.2 to generate all the models. The available information for the 25,588 residues was split into 90% and 10% between the training and test sets, respectively. For the training, a 3-fold cross-validation was performed to select the best hyperparameters (n_estimators = 75, max_depth = 15, min_samples_split = 5, min_samples_leaf = 1, and bootstrap = False). The RF model is trained using those hyperparameters and finally tested on the remaining 10% of the data (test set), from which our model is completely agnostic.

2.2.3 Combined random forest

The combined_set was generated by merging the ordered (O) and same (S) categories, and the transition (T) and convert (C) categories from the disprot_codnas_dataset and the foldswitch_set, respectively (for details, see c. f. *Datasets*). Again, the same biophysical predictions were used at the residue level as features for an RF classifier (combined_RF). The data was split 70% to 30% into train and test sets, respectively. The best hyper-parameters were retrieved using a 3-fold cross-validation (n_estimators = 25, max_depth = 15, min_samples_split = 5, min_samples_leaf = 5, and bootstrap =

TABLE 1 Performances of the trained random forest predictors.

Dataset	Label	Number	Precision	Recall	F1 score
disprot_codnas_set	Order	11,947	0.72	0.84	0.78
	Transition	9,409	0.65	0.6	0.62
	Disorder	4,232	0.72	0.5	0.59
foldswitch_set	Same	3,751	0.79	0.96	0.87
	Convert	1,341	0.72	0.26	0.38
combined_set	Order/Same	15,698	0.72	0.86	0.78
	Transition/convert	10,750	0.62	0.53	0.57
	Disorder	4,232	0.72	0.46	0.56

True) and the model was further validated by testing it on the test set that contains 30% of the original data.

2.2.4 Interpretation of random forest models

The RF models were interpreted using a surrogate model trained over the predictions for each of the models. To generate these models, we used the Weka (Eibe et al., 2016) implementation of the Ripper algorithm (Cohen, 1995) (Repeated Incremental Pruning to Produce Error Reduction) that works as a rule-based classification algorithm and supports multi-classification tasks. As a result, we obtained a limited set of rules that summarize the key information on the RF models to classify the residues into different categories. The surrogate models simplify the complexity of the original RF, making them easier to interpret, as the decision trees derived from the raw RF models are often too big and diverse to interpret without any further actions.

3 Results

In the first section, we describe the RF predictors of "ambiguous residues." We did not develop these predictors for optimal performance, but instead for interpretability in relation to the "biophysical" input features. Comparing the predictors, which are each trained on different classifications of ambiguity, enables us to detect whether they seem to recognize the same features (or not), with the aim of identifying whether the different ambiguity definitions (order/disorder transitions or residues that can change conformation in metamorphic/fold-switching proteins) seem to have the same origin. To further contextualize the input features and the classifications, we also describe the relationship of the ambiguous residues to the AlphaFold2 output, as well as information about post-translational modifications and deleterious amino acid variants.

3.1 Random forest model interpretation

The F1 scores for the folding_upon_binding_RF model to recognize folding-upon-binding regions of proteins based on the combined disprot_codnas_set are lowest for the disorder class (D), where especially the recall is significantly lower (0.67) (Table 1). The performances are overall acceptable and indicate that the model is predictive and captures essential information from the input biophysical features. These features were then ranked by importance (Figure 2), with the early folding (EFoldMine), disorder (DisoMine), and backbone dynamics (DynaMine) being the most relevant. The secondary structure propensities and side-chain dynamics were less relevant for this prediction.

The fold_switching_RF model, based on the foldswitch_set, has a high F1 score for retrieving residues that remain the same when the fold switches (S), but for the residues that convert to secondary structure (C), the F1 prediction performance is very low (0.36) due to very low recall (0.26) (Table 1). This indicates that the biophysical features, which essentially capture local sequence information, are insufficient to detect such residues, or alternatively, that there is little difference between the S and C categories. The amino acid content of fold-switching proteins is similar to those of ordered proteins with a few important differences, including higher valine/phenylalanine and lower proline content for the metamorphic regions (Figure 3). In these regards, this class of proteins is significantly different from intrinsically disordered proteins that have fewer valine and phenylalanine residues but more prolines (Figure 3). In terms of feature importance, the disorder content is the most relevant (Figure 2), indicating that a tendency toward flexibility and/or conformational ambiguity does play a role in distinguishing between the categories, however poor this distinction is.

Finally, the combined_RF model, where the O/S classes and the T/C classes were combined (combined_set), shows overall poorer F1 performances for the O/S classes compared to O and S separately, indicating that the definitions of O and S are likely different, while the T/C class F1 performance is in between the T and C classes, and the D performance drops (Table 1). The feature importance is similar to the one for the disprot_codnas_set (Figure 2). Although there is an imbalance in the absolute numbers of the O compared to S,

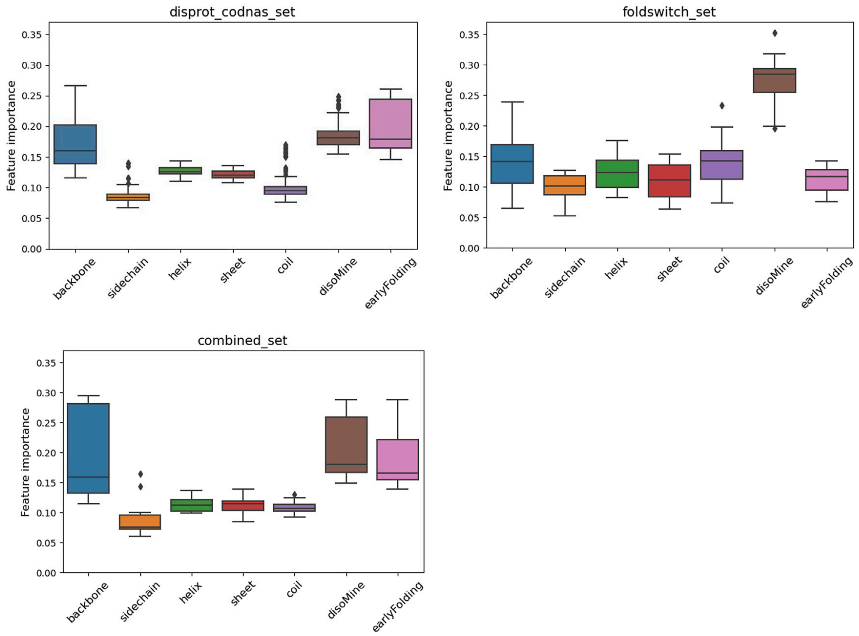


FIGURE 2
Feature importance variation for the RF classifier for the disprot_codnas_set (top left), the foldswitch_set (top right), and the combined_set (bottom left).

and T compared to C, classes, the sharp drop in overall performances indicates that the biophysical characteristics required for folding-upon-binding and for fold switching are fundamentally quite different.

The surrogate models generated from each of the RF models provide a perspective on the complexity of the data within. While both the codnas_disprot_set and combined_set surrogate models generate a large number of rules (84 and 89 rules, respectively), the surrogate model trained on the foldswitch_set is much simpler, with just 11 rules, which makes it easier to interpret. We observed that the most disordered residues (DisoMine >= 0.897) are all predicted a transition (ambiguous behavior). Less disordered residues (DisoMine > 0.256) that present a low backbone rigidity (backbone <= 0.724 with DynaMine) are also classified as transition, as are residues with low backbone rigidity (backbone <= 0.754) and a high coil propensity (coil >= 0.505). The rest of the rules are often the combination of three or more biophysical features, with the disorder by DisoMine and backbone dynamics by DynaMine being the most prevalent ones, as already observed in the RF feature importance analysis (Figure 2).

3.2 Assessments on independent MFIB dataset

To assess to what extent the RF predictor can recognize the conditional fold of IDPs undergoing mutual folding-upon-binding, we assembled a validation set based on the MFIB database (Fichó et al., 2017) with structural filtering and removal of overlap with other training datasets (for details, see Methods). These proteins are quite different from the classical IDPs, as they are only disordered in the absence of their binding partner or under conditions that prevent their homo-oligomerization. Otherwise, they fold into compact domain-like structures. Thus, we expected to see an enrichment of the predicted ordered and ambiguous conformational class as opposed to the enrichment of the disordered classes.

For the residues in regions undergoing synergistic folding, the disordered class, without ambiguous folding propensity, was shown to be depleted in the output of the combined_RF predictor (<1%), while the ordered class was predicted to be the most represented (79.6%). The ambiguous class was predicted for 20%

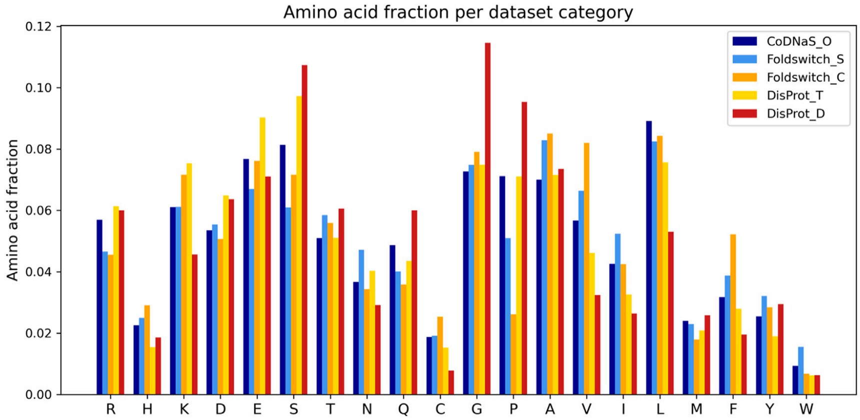


FIGURE 3
Amino acid fractions observed in the disprot_codnas_set for ordered residues (CoDNaS_O), transition (DisProt_T), and disordered (DisProt_D) and the foldswitch_set for fold-switching residues (Foldswitch_C) and residues that stay in the same fold (Foldswitch_S).

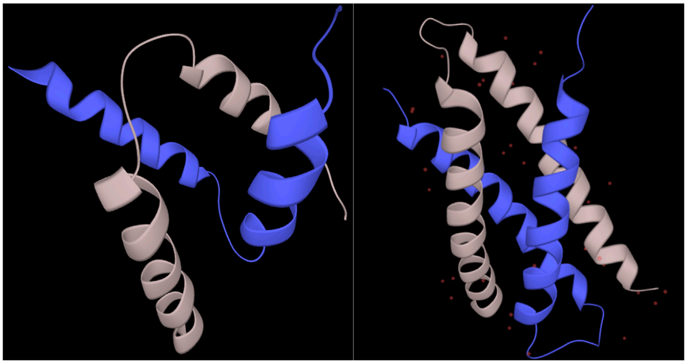


FIGURE 4
SinR and SH2B2 dimerization domains from MFIB (MF2120029, MF2100004). The SinR (left) dimerization domain (PDB:2YAL) is predicted to have only ambiguous residues, while the SH2B2 (right) dimerization domain (PDB:1Q2H) is predicted to be fully ordered based on the combined_RF model.

of cases, indicating that the folding mechanism of complexes in MFIB, in terms of biophysics, resembles folded domains. This resemblance between folded domains and mutually folded IDPs has already been recognized earlier from the structural and coevolution point of view (Iserte et al., 2020). A significant proportion of ambiguous behavior is still present, however, though fewer than the disorder-to-order transitions of IDPs upon binding or to metamorphic fold-switchers. For individual cases, predictions of regions with ambiguous conformations had significant variation. For example, the SinR dimerization domain of *B. subtilis* (MFIB:MF2120029; PDB: 2YAL) is predicted to have ambiguous confirmation with 94% coverage of the domain. On the other hand, the dimerization domain of the human SH2B adapter protein 2 (MFIB: MF2100004; PDB:1Q2H) is predicted to be 100% ordered despite the structural resemblance to the other case (Figure 4).

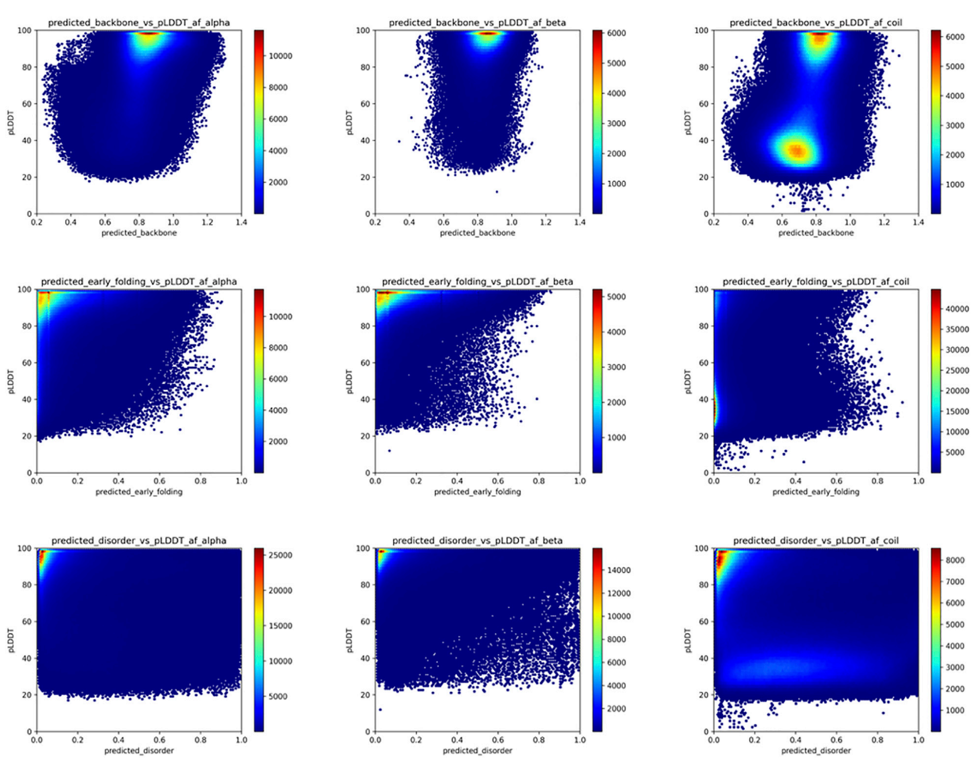


FIGURE 5
AlphaFold2 pLDDT versus backbone rigidity, early folding and disorder predictions for human proteome residues. Based on the "selected_human_dataset," we show heat maps of the relation between AlphaFold2 pLDDT value and the backbone rigidity (top), early folding (middle), and disorder (bottom) predictions for residues designed at alpha-helix (left), beta-strand (middle) and coil (right) by DSSP based on the AlphaFold2 models.

The complete prediction file is available from https://bitbucket.org/bio2byte/protein_ambiguity/.

3.3 Relation to AlphaFold2 human proteome models

AlphaFold2 (Jumper et al., 2021, p. 2, p. 2) can predict single low-energy conformations of proteins with unprecedented accuracy and provides excellent indications of the confidence with which this is done through the per-residue pLDDT values. However, possible conformational ambiguity is not well captured by the AlphaFold2 models (AlphaFold2 fails to predict protein fold switching—Chakravarty—2022—Protein Science—Wiley Online Library, no date), indicating the need to understand how the characteristics of these models relate to conformational ambiguity and dynamics. We, therefore, related the key biophysical predictions of the selected_human_set with the respective pLDDT values of the AlphaFold2 models, subdivided by secondary structure category in the model as determined by DSSP, to understand how these are related, and how this can give insights into the ambiguous residue categories. Figure 5 shows that for the backbone dynamics predictions (first row), the confidently predicted alpha-helix or beta-strand residues, with pLDDT scores close to 100%, have high predicted rigidity (>0.8 DynaMine score); for DynaMine, residues with values above 0.8 are expected to be well folded (Cilia et al., 2014). Residues with a coil classification according to DSSP are either similar to the secondary structure categories (pLDDT confident/backbone rigid), indicating folded residues that do not fall into

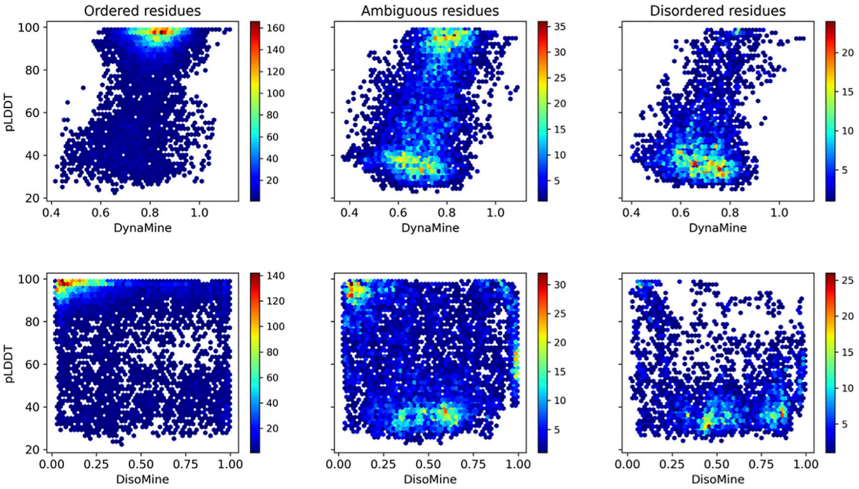


FIGURE 6
Relation between pLDDT score and backbone dynamics (top) and disorder (bottom) for the O/T/D classes from the disprot_codnas_set.

regular secondary structure categories, or they have low pLDDT confidence and are in the "context-dependent" (DynaMine scores between 0.69 and 0.80), or in the flexible region (<0.69). The pLDDT and DynaMine scores are, therefore, aligned, with high backbone dynamics (lower DynaMine scores) indicating multiple conformations correlating with AlphaFold2 predictions of lower confidence, as it is not able to confidently predict a single low-energy conformation for these residues. The early folding propensity predictions (Figure 5, second row) show that residues with increased early folding propensity are also typically residues predicted with high confidence by AlphaFold2, although AlphaFold2 cannot distinguish between these residues and ones that do not initiate folding pathways, as already indicated by other studies (Outeiral, Nissley, and Deane, 2022). Finally, for disorder predictions (Figure 5, third row), regions with high pLDDT are enriched with residues predicted to have disorder scores of 0 (no disorder), whereas residues predicted to be a coil by AlphaFold2 feature a low pLDDT region that has a wide dispersion of datapoints covering a range of disorder propensity values. Similar to backbone dynamics, this indicates residues that might have ambiguous conformational behavior.

When subdividing these plots in relation to our datasets that indicate ambiguous residues (Figure 6), these trends are more obvious. The ordered residues cluster at high pLDDT values (>80%) and high backbone rigidity (>0.8), the disordered residues at very low pLDDT values (<40%), and high backbone dynamics (<0.8). The ambiguous residues fall in between these categories, with many lower confidence pLDDT values between 80% and 40%, and backbone dynamics between 0.70–0.80, as well as significant overlap with the ordered and disordered categories. The disorder values confirm this trend, with few ordered residues predicted as having high disorder scores and most disordered residues correctly predicted with high disorder scores. The ambiguous residues again give an intermediate picture, with more residues having scores intermediate between the typical scores for order and disorder.

For the fold_switch_set only (Figure 7), there are interesting differences, especially the AlphaFold2 pLDDT scores, which tend to be below 90% for the residues that change conformation. The backbone dynamics also contain fewer high values, while more residues are predicted with high disorder.

3.4 Relation to post-translational modification data

Post-translational modifications (PTMs) of amino acid residues are important for regulation and can have a significant impact on protein conformation and function. Based on the ptm_set, which contains information for sumoylation, methylation, acetylation, ubiquitination, and phosphorylation, or a combination of these (Figure 8, log scale), we subdivided the observed PTMs by the different datasets. For the disprot_codnas_set, the majority of PTMs

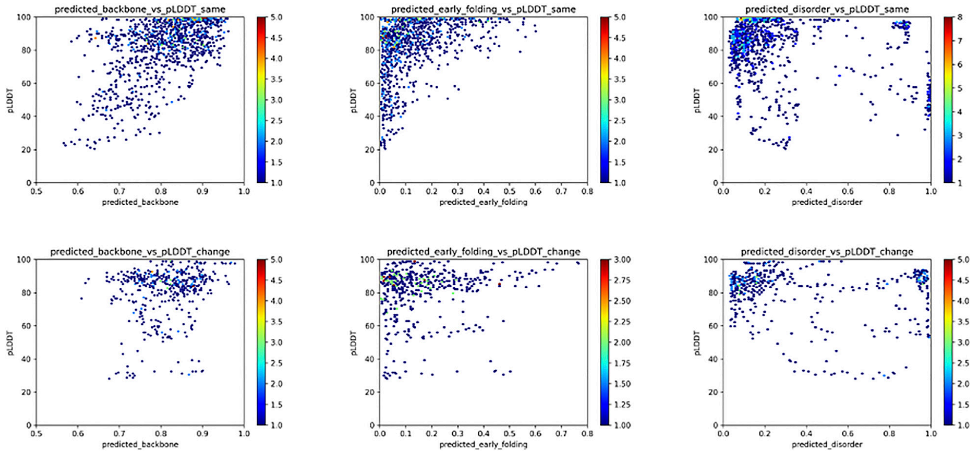


FIGURE 7
Relation between the pLDDT score and backbone dynamics (left), early folding (middle), disorder (right) for the fold_switch_set same (top row), and convert (bottom row) residues.

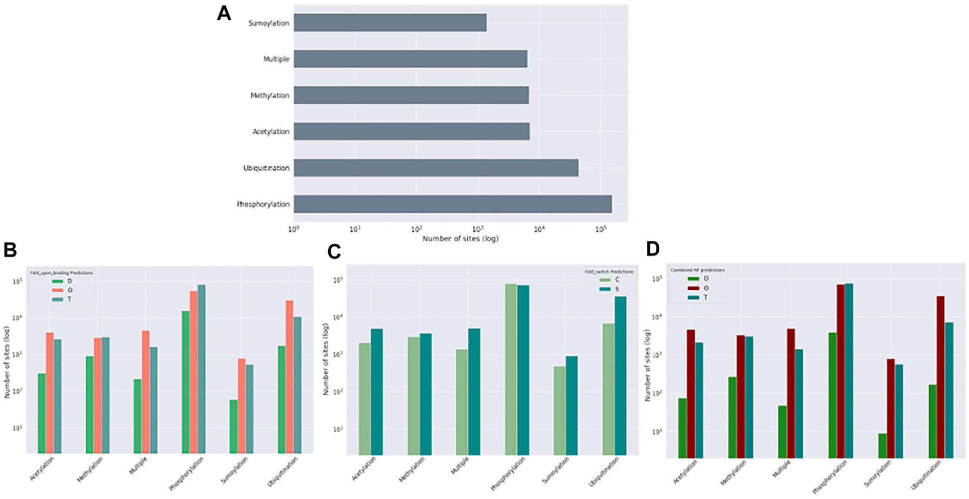


FIGURE 8
Post-translational modification (PTM) sites from the ptm_set in relation to datasets. The total number of included PTMs **(A)**, subdivided by disordered, ordered, and transition based on disprot_codnas_set **(B)**, by ordered and convert based on foldswitch_set **(C)**, and based on the combined order, disorder, and transition classes **(D)**.

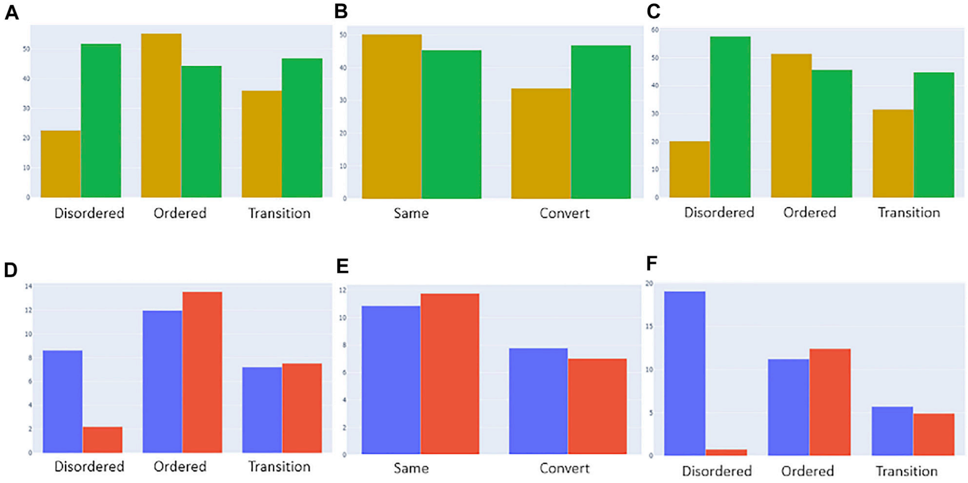


FIGURE 9
Categorization of deleterious (yellow) and benign (green) mutations from the "canonical_mut" dataset classified as disordered, ordered, and ambiguous based on models 1 [**(A)**, top left], 2 [**(B)**, top middle], and 3 [**(C)**, top right], respectively. The distributions were normalized by the total number of assigned ordered, disordered and ambiguous residues in the dataset. Categorization of germline (blue) and somatic (red) mutations from the "germline_somatic_deleterious" dataset classified as disordered, ordered, and ambiguous based on models 1 [**(D)**, bottom left], 2 [**(E)**, bottom middle], and 3 [**(F)**, bottom right], respectively. The distributions were normalized by the total number of assigned ordered, disordered and ambiguous residues in the dataset and the number of somatic/germline ratios for better comparison.

are observed in the order and transition classes, with phosphorylation overrepresented in residues with transition properties, and with ubiquitination and sumoylation underrepresented (Figure 8B). In the foldswitch_set, residues that remain in the same secondary structure state (S) have again increased ubiquitination and sumoylation compared to residues that convert (C), with a slight increase in acetylation and especially multiple modifications, indicating a possible role in fold-switching processes or more availability of these residues to be modified by smaller PTMs. The trends for the combined_set are very similar to the disprot_codnas_set, which constitutes the bulk of the data.

3.5 Relation to deleterious amino acid variants

We also investigated whether residues in ambiguous regions, again given their likely role in conformational rearrangements and allostery, are more likely to contain deleterious or benign mutations, as classified in the canonical_mut dataset. Figure 9A shows that for the disprot_codnas_set (RF model 1), the ordered residues contain, as expected, relatively more deleterious mutations. Although the ambiguous residues contain more benign mutations, they still contain a high proportion of deleterious mutations, especially compared to the ratio observed for disordered residues. This situation is similar to somatic versus germline cancer mutations (Figure 9D). For the foldswitch_set (RF model 2), the ambiguous metamorphic residues contain a higher amount of deleterious mutations than the residues that retain their secondary structure (Figure 9B), whereas there is no difference between somatic versus germline mutations (Figure 9E). For the combined RF model 3, the trends are very similar to RF model 1 (Figure 9C–F).

4 Discussion

In this exploratory analysis, we use two datasets that try to capture amino acid residues in proteins that display different "ambiguous" behaviors either by folding-upon-binding (disprot_codnas_set) or by changing secondary structure in metamorphic proteins (foldswitch_set). This definition of "ambiguous" residues is highly relevant given the ready availability of predicted AlphaFold2 protein structure models with qualities comparable to experimentally derived structures. Given the dynamic nature of proteins, and their capacity to change conformation and transmit signals through allostery

(Tompa, 2014, 2016), annotations of the AlphaFold2 models indicate where such conformational changes are more likely to happen, which will help in interpreting such models. Our results indicate that AlphaFold2, based on the per-residue pLDDT prediction confidence values, captures ordered and disordered residues very well, and while for ambiguous regions intermediate pLDDT values are observed, many of these ambiguous residues fall into the "traditional" ordered or disordered regions (Figure 6). The RF models we created and their interpretation show that sequence-predicted disorder is the most important factor predicting fold switching residues (from order to order), as well as folding-upon-binding (from disorder to order), with backbone dynamics and early folding also important for the last category. Specific amino acids are also a likely factor, such as valine and phenylalanine for the fold switching residues. Although the recognition by the combined_RF model of the MFIB dataset, which contains dimers that form domain-like structures, is of limited sensitivity (see https://bitbucket.org/bio2byte/protein_ambiguity/), there are indications that ambiguous residues can also be picked up in these cases. This illustrates the complexity of protein behavior in relation to its (local) environment; in this case, and expressed in terms of ambiguous behavior, the local sequence context of the protein is strongly geared toward order, but enough ambiguous residues are present that the individual proteins cannot fold.

Previous AlphaFold2-related studies in this area have given similar indications. AlphaFold2 is a good predictor of intrinsically disordered regions (IDRs) based on the CAID PDB-DisProt dataset (Piovesan, Monzon, and Tosatto, 2022), a study on conditionally folded IDRs (Alderson et al., 2022) showed that many IDRs are in the high ($70 \leq \times < 90$) or very high (≥ 90) pLDDT regions, similar to what we report, with enrichment in helical conformations, and with long, extended single α-helix domains not stabilized by tertiary contacts identified. For a subset of IDRs that fold under specific conditions and have been extensively characterized by NMR spectroscopy, the IDRs resemble the conformation of the folded state, even if there is no stable secondary structure observed with only a fractional preference to populate secondary structures from the experimental NMR data. The combination of higher relative solvent accessibility in the AlphaFold2 models, which indicates a lack of overall structure, and high pLDDT scores, which indicate confident structure predictions, does, however, seem to be a good indicator of regions with a tendency for ambiguous behavior (Piovesan, Monzon, and Tosatto, 2022). These results show again that AlphaFold2 is excellent at defining a single low-energy state for a given protein sequence if it exists, but that the context of the protein and possible ambiguous behavior is more difficult to capture. Indeed, in relation to conformational diversity as observed in the PDB from apo-holo pairs of conformers for the same protein (Saldaño et al., 2022), AlphaFold2 predicts the holo form in ~70% of cases but is unable to capture both states. As the conformational diversity between the apo/holo states increases, its prediction performance also worsens. A similar picture is observed for proteins that can switch folds (AlphaFold2 fails to predict protein fold switching—Chakravarty—2022—Protein Science—Wiley Online Library, no date), with 94% of AlphaFold2 predictions capturing one experimentally determined conformation but not the other, and with moderate-to-high pLDTT scores for 74% of fold-switching residues, similar to our study. Finally, although AlphaFold2 and RoseTTAfold models seem to carry overall foldability information (Liu, Wu, and Chen, 2022), the folding process itself is not well captured (Outeiral, Nissley, and Deane, 2022), if at all.

Overall, it remains very difficult to capture the dynamic properties of proteins; despite the availability of molecular dynamics simulations of increasing length, limited direct dynamics measurements from NMR and other structural biology approaches, and the observed conformational diversity in the PDB, the complexity of possible protein movements and their likelihood within the *in vivo* environment of proteins, in general, precludes the generation of relevant all-encompassing datasets. The increasing amount of data that indirectly indicates such behavior, from mass spectrometry proteomics (Britt, Cragnolini, and Thalassinos, 2021) as well as from evolutionary and disease mutation sources, will be in this respect invaluable, as already indicated in our limited study. The challenge here lies in interconnecting the various diverse data sources and analyzing the resulting complex information, which is beyond direct human understanding and requires machine learning approaches, preferably interpretable so that concepts and first principles can be derived from them. Furthermore, methodology development in the more traditional sense is also key, for example, improved ensemble representations of proteins and especially IDRs, as already indicated in other studies such as the ones discussed here (Alderson et al., 2022; AlphaFold2 fails to predict protein fold switching—Chakravarty—2022—Protein Science—Wiley Online Library, no date), as well as more accurate sequence-based predictors, with the combination of structure and sequence-based approaches likely giving the most relevant results.

5 Conclusion

In our view, it is essential that we move away from the two-state view of proteins (one single well-defined static fold, or complete disorder) to a more nuanced probabilistic view, where the "probability space" of proteins is defined—as the possible states of a protein can adopt. The definition of the different kinds of ambiguity observed in protein behavior, and their interpretation is an important step to help the field move in this direction. Ongoing ELIXIR implementation projects, for example, are also focusing on related topics, highlighting the community's need for this kind of probabilistic interpretation of protein behavior. We hope that the datasets and analyses we assembled here provide additional reference points to further explore and define residues with ambiguous behavior in proteins.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://bitbucket.org/bio2byte/protein_ambiguity/.

Author contributions

JR-M trained the RF models; IG contributed the RF interpretation code; JR-M, TL, RP, PR, and KT contributed datasets; JG-G contributed the analysis of the AlphaFold2 models; KT contributed the analysis of deleterious mutants; PR contributed the analysis of PTMs; WV provided the manuscript concept and organization of results; JR-M, TL, JG-G, DB, BD, KT, PR, MS-F, and WV contributed to the writing.

Funding

This study was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (813239 to JR-M and JG-G); Research Foundation Flanders (FWO) (G.032816N to BD and G.028821N to DB); Vrije Universiteit Brussel Research Council under the Interdisciplinary Research Program TumorScope (IRP20 to KT); Tempus Public Foundation postdoctoral fellowships (158534 and 166538 to RP); and National Research, Development and Innovation Office research grant (FK128133 to RP). Funding for open access charge: European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (813239).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abriata, L. A., and Dal Peraro, M. (2021). Assessment of transferable forcefields for protein simulations attests improved description of disordered states and secondary structure propensities, and hints at multi-protein systems as the next challenge for optimization. *Comput. Struct. Biotechnol. J.* 19, 2626–2636. doi:10.1016/j.csbj.2021.04.050

Adamczak, R., Porollo, A., and Meller, J. (2004). Accurate prediction of solvent accessibility using neural networks–based regression. *Proteins* 56 (4), 753–767. doi:10.1002/prot.20176

Alderson, T. R., Pritišanac, I., Moses, A. M., and Forman-Kay, J. D. (2022). *Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2*preprint. *Biochemistry*. doi:10.1101/2022.02.18.481080

Armstrong, K. M., Piepenbrink, K. H., and Baker, B. M. (2008). Conformational changes and flexibility in T-cell receptor recognition of peptide–MHC complexes. *Biochem. J.* 415 (Pt 2), 183–196. doi:10.1042/BJ20080850

Bah, A., and Forman-Kay, J. D. (2016). Modulation of intrinsically disordered protein function by post-translational modifications. *J. Biol. Chem.* 291 (13), 6696–6705. doi:10.1074/jbc.R115.695056

Bai, B., Vanderwall, D., Li, Y., Wang, X., Poudel, S., Wang, H., et al. (2021). Proteomic landscape of Alzheimer's disease: novel insights into pathogenesis and biomarker discovery. *Mol. Neurodegener.* 16 (1), 55. doi:10.1186/s13024-021-00474-z

Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 35, D301–D303. Database issue). doi:10.1093/nar/gkl971

Bolognesi, B., Lorenzo Gotor, N., Dhar, R., Cirillo, D., Baldrighi, M., Tartaglia, G. G., et al. (2016). A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep.* 16 (1), 222–231. doi:10.1016/j.celrep.2016.05.076

Bonucci, A., Palomino-Schatzlein, M., Malo de Molina, P., Arbe, A., Pierattelli, R., Rizzuti, B., et al. (2021). Crowding effects on the structure and dynamics of the intrinsically disordered nuclear chromatin protein NUPR1. *Front. Mol. Biosci.* 8, 684622. doi:10.3389/fmolb.2021.684622

Britt, H. M., Cragnolini, T., and Thalassinos, K. (2021). Integration of mass spectrometry data for structural biology. *Chem. Rev.* 122 (8), 7952–7986. doi:10.1021/acs.chemrev.1c00356

Chang, M. T., Bhattarai, T. S., Schram, A. M., Bielski, C. M., Donoghue, M. T. A., Jonsson, P., et al. (2018). Accelerating discovery of functional mutant alleles in cancer. *Cancer Discov.* 8 (2), 174–183. doi:10.1158/2159-8290.CD-17-0321

Cheng, Y., Oldfield, C. J., Meng, J., Romero, P., Uversky, V. N., and Dunker, A. K. (2007). Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* 46 (47), 13468–13477. doi:10.1021/bi7012273

Childers, M. C., and Daggett, V. (2018). Validating molecular dynamics simulations against experimental observables in light of underlying conformational ensembles. *J. Phys. Chem. B* 122 (26), 6673–6689. doi:10.1021/acs.jpcb.8b02144

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2013). From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.* 4, 2741. doi:10.1038/ncomms3741

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2014). The DynaMine webserver: Predicting protein dynamics from sequence. *Nucleic Acids Res.* 42 (W1), W264–W270. doi:10.1093/nar/gku270

Cohen, W. W. (1995). Fast effective rule induction, *Mach. Learn. Proc.*, 1995, 115–123. San Francisco (CA): Morgan Kaufmann. doi:10.1016/B978-1-55860-377-6.50023-2

Daggett, V., and Fersht, A. R. (2003). Is there a unifying mechanism for protein folding? *Trends biochem. Sci.* 28 (1), 18–25. doi:10.1016/s0968-0004(02)00012-9

Dass, R., Mulder, F. A., and Nielsen, J. T. (2020). ODiNPred: Comprehensive prediction of protein order and disorder. *Sci. Rep.* 10 (1), 14780. doi:10.1038/s41598-020-71716-1

De Gieter, S., Konijnenberg, A., Talavera, A., Butterer, A., Haesaerts, S., De Greve, H., et al. (2014). The intrinsically disordered domain of the antitoxin phd chaperones the toxin doc against irreversible inactivation and misfolding. *J. Biol. Chem.* 289 (49), 34013–34023. doi:10.1074/jbc.M114.572396

DeForte, S., and Uversky, V. N. (2016). Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. *Protein Sci.* 25 (3), 676–688. A Publication of the Protein Society. doi:10.1002/pro.2864

Dobson, C. M. (2019). Biophysical techniques in structural biology. *Annu. Rev. Biochem.* 88, 25–33. doi:10.1146/annurev-biochem-013118-111947

Dobson, C. M. (2003). Protein folding and misfolding. *Nature* 426 (6968), 884–890. doi:10.1038/nature02261

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21 (16), 3433–3434. doi:10.1093/bioinformatics/bti541

Eibe, F., Hall, A., and Witten, I. H. (2016). *The WEKA workbench. Online appendix for "data mining: Practical machine learning tools and techniques"*. 4 Edn. Burlington, MA: Morgan Kaufmann.

Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 22 (10), 1302–1306. doi:10.1038/nbt1012

Fichó, E., Remenyi, I., Simon, I., and Meszaros, B. (2017). Mfib: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* 33 (22), 3682–3684. doi:10.1093/bioinformatics/btx486

Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., et al. (2008). The Catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.* Chapter 10, Unit 10.11. doi:10.1002/0471142905.hg1011s57

Gall, T. L., Romero, P. R., Cortese, M. S., Uversky, V. N., and Dunker, A. K. (2007). Intrinsic disorder in the protein Data Bank. *J. Biomol. Struct. Dyn.* 24 (4), 325–342. doi:10.1080/07391102.2007.10507123

Gerez, J. A., Prymaczok, N. C., and Riek, R. (2020). In-cell NMR of intrinsically disordered proteins in mammalian cells. *Methods Mol. Biol.* 2141, 873–893. doi:10.1007/978-1-0716-0524-0_45

Glazer, D. S., Radmer, R. J., and Altman, R. B. (2009). Improving structure-based function prediction using molecular dynamics. *Structure* 17 (7), 919–929. doi:10.1016/j.str.2009.05.010

Hanson, J., Litfin, T., Paliwal, K., and Zhou, Y. (2020). Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning. *Bioinforma. Oxf. Engl.* 36 (4), 1107–1113. doi:10.1093/bioinformatics/btz691

Hanson, J., Paliwal, K. K., Litfin, T., and Zhou, Y. (2019). SPOT-Disorder2: Improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics Proteomics Bioinforma.* 17 (6), 645–656. doi:10.1016/j.gpb.2019.01.004

Hatos, A., Hajdu-Soltesz, B., Monzon, A. M., Palopoli, N., Alvarez, L., Aykac-Fas, B., et al. (2020). DisProt: Intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* 48 (D1), D269-D276. doi:10.1093/nar/gkz975

Hilger, D., Masureel, M., and Kobilka, B. K. (2018). Structure and dynamics of GPCR signaling complexes. *Nat. Struct. Mol. Biol.* 25 (1), 4–12. doi:10.1038/s41594-017-0011-7

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* 43, D512–D520. Database issue). doi:10.1093/nar/gku1267

Horvath, A., Miskei, M., Ambrus, V., Vendruscolo, M., and Fuxreiter, M. (2020). Sequence-based prediction of protein binding mode landscapes. *PLoS Comput. Biol.* 16 (5), e1007864. doi:10.1371/journal.pcbi.1007864

Hsin, J., Strumpfer, J., Lee, E. H., and Schulten, K. (2011). Molecular origin of the hierarchical elasticity of titin: simulation, experiment, and theory. *Annu. Rev. Biophys.* 40, 187–203. doi:10.1146/annurev-biophys-072110-125325

Huang, J., and MacKerell, A. D. (2018). Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 48, 40–48. doi:10.1016/j.sbi.2017.10.008

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., et al. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14 (1), 71–73. doi:10.1038/nmeth.4067

Huang, K.-Y., Lee, T. Y., Kao, H. J., Ma, C. T., Lee, C. C., Lin, T. H., et al. (2019). dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic Acids Res.* 47 (D1), D298-D308. doi:10.1093/nar/gky1074

Hummer, G., and Köfinger, J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* 143 (24), 243150. doi:10.1063/1.4937786

Hunkapiller, M. W., Strickler, J. E., and Wilson, K. J. (1984). Contemporary methodology for protein structure determination. *Science* 226, 304–311. doi:10.1126/science.6385254

Iserte, J. A., Lazar, T., Tosatto, S. C. E., Tompa, P., and Marino-Buslje, C. (2020). Chasing coevolutionary signals in intrinsically disordered proteins complexes. *Sci. Rep.* 10 (1), 17962. doi:10.1038/s41598-020-74791-6

Jones, D. T., and Cozzetto, D. (2015). DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinforma. Oxf. Engl.* 31 (6), 857–863. doi:10.1093/bioinformatics/btu744

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292 (2), 195–202. doi:10.1006/jmbi.1999.3091

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12), 2577–2637. doi:10.1002/bip.360221211

Kagami, L. P., Orlando, G., Raimondi, D., Ancien, F., Dixit, B., Gavalda-Garcia, J., et al. (2021). b2bTools: online predictions for protein biophysical features and their conservation. *Nucleic Acids Res.* 49 (W1), W52–W59. doi:10.1093/nar/gkab425

Karplus, M., and Kuriyan, J. (2005). Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S. A.* 102 (19), 6679–6685. doi:10.1073/pnas.0408930102

Katuwawala, A., Peng, Z., Yang, J., and Kurgan, L. (2019). Computational prediction of MoRFs, short disorder-to-order transitioning protein binding regions. *Comput. Struct. Biotechnol. J.* 17, 454–462. doi:10.1016/j.csbj.2019.03.013

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sonderby, C. K., et al. (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* 87 (6), 520–527. doi:10.1002/prot.25674

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46 (D1), D1062-D1067. doi:10.1093/nar/gkx1153

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (20031993)., 11. London, England, 1453–1459. doi:10.1016/j.str.2003.10.002Protein disorder prediction: Implications for structural proteomics*Structure*11

Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature* 433 (7022), 128–132. doi:10.1038/nature03199

Liu, S., Wu, K., and Chen, C. (2022). *The computational models of AlphaFold2 and RoseTTAfold carry protein foldability information*. preprint. *Bioinformatics*. doi:10.1101/2022.01.27.477978

Magnan, C. N., and Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinforma. Oxf. Engl.* 30 (18), 2592–2597. doi:10.1093/bioinformatics/btu352

Malhis, N., Jacobson, M., and Gsponer, J. (2016). MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res.* 44 (W1), W488–W493. doi:10.1093/nar/gkw409

McLaughlin, R. J., Spindler, M. P., van Lummel, M., and Roep, B. O. (2016). Where, how, and when: Positioning posttranslational modification within type 1 diabetes pathogenesis. *Curr. Diab. Rep.* 16 (7), 63. doi:10.1007/s11892-016-0752-4

Mészáros, B., Erdos, G., and Dosztányi, Z. (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46 (W1), W329-W337. doi:10.1093/nar/gky384

Mir, S., Alhroub, Y., Anyango, S., Armstrong, D. R., Berrisford, J. M., Clark, A. R., et al. (2018). PDBe: towards reusable data delivery infrastructure at protein data bank in europe. *Nucleic Acids Res.* 46 (D1), D486-D492. doi:10.1093/nar/gkx1070

Miskei, M., Horvath, A., Vendruscolo, M., and Fuxreiter, M. (2020). Sequence-based prediction of fuzzy protein interactions. *J. Mol. Biol.* 432 (7), 2289–2303. doi:10.1016/j.jmb.2020.02.017

Mizianty, M. J., Uversky, V., and Kurgan, L. (2014). Prediction of intrinsic disorder in proteins using MFDp2. *Methods Mol. Biol.* 1137, 147–162. doi:10.1007/978-1-4939-0366-5_11

Monzon, A. M., Rohr, C. O., Fornasari, M. S., and Parisi, G. (2016). CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database.* 2016, baw038. doi:10.1093/database/baw038

Mu, J., Liu, H., Zhang, J., Luo, R., and Chen, H. F. (2021). Recent force field strategies for intrinsically disordered proteins. *J. Chem. Inf. Model.* 61 (3), 1037–1047. doi:10.1021/acs.jcim.0c01175

Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., and Dunker, A. K. (2005). Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 44 (37), 12454–12470. doi:10.1021/bi050736e

Orioli, S., Larsen, A. H., Bottaro, S., and Lindorff-Larsen, K. (2020). "Chapter Three - how to learn from inconsistencies: Integrating molecular simulations with experimental data, *Prog. Mol. Biol. Transl. Sci.*, 170, 123–176. Academic Press. doi:10.1016/bs.pmbts.2019.12.006

Orlando, G., Raimondi, D., Codice, F., Tabaro, F., and Vranken, W. (2022). Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. *J. Mol. Biol.* 434 (12), 167579. doi:10.1016/j.jmb.2022.167579

Orlando, G., Raimondi, D., Tabaro, F., Codice, F., Moreau, Y., and Vranken, W. F. (2019). Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates. *Bioinforma. Oxf. Engl.* 35 (22), 4617–4623. doi:10.1093/bioinformatics/btz274

Orlando, G., Silva, A., Macedo-Ribeiro, S., Raimondi, D., and Vranken, W. (2020). Accurate prediction of protein beta-aggregation with generalized statistical potentials. *Bioinforma. Oxf. Engl.* 36 (7), 2076–2081. doi:10.1093/bioinformatics/btz912

Outeiral, C., Nissley, D. A., and Deane, C. M. (2022). Current structure predictors are not learning the physics of protein folding. *Bioinformatics* 38, 1881–1887. doi:10.1093/bioinformatics/btab881

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi:10.48550/arXiv.1201.0490

Piana, S., Robustelli, P., Tan, D., Chen, S., and Shaw, D. E. (2020). Development of a force field for the simulation of single-chain proteins and protein–protein complexes. *J. Chem. Theory Comput.* 16 (4), 2494–2507. doi:10.1021/acs.jctc.9b00251

Pintado, C., Santos, J., Iglesias, V., and Ventura, S. (2021). SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins. *Bioinformatics* 37 (11), 1602–1603. doi:10.1093/bioinformatics/btaa909

Piovesan, D., Monzon, A. M., and Tosatto, S. C. E. (2022). Intrinsic protein disorder, conditional folding and AlphaFold2. *bioRxiv* 2022, 482768. doi:10.1101/2022.03.03.482768

Raimondi, D., Orlando, G., Michiels, E., Pakravan, D., Bratek-Skicki, A., Van Den Bosch, L., et al. (2021)., 37. Oxford, England, 3473–3479. doi:10.1093/bioinformatics/btab350*In-silico* prediction of *in-vitro* protein liquid-liquid phase separation experiments outcomes with multi-head neural attention*Bioinformatics*

Raimondi, D., Orlando, G., Pancsa, R., Khan, T., and Vranken, W. F. (2017). Exploring the sequence-based prediction of folding initiation sites in proteins. *Sci. Rep.* 7 (1), 8826. doi:10.1038/s41598-017-08366-3

Ramasamy, P., Turan, D., Tichshenko, N., Hulstaert, N., Vandermarliere, E., Vranken, W., et al. (2020). Scop3P: A comprehensive resource of human phosphosites within their full context. *J. Proteome Res.* 19 (8), 3478–3486. doi:10.1021/acs.jproteome.0c00306

Saldaño, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Velez Rueda, A. J., et al. (2022). 'Impact of protein conformational diversity on AlphaFold predictions'. *Bioinformatics* 38 (10), 2742–2748. doi:10.1093/bioinformatics/btac202

Santos, J., Iglesias, V., Pintado, C., Santos-Suarez, J., and Ventura, S. (2020a). DispHred: A server to predict pH-dependent order–disorder transitions in intrinsically disordered proteins. *Int. J. Mol. Sci.* 21 (16), 5814. doi:10.3390/ijms21165814

Santos, J., Iglesias, V., Santos-Suárez, J., Mangiagalli, M., Brocca, S., Pallares, I., et al. (2020b). pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity. *Cells* 9 (1), 145. doi:10.3390/cells9010145

Shental-Bechor, D., and Levy, Y. (2008). Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. *Proc. Natl. Acad. Sci. U. S. A.* 105 (24), 8256–8261. doi:10.1073/pnas.0801340105

Singh, J., Litfin, T., Paliwal, K., Singh, Jaspreet, Singh, J., Hanumanthappa, A. K., et al. (2021). SPOT-1D-Single: Improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning. *Bioinformatics* 37 (20), 3464–3472. doi:10.1093/bioinformatics/btab316

Song, L., and Luo, Z.-Q. (2019). Post-translational regulation of ubiquitin signaling. *J. Cell Biol.* 218 (6), 1776–1786. doi:10.1083/jcb.201902074

Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015). The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* 427 (2), 478–490. doi:10.1016/j.jmb.2014.09.026

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23 (10), 1282–1288. doi:10.1093/bioinformatics/btm098

Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M. P., Vivancos, A., Rovira, A., et al. (2018). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* 10 (1), 25. doi:10.1186/s13073-018-0531-8

Tian, C., Kasavajhala, K., Belfon, K. A. A., Raguette, L., Huang, H., Migues, A. N., et al. (2020). ff19SB: Amino-Acid-Specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.* 16 (1), 528–552. doi:10.1021/acs.jctc.9b00591

Tompa, P. (2014). Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery. *Chem. Rev.* 114 (13), 6715–6732. doi:10.1021/cr4005082

Tompa, P. (2016). The principle of conformational signaling. *Chem. Soc. Rev.* 45 (15), 4252–4284. doi:10.1039/c6cs00011h

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596 (7873), 590–596. doi:10.1038/s41586-021-03828-1

UniProt Consortium, The (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1), D480–D489. doi:10.1093/nar/gkaa1100

Uversky, V. N. (2019). "Chapter One - protein intrinsic disorder and structure-function continuum, *Prog. Mol. Biol. Transl. Sci.*, 166, 1–17. Academic Press(Dancing protein clouds: Intrinsically disordered proteins in health and disease, Part A). doi:10.1016/bs.pmbts.2019.05.003

Uversky, V. N. (2013). Unusual biophysics of intrinsically disordered proteins. *Biochim. Biophys. Acta* 1834 (5), 932–951. doi:10.1016/j.bbapap.2012.12.008

Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., et al. (2018). Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *eLife* 7, e31486. doi:10.7554/eLife.31486

Vu, L. D., Gevaert, K., and De Smet, I. (2018). Protein language: Post-translational modifications talking to each other. *Trends Plant Sci.* 23 (12), 1068–1080. doi:10.1016/j.tplants.2018.09.004

Walsh, I., Martin, A. J. M., Di Domenico, T., and Tosatto, S. C. E. (2012). ESpritz: accurate and fast prediction of protein disorder. *Bioinforma. Oxf. Engl.* 28 (4), 503–509. doi:10.1093/bioinformatics/btr682

Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014). Pasta 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.* 42, W301–W307. Web Server issue). doi:10.1093/nar/gku399

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinforma. Oxf. Engl.* 20 (13), 2138–2139. doi:10.1093/bioinformatics/bth195

Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K., and Uversky, V. N. (2010). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta* 1804 (4), 996–1010. doi:10.1016/j.bbapap.2010.01.011

Yan, J., Dunker, A. K., Uversky, V. N., and Kurgan, L. (2016). Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.* 12 (3), 697–710. doi:10.1039/c5mb00640f

Yang, Y. I., Shao, Q., Zhang, J., Yang, L., and Gao, Y. Q. (2019). Enhanced sampling in molecular dynamics. *J. Chem. Phys.* 151 (7), 070902. doi:10.1063/1.5109531

Zapletal, V., Mladek, A., Melkova, K., Lousa, P., Nomilner, E., Jasenakova, Z., et al. (2020). Choice of force field for proteins containing structured and intrinsically disordered regions. *Biophys. J.* 118 (7), 1621–1633. doi:10.1016/j.bpj.2020.02.019

Zhang, T., Faraggi, E., Li, Z., and Zhou, Y. (2017). Intrinsic disorder and semi-disorder prediction by SPINE-D. *Methods Mol. Biol.* 1484, 159–174. doi:10.1007/978-1-4939-6406-2_12

Zhang, T., Faraggi, E., Li, Z., and Zhou, Y. (2013). Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell biochem. Biophys.* 67 (3), 1193–1205. doi:10.1007/s12013-013-9638-0

Chapter 7

Conclusion and perspectives

With the goal of uncovering protein flexibility and dynamics, this thesis investigated proteins in various physiological contexts, focusing on the effect of N-glycosylation and in-solution biophysical behaviour. Through the use of diverse methods including MD simulations, NMA, and NMR based metrics of flexibility, we examined a variety of proteins to achieve these insights.

7.1 Effect of glycosylation on protein dynamics

Glycosylation is a common and generally conserved PTM essential for a protein's biophysical, kinetic, and thermodynamic stability. However, understanding glycoprotein dynamics remains a 'black box' due to the challenges of resolving glycans experimentally, which stem from their high flexibility, disease-specificity, and microheterogeneity. As a result, glycans are often neglected in studies related to protein dynamics. This thesis aimed to explore the effects of glycosylation and mutations on AGP, a heavily glycosylated plasma protein, and to address the complex impact of glycans on protein's backbone dynamics. Through triplicate MD simulations (100 ns per system), which ensured efficient sampling, it was demonstrated that **glycosylation reduces local backbone flexibility at the glycosylation site while increasing flexibility in structurally distant regions**. One proposed hypothesis was that this allosteric effect results from a **reordering of dynamics** to compensate for the potential entropy loss introduced by PTMs.

7.2 Combined effect of glycosylation and mutations on protein's conformational dynamics

Mutations were shown to alter the local flexibility of the protein through their long-range conformational effects. The combined effects of mutations and glycosylation on AGP's behaviour were found to be complex. Glycosylated mutants exhibited similar backbone dynamics, yet no group of mutants demonstrated consistent biophysical effects. Mutation-induced changes in glycan flexibility also resulted in significant variations in solvent accessibility among different glycosylated mutants, depending on probe size. **Among the analyzed mutants, the R101W variant emerged as the mutant with the most pronounced impact of glycosylation and mutations on AGP's conformational dynamics**. Strategically positioned near three out of five glycosylation sites—site I (N33), site V (N103), and site III (N72)—this mutation emerged as structurally significant, profoundly influencing glycan interactions and, consequently, the overall dynamics of both the glycans and the protein. This mutation site enabled the mutant to more effectively disrupt interactions, indicating a potential immunomodulatory role in cancer-related processes. These findings underscore the necessity of considering in vivo glycosylation patterns when assessing the impact of mutations. In conclusion, the combination of glycosylation and mutations can create a more varied conformational landscape for a protein than the 20 amino acids alone, potentially affecting its function and ligand-binding properties.

7.3 Predicting the effect of glycans on protein's backbone dynamics

In this thesis, we used $\Delta RMSF$, which describes the difference in RMSF between the glycosylated and unglycosylated systems, as a metric to assess the change in flexibility for the protein fragments of interest due to glycosylation. **This metric can be invaluable to assess the effect of glycans on protein conformations**. The metric provided a detailed understanding of the specific regions most impacted by glycan addition, proving useful for pinpointing changes in flexibility around glycosylation sites and functionally important areas, such as the ligand-binding site of AGP.

7.4 Convoluting experimental and computational metrics of protein flexibility

To address the second research question related to the relationship between protein flexibility predicted by computational methods and observed through experimental techniques, a large-scale analysis of AlphaFold2 models and their NMR ensembles was carried out, comparing multiple metrics of protein flexibility and also assessing their relationship with the pLDDT score. The pLDDT score for a residue estimates how closely the predicted structure matches an experimentally determined protein structure, with higher scores indicating that the residue is likely well-folded, ordered, and likely to be rigid. On the other hand, these flexibility metrics obtained from MD simulations, S^2_{RCI}, S^2 order parameters, and NMA-derived RMSF carry information on distinct timescales. The NMR-based metrics capture dynamics on upto low ms timescales as an ensemble average, while MD simulations typically capture dynamics upto 1 μs, and NMA-derived RMSF, on the other hand, provides insights into the low-frequency, collective motions accessible to the protein, without absolute timescales. Thus, **for a protein, if its dynamics are accurately captured, these metrics can provide the correct distributions of flexibility**, or close to correct in the case of NMA. Since the dynamics of a protein are structure-encoded, any inaccuracies in the structural model or ensemble can impact the accuracy of the flexibility estimations or predictions. In simple terms, coils in a protein are more flexible and typically found in disordered or partially disordered regions and more challenging to emulate and interpret in flexibility predictions. Any deviations from coils to more rigid helices and sheets between an experimental structure and its predicted model could significantly impact the accuracy of capturing the dynamics, as observed in AlphaFold2 and their NMR models. When mismatches occur in well-folded regions like helices or sheets, AlphaFold2 typically provides more reliable secondary structure predictions than NMR ensembles. Even with non-unique assignments, AlphaFold2 often captures rigid in-solution helices and sheets more accurately, which may explain why AlphaFold2 models align better with experimental NMR data than the calculated NMR ensembles. The deviations or mismatches in the NMR ensembles and AlphaFold2 assignments are indicators of dynamic behaviour and multiple conformations in solution. The hard boundary between order and disorder in the rigid, well-folded regions of the protein is easily identified by NMA-derived RMSF, S^2_{RCI}, S^2 order parameters, and high pLDDT values, at the residue level. However, at the protein level, eveything becomes more complex. The interpretation of regions with lower pLDDT values depends on whether they link folded regions of the protein, as there is no direct correlation between NMA RMSF and experimentally determined dynamics. However, this correlation improves when NMR ensemble models are used as input for the NMA. For the second research question, it can be concluded that while **AlphaFold2 can identify a clear order/disorder boundary, it does not capture the dynamic behaviour of individual residues or the gradations in protein dynamics. NMA on the AlphaFold2 models equally does not capture such information.** Experimental techniques, such as NMR, and more detailed computational approaches like MD simulations, therefore, remain essential for accurately assessing protein motions and conformational states.

7.5 AlphaFold vs NMR: Implications for capturing protein dynamics

With the widespread accessibility of protein structures enabled by AlphaFold, experimentalists and biologists are increasingly relying on predicted structures of their proteins of interest, often treating high pLDDT scores as definitive and accurate. While low pLDDT values warrant scrutiny and validation, high pLDDT scores are frequently accepted without question, despite the need for critical evaluation in certain contexts. Therefore, the key implication is that AlphaFold2 or AlphaFold3's high pLDDT values indicate reliable structural models but do not always reflect true conformational dynamics. This raises important questions:

1. How should then the accuracy of AlphaFold structures be interpreted in the context of dynamics?
2. How should the results of this thesis be utilized, especially when no experimental NMR structures are available for comparison?

To address the first question, in cases where NMR structures are available for comparison with AlphaFold models, a straightforward comparison of secondary structures predicted by AlphaFold with those obtained from NMR using tools like STRIDE can highlight discrepancies between the experimental and predicted models. Additionally, experimental data (e.g., NMR S^2 order parameters) and computational flexibility measures, such as RMSF from NMA or MD simulations, can further reveal instances where AlphaFold underestimates protein dynamics or overrepresent certain secondary structures. Furthermore, data from X-ray crystallography and cryo-EM can further strengthen the validation process, aiding in the interpretation of the accuracy of AlphaFold structures. In addition, crucial insights can be inferred by comparing fold-switching proteins for which both NMR and AlphaFold structures are available. These proteins, which exhibit different conformations depending on the environmental conditions or functional states, can provide a valuable context for evaluating the accuracy of AlphaFold's structural predictions. By examining how well AlphaFold models capture these conformational changes, researchers can better understand its limitations in predicting dynamic or flexible regions, as well as its ability to account for structural transitions in functionally relevant contexts.

In cases where no NMR structures are available, an integrative approach combining sequence-based predictions of dynamics—such as backbone flexibility predictions from DynaMine, and per-residue conformational state propensity from Constava—with computational methods like NMA and MD simulations can provide valuable insights into protein flexibility and dynamic behavior. Another practical implication of this study is the introduction of a standardized computational workflow (Docker pipeline, Pandas integration) for systematically analyzing protein dynamics. This pipeline enables large-scale benchmarking of AlphaFold models against experimental and flexibility data, improving conformer selection for structural biology applications such protein design. Additionally, all datasets in this work were processed into an interactive resource, providing dynamic 3D mapping, biophysical metrics along the sequence, and downloadable structures and source data for further analysis.

7.6 NMA: strengths and shortcomings in capturing protein flexibility and dynamics

Building on the previous discussion, an important question arises regarding the application of NMA: does it tend to overestimate or underestimate protein dynamics, or is the issue rooted in the AlphaFold2 models themselves? The findings of this thesis suggest that NMA-derived RMSF values from AlphaFold2 models are less effective at capturing conformational diversity compared to those derived from NMR ensembles. This discrepancy does not indicate an inherent limitation of the NMA method itself but rather highlights the influence of the input structure's quality and definition. AlphaFold2 models tend to interpret dynamic or flexible regions as either fully disordered or adopting rigid secondary structures, overlooking partial, transient, or population-weighted conformations often observed through experimental techniques like solution NMR. For instance, Fowler et al.[197] demonstrated such behaviour in protein T1027, where AlphaFold2 predicted overly structured regions that misrepresented their true dynamics in solution. As global modes in NMA are inherently structure-dependent, applying NMA to AlphaFold2 models may lead to overestimated rigidity or underestimated flexibility in inherently dynamic regions. This occurs because AlphaFold2 often models these regions as overly defined or disordered. In contrast, NMR-derived models or ensembles, which integrate experimentally observed conformational heterogeneity, provide a more nuanced starting point, enabling NMA to more accurately reflect the actual conformational diversity of the protein, capturing the gradations of protein dynamics in NMR models. This is not to suggest that NMA is without shortcomings, as it relies on the assumption of harmonic fluctuations around a single equilibrium conformation, restricting its accuracy to states near equilibrium. NMA also ignores internal distance constraints like fixed bond lengths or angles. Without these constraints, larger movements may lead to unrealistic distortions of the molecule. Lastly, NMA does not account for local interactions. In the study, these unrealistic distortions were therefore treated by truncating the termini with a general truncation criteria, in which terminal residues with fewer than 13 Cα contacts within a distance of 10 Å were removed. This approach effectively removed unstructured terminal residues while retaining highly connected regions, ensuring that the resulting models accurately reflected flexibility. In addition, the current study was limited to the first lowest 200 non-trivial normal modes, due to computational limitations, which may have excluded high-frequency motions, potentially leading to an incomplete representation of the protein's full dynamic behaviour. In conclusion, NMA can provide valuable insights when combined with traditional MD simulations or other experimental techniques to investigate low-frequency motions while accounting for detailed atomic interactions.

7.7 Current limitations

While this thesis provides valuable insights into protein flexibility and dynamics, several inherent limitations should be acknowledged.

These challenges stem from both the complexity of the biological systems studied and the approximations inherent in computational and experimental techniques used to investigate them.

AGP dynamics: The study examined AGP and its mutants, focusing on glycosylation. However, glycosylation is highly heterogeneous, context-dependent, and influenced by cellular conditions. Experimental validation of its precise impact on protein dynamics remains difficult due to: 1) microheterogeneity of glycan structures, which are hard to control in vivo, and 2) the lack of high-resolution experimental techniques that can fully capture glycan dynamics at atomic detail. Therefore, while the study provides insights into AGP and its mutants, the findings may not generalize to all glycosylated proteins, particularly those with more complex conformational changes or membrane-associated properties.

Gradations of protein dynamics: The panoramic approach aims to integrate large-scale flexibility predictions with experimental data. However, capturing global protein dynamics across multiple proteins remains an ongoing challenge due to:

1. **Integration challenges between computational and experimental data**: These data are derived under different conditions and timescales (such as MD simulations in explicit solvent vs NMR in solution or crystallographic data in solid state), making direct comparisons non-trivial.

2. **Limitations of NMA**: As discussed previously, NMA assumes harmonic motion, meaning it may oversimplify large-scale conformational changes by treating them as small, linear deformations rather than capturing full-scale dynamics.

3. **Insufficient sampling by MD simulations**: Despite extended simulation times, MD can struggle to fully explore the entire conformational landscape of flexible proteins, especially slow motions that occur on ms or longer timescales.

4. **Generalizability and biological Context**: Many simulations and experiments are performed under artificial conditions (e.g., simplified buffer systems, absence of crowding effects, or PTMs like phosphorylation), potentially limiting biological relevance.

Despite these limitations, this thesis makes significant progress in understanding protein flexibility by combining experimental and computational approaches. Future work should focus on refining computational models, integrating more experimental validation, and improving how we connect static structure predictions with dynamic behavior in biological systems. This could include exploring integrative ensemble modeling to distinguish noise from conformational heterogeneity and leveraging open-source protein dynamics databases more accurate insights.

7.8 Future perspectives

With the advancement of tools like AlphaFold2 and AlphaFold2-based AlphaFlow for ensemble generation, the scientific community is increasingly embracing artificial intelligence (AI)-driven approaches. These tools have demonstrated impressive capabilities, particularly in detecting regions with low pLDDT scores, which might suggest regions of potential structural dynamics. However, it is important to recognize that the data used to train these models may still be incomplete or insufficient when it comes to accurately accounting for protein flexibility. Protein flexibility is a complex phenomenon influenced by factors like glycosylation, environmental conditions, and interactions, which may not yet be fully modeled by AI tools. Bridging this gap could lead to a more comprehensive understanding of protein behaviour and function. While we have demonstrated general trends using NMA and AlphaFold2, further validation through experiments and computational studies is needed to strengthen these findings. The work presented in this thesis offers several avenues for advancing protein structure refinement and design and, therefore, can complement AI-driven predictions of protein structure and dynamics. A key step toward incorporating protein dynamics into routine protein structure predictions can be achieved through computational approaches, such as:

1. Achieving a more precise interpretation and representation of flexibility, including data from computational methods (MD simulations, NMA) and experimental metrics such as model-free parameters of flexibility.

2. Refining NMA-based approaches, such as determining the optimal number of modes for accurately modeling flexibility while accounting for size differences in proteins.

3. Performing NMA on a large scale using internal coordinates, such as torsional angle space.
4. Incorporating the effects of PTMs, such as glycosylation, into flexibility predictions.

In addition, this thesis offers computational solutions, including Python scripts and packages for correlating NMA, AlphaFold2, and NMR data in routine predictions. It also addresses a gap in glycoinformatics resources by providing Python scripts for analyzing complex glycans and glycoproteins, which remain limited and inaccessible to early-stage researchers. Further, based on the conclusions of this Ph.D. thesis, future perspectives may revolve around the question: Can we effectively model and capture the complex dynamic behavior and flexibility of proteins, or the reorganization of their dynamics induced by PTMs, using:

1. MD simulations, or
2. predictive approaches based solely on sequence and structural information?

Indeed, MD simulations are a valuable tool for investigating glycoprotein dynamics. Glycans are significantly more flexible than proteins and require longer timescales to adequately sample their conformational space. However, exploring shorter timescales can often provide suffcient answers to many questions. Moreover, a potential approach to predict the effect of glycosylation on protein backbone flexibility using $\Delta RMSF$ could involve creating a distance matrix that estimates the proximity of each residue to glycosylation sites. This would allow us to classify residues as "local" or "distant" from the glycosylation sites. Typically, glycosylation reduces flexibility near the modification sites (local effects) while potentially increasing flexibility in distant regions (global effects). By combining the distance matrix with $\Delta RMSF$ values and other structural features (such as secondary structure), this information could serve as a feature set for predictive models. These models could estimate changes in flexibility due to glycosylation, helping to understand how glycan attachment influences protein dynamics. However, implementing this potential approach would require a substantial amount of simulation data or the use of normal mode analysis, which would necessitate specialized parameterized force fields to represent glycans as elastic model networks.

Sequence-based approaches like DynaMine, which utilize model-free NMR order parameters, are capable of distinguishing the biophysical impact of mutations on a protein's backbone based solely on its amino acid sequence. However, to accurately predict the effects of glycans on proteins, these approaches still require relevant data on glycoproteins in solution. Integrative NMR approaches are making significant progress in resolving complex glycans in glycoprotein complexes. Studies have shown, the relaxation process occurring in the ns to ps range in glycans can lead to broadening of NMR peaks [198]. This broadening helps reveal the flexibility and dynamics of different glycan segments. Broader peaks are usually linked to faster rotational motion and greater flexibility, while narrower peaks indicate more restricted movement or rigidity. In the future, this could enable predictions of glycan conformational states based purely on their composition.

Bibliography

[1] J. E. Murray, N. Laurieri and R. Delgoda, ‘Chapter 24 - Proteins’, in *Pharmacognosy*, S. Badal and R. Delgoda, Eds., Boston: Academic Press, 2017, pp. 477–494.

[2] M. Feig, I. Yu, P.-h. Wang, G. Nawrocki and Y. Sugita, ‘Crowding in cellular environments at an atomistic level from computer simulations’, *The Journal of Physical Chemistry B*, vol. 121, no. 34, pp. 8009–8025, 2017.

[3] S. Cooper, ‘The central dogma of cell biology’, *Cell Biology International Reports*, vol. 5, no. 6, pp. 539–549, 1981.

[4] S. Minchin and J. Lodge, ‘Understanding biochemistry: Structure and function of nucleic acids’, *Essays in biochemistry*, vol. 63, no. 4, pp. 433–456, 2019.

[5] V. Body, *Visible body: Learn biology*, en.

[6] L. Snider, *Dna and rna basics: Replication, transcription, and translation*, en-us.

[7] Y.-M. Yu and N. K. Fukagawa, ‘Chapter 2 - Protein and amino acids’, in *Present Knowledge in Nutrition (Eleventh Edition)*, B. P. Marriott, D. F. Birt, V. A. Stallings and A. A. Yates, Eds., Academic Press, 2020, pp. 15–35.

[8] M. Cox and D. Nelson, ‘Lehninger Principles of Biochemistry’. 2000, vol. 5, Journal Abbreviation: Wh Freeman Publication Title: Wh Freeman.

[9] R. Swanson, ‘A unifying concept for the amino acid code’, *Bulletin of Mathematical Biology*, vol. 46, no. 2, pp. 187–203, 1984.

[10] G. A. Petsko and D. Ring, ‘Protein Structure and Function’. New Science Press Ltd, 2004, ch. 1.

[11] H. K. Schachman, ‘Considerations on the Tertiary Structure of Proteins’, en, *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 28, pp. 409–430, 1963, Publisher: Cold Spring Harbor Laboratory Press.

[12] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, ‘Protein Function’, en, in *Molecular Biology of the Cell. 4th edition*, Garland Science, 2002.

[13] M. Levitt and R. Huber, ‘Molecular dynamics of native protein: II. Analysis and nature of motion’, *Journal of Molecular Biology*, vol. 168, no. 3, pp. 621–657, 1983.

[14] D. M. Zuckerman, ‘- Proteins Don’t Know Biology’, in *Statistical Physics of Biomolecules*, Num Pages: 20, CRC Press, 2010.

[15] G. N. Ramachandran and V. Sasisekharan, ‘Conformation of Polypeptides and Proteins**The literature survey for this review was completed in September 1967, with the journals which were then available in Madras and the preprinta which the authors had received.††By the authors’ request, the publishers have left certain matters of usage and spelling in the form in which they wrote them.’ In *Advances in Protein Chemistry*, C. B. Anfinsen, M. L. Anson, J. T. Edsall and F. M. Richards, Eds., vol. 23, Academic Press, 1968, pp. 283–437.

[16] J. Janin, S. Wodak, M. Levitt and B. Maigret, ‘Conformation of amino acid side-chains in proteins’, *Journal of Molecular Biology*, vol. 125, no. 3, pp. 357–386, 1978.

[17] G. Fischer, ‘Chemical aspects of peptide bond isomerisation’, en, *Chemical Society Reviews*, vol. 29, no. 2, pp. 119–127, 2000, Publisher: The Royal Society of Chemistry.

[18] C. J. Williams *et al.*, ‘MolProbity: More and better reference data for improved all-atom structure validation’, *Protein Science : A Publication of the Protein Society*, vol. 27, no. 1, pp. 293–315, 2018.

[19] L. Pauling and R. B. Corey, ‘Atomic coordinates and structure factors for two helical configurations of polypeptide chains’, eng, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 37, no. 5, pp. 235–240, 1951.

[20] W. G. Miller and C. V. Goebel, ‘Dimensions of protein random coils’, *Biochemistry*, vol. 7, no. 11, pp. 3925–3935, 1968.

[21] Y. Choi, S. Agarwal and C. M. Deane, ‘How long is a piece of loop?’, *PeerJ*, vol. 1, e1, 2013.

[22] A. Tramontano, ‘- The Architecture of Loops in Proteins’, in *Advances in Computational Biology*, ser. Advances in Computational Biology, H. O. Villar, Ed., vol. 2, Amsterdam: Elsevier Science B.V., 1996, pp. 239–259.

[23] L. Pauling and R. B. Corey, ‘The pleated sheet, a new layer configuration of polypeptide chains’, eng, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 37, no. 5, pp. 251–256, 1951.

[24] M. Novotny and G. J. Kleywegt, ‘A survey of left-handed helices in protein structures’, eng, *Journal of Molecular Biology*, vol. 347, no. 2, pp. 231–241, 2005.

[25] A. V. Efimov, ‘Patterns of loop regions in proteins’, *Current Opinion in Structural Biology*, vol. 3, no. 3, pp. 379–384, 1993.

[26] H. T. Kristensen, M. Christensen, M. S. Hansen, M. Hammershøj and T. K. Dalsgaard, ‘Protein–protein interactions of a whey–pea protein co-precipitate’, en, *International Journal of Food Science & Technology*, vol. 56, no. 11, pp. 5777–5790, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijfs.15165.

[27] J. Janin, R. P. Bahadur and P. Chakrabarti, ‘Protein–protein interaction and quaternary structure’, en, *Quarterly Reviews of Biophysics*, vol. 41, no. 2, pp. 133–180, 2008, Publisher: Cambridge University Press.

[28] K. A. Dill, S. B. Ozkan, M. S. Shell and T. R. Weikl, ‘The Protein Folding Problem’, *Annual review of biophysics*, vol. 37, pp. 289–316, 2008.

[29] C. M. Dobson, ‘Protein folding and misfolding’, eng, *Nature*, vol. 426, no. 6968, pp. 884–890, 2003.

[30] P. Ciryam, R. I. Morimoto, M. Vendruscolo, C. M. Dobson and E. P. O’Brien, ‘In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome’, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 2, E132–E140, 2013.

[31] C. A. Waudby, C. M. Dobson and J. Christodoulou, ‘Nature and Regulation of Protein Folding on the Ribosome’, *Trends in Biochemical Sciences*, vol. 44, no. 11, pp. 914–926, 2019.

[32] S. O. Garbuzynskiy, D. N. Ivankov, N. S. Bogatyreva and A. V. Finkelstein, ‘Golden triangle for folding rates of globular proteins’, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 1, pp. 147–150, 2013.

[33] Y. E. Kim, M. S. Hipp, A. Bracher, M. Hayer-Hartl and F. U. Hartl, ‘Molecular chaperone functions in protein folding and proteostasis’, eng, *Annual Review of Biochemistry*, vol. 82, pp. 323–355, 2013.

[34] P. Kulkarni, K. Rajagopalan, D. Yeater and R. H. Getzenberg, ‘Protein folding and the order/disorder paradox’, *Journal of cellular biochemistry*, vol. 112, no. 7, pp. 1949–1952, 2011.

[35] A. Schlessinger and B. Rost, ‘Protein flexibility and rigidity predicted from sequence’, eng, *Proteins*, vol. 61, no. 1, pp. 115–126, 2005.

[36] V. N. Uversky and A. K. Dunker, ‘Understanding protein non-folding’, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1804, no. 6, pp. 1231–1264, 2010.

[37] P. Lieutaud, F. Ferron, A. V. Uversky, L. Kurgan, V. N. Uversky and S. Longhi, ‘How disordered is my protein and what is its disorder for? a guide through the “dark side” of the protein universe’, *Intrinsically disordered proteins*, vol. 4, no. 1, e1259708, 2016.

[38] A. L. Fink, ‘Compact intermediate states in protein folding.’ *Annual review of biophysics and biomolecular structure*, vol. 24, pp. 495–522, 1995.

[39] N. Mitra, S. Sinha, T. N. C. Ramya and A. Surolia, ‘N-linked oligosaccharides as outfitters for glycoprotein folding, form and function’, English, *Trends in Biochemical Sciences*, vol. 31, no. 3, pp. 156–163, 2006, Publisher: Elsevier.

[40] S. Ramazi and J. Zahiri, ‘Post-translational modifications in proteins: Resources, tools and prediction methods’, *Database: The Journal of Biological Databases and Curation*, vol. 2021, baab012, 2021.

[41] H. S. Lee, Y. Qi and W. Im, ‘Effects of N-glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study’, en, *Scientific Reports*, vol. 5, no. 1, p. 8926, 2015, Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biophysics;Data mining Subject_term_id: computational-biophysics;data-mining.

[42] C. Boscher, J. W. Dennis and I. R. Nabi, ‘Glycosylation, galectins and cellular signaling’, *Current Opinion in Cell Biology*, Membranes and organelles, vol. 23, no. 4, pp. 383–392, 2011.

[43] C. B. Anfinsen, ‘Principles that govern the folding of protein chains’, eng, *Science (New York, N.Y.)*, vol. 181, no. 4096, pp. 223–230, 1973.

[44] D. Baker and D. A. Agard, ‘Kinetics versus thermodynamics in protein folding’, eng, *Biochemistry*, vol. 33, no. 24, pp. 7505–7509, 1994.

[45] M. Karplus and D. L. Weaver, ‘Protein folding dynamics: The diffusion-collision model and experimental data’, en, *Protein Science*, vol. 3, no. 4, pp. 650–668, 1994, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.5560030413.

[46] K. A. Dill *et al.*, ‘Principles of protein folding — A perspective from simple exact models’, en, *Protein Science*, vol. 4, no. 4, pp. 561–602, 1995, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.5560040401.

[47] P. E. Leopold, M. Montal and J. N. Onuchic, ‘Protein folding funnels: A kinetic approach to the sequence-structure relationship.’ *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 18, pp. 8721–8725, 1992.

[48] G. Gambardella *et al.*, ‘The Anfinsen Dogma: Intriguing Details Sixty-Five Years Later’, *International Journal of Molecular Sciences*, vol. 23, no. 14, p. 7759, 2022.

[49] M. C. Deller, L. Kong and B. Rupp, ‘Protein stability: A crystallographer’s perspective’, en, *Acta Crystallographica Section F: Structural Biology Communications*, vol. 72, no. 2, pp. 72–95, 2016, Publisher: International Union of Crystallography.

[50] K. A. Dill and J. L. MacCallum, ‘The Protein-Folding Problem, 50 Years On’, *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012, Publisher: American Association for the Advancement of Science.

[51] S.-Q. Liu, D.-Y. Tan, K.-Q. Zhang, X.-L. Ji, Y. Tao and Y.-X. Fu, ‘Protein folding, binding and energy landscape: A synthesis’. INTECH Open Access Publisher Retrieved from, 2012.

[52] K. A. Dill and H. S. Chan, ‘From Levinthal to pathways to funnels’, eng, *Nature Structural Biology*, vol. 4, no. 1, pp. 10–19, 1997.

[53] N. D. Socci, J. N. Onuchic and P. G. Wolynes, ‘Protein folding mechanisms and the multidimensional folding funnel’, en, *Proteins: Structure, Function, and Bioinformatics*, vol. 32, no. 2, pp. 136–158, 1998, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-0134%2819980801%2932%3A2%3C136%3A%3AAID-PROT2%3E3.0.CO%3B2-J.

[54] J. Kubelka, J. Hofrichter and W. A. Eaton, ‘The protein folding ’speed limit’’, *Current opinion in structural biology*, vol. 14, no. 1, pp. 76–88, 2004.

[55] A. Gupta, A. Singh, N. Ahmad, T. P. Singh, S. Sharma and P. Sharma, ‘Chapter 12 - Experimental techniques to study protein dynamics and conformations’, in *Advances in Protein Molecular and Structural Biology Methods*, T. Tripathi and V. K. Dubey, Eds., Academic Press, 2022, pp. 181–197.

[56] A. V. Finkelstein, N. S. Bogatyreva, D. N. Ivankov and S. O. Garbuzynskiy, ‘Protein folding problem: Enigma, paradox, solution’, eng, *Biophysical Reviews*, vol. 14, no. 6, pp. 1255–1272, 2022.

[57] K. A. Dill, S. B. Ozkan, T. R. Weikl, J. D. Chodera and V. A. Voelz, ‘The protein folding problem: When will it be solved?’, eng, *Current Opinion in Structural Biology*, vol. 17, no. 3, pp. 342–346, 2007.

[58] K. Henzler-Wildman and D. Kern, ‘Dynamic personalities of proteins’, *Nature*, vol. 450, no. 7172, pp. 964–972, 2007.

[59] Behance, *Protein Folding Funnel*, 2013.

[60] J. Jumper *et al.*, ‘Highly accurate protein structure prediction with AlphaFold’, en, *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, Number: 7873 Publisher: Nature Publishing Group.

[61] M. Baek *et al.*, ‘Accurate prediction of protein structures and interactions using a three-track neural network’, *Science*, vol. 373, no. 6557, pp. 871–876, 2021, Publisher: American Association for the Advancement of Science.

[62] Z. Lin *et al.*, ‘Evolutionary-scale prediction of atomic-level protein structure with a language model’, eng, *Science (New York, N.Y.)*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[63] K. M. Ruff and R. V. Pappu, ‘AlphaFold and Implications for Intrinsically Disordered Proteins’, *Journal of Molecular Biology*, From Protein Sequence to Structure at Warp Speed: How Alphafold Impacts Biology, vol. 433, no. 20, p. 167 208, 2021.

[64] J. Jumper *et al.*, ‘Applying and improving AlphaFold at CASP14’, en, *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 12, pp. 1711–1721, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26257.

[65] V. Mariani, M. Biasini, A. Barbato and T. Schwede, ‘lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests’, *Bioinformatics*, vol. 29, no. 21, pp. 2722–2728, 2013.

[66] Y. J. Huang *et al.*, ‘Assessment of prediction methods for protein structures determined by nmr in casp14: Impact of alphafold2’, *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 12, pp. 1959–1976, 2021.

[67] T. Saldaño *et al.*, ‘Impact of protein conformational diversity on alphafold predictions’, *Bioinformatics*, vol. 38, no. 10, pp. 2742–2748, 2022.

[68] C. Marino-Buslje, A. M. Monzon, D. J. Zea, M. S. Fornasari and G. Parisi, ‘On the dynamical incompleteness of the protein data bank’, *Briefings in Bioinformatics*, vol. 20, no. 1, pp. 356–359, 2019.

[69] B. Halle, ‘Biomolecular cryocrystallography: Structural changes during flash-cooling’, *Proceedings of the National Academy of Sciences*, vol. 101, no. 14, pp. 4793–4798, 2004.

[70] H. Na, K. Hinsen and G. Song, ‘The amounts of thermal vibrations and static disorder in protein x-ray crystallographic b-factors’, *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 11, pp. 1442–1457, 2021.

[71] J. S. Fraser, K. Lindorff-Larsen and M. Bonomi, ‘What will computational modeling approaches have to say in the era of atomistic cryo-em data?’, *Journal of chemical information and modeling*, vol. 60, no. 5, pp. 2410–2412, 2020.

[72] J. Gavalda-Garcia, B. Dixit, A. Díaz, A. Ghysels and W. Vranken, ‘Gradations in protein dynamics captured by experimental nmr are not well represented by alphafold2 models and other computational metrics.’ *Journal of Molecular Biology*, p. 168 900, 2024.

[73] J. Abramson *et al.*, ‘Accurate structure prediction of biomolecular interactions with AlphaFold 3’, en, *Nature*, vol. 630, no. 8016, pp. 493–500, 2024, Publisher: Nature Publishing Group.

[74] J. Roca-Martinez *et al.*, ‘Challenges in describing the conformation and dynamics of proteins with ambiguous behavior’, English, *Frontiers in Molecular Biosciences*, vol. 9, 2022, Publisher: Frontiers.

[75] P. W. Fenimore, H. Frauenfelder, B. H. McMahon and F. G. Parak, ‘Slaving: Solvent fluctuations dominate protein dynamics and functions’, *Proceedings of the National Academy of Sciences*, vol. 99, no. 25, pp. 16 047–16 051, 2002, Publisher: Proceedings of the National Academy of Sciences.

[76] J. M. Krieger, C. O. S. Sorzano, J. M. Carazo and I. Bahar, ‘Protein dynamics developments for the large scale and cryoEM: Case study of ProDy 2.0’, en, *Acta Crystallographica Section D: Structural Biology*, vol. 78, no. 4, pp. 399–409, 2022, Publisher: International Union of Crystallography.

[77] K. Nam and M. Wolf-Watz, ‘Protein dynamics: The future is bright and complicated!’, *Structural Dynamics*, vol. 10, no. 1, p. 014 301, 2023.

[78] C. C. Hsu, M. J. Buehler and A. Tarakanova, ‘The Order-Disorder Continuum: Linking Predictions of Protein Structure and Disorder through Molecular Simulation’, en, *Scientific Reports*, vol. 10, no. 1, p. 2068, 2020, Publisher: Nature Publishing Group.

[79] Y. Xu and M. Havenith, ‘Perspective: Watching low-frequency vibrations of water in biomolecular recognition by thz spectroscopy’, *The Journal of chemical physics*, vol. 143, no. 17, 2015.

[80] J. A. McCammon and M. Karplus, 'Simulation of Protein Dynamics', *Annual Review of Physical Chemistry*, vol. 31, no. 1, pp. 29–45, 1980, _eprint: https://doi.org/10.1146/annurev.pc.31.100180.000333.

[81] I. Bahar, T. R. Lezon, A. Bakan and I. H. Shrivastava, 'Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins', *Chemical reviews*, vol. 110, no. 3, pp. 1463–1497, 2010.

[82] T. Xie, T. Saleh, P. Rossi and C. G. Kalodimos, 'Conformational states dynamically populated by a kinase determine its function', *Science*, vol. 370, no. 6513, eabc2754, 2020.

[83] E. Rennella *et al.*, 'Dynamic conformational equilibria in the active states of kras and nras', *RSC Chemical Biology*, vol. 6, no. 1, pp. 106–118, 2025.

[84] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. Parrish, H. Wyckoff and D. C. Phillips, 'A three-dimensional model of the myoglobin molecule obtained by x-ray analysis', *Nature*, vol. 181, no. 4610, pp. 662–666, 1958.

[85] M. L. Quillin, R. M. Arduini, J. S. Olson and G. N. Phillips Jr, 'High-resolution crystal structures of distal histidine mutants of sperm whale myoglobin', *Journal of molecular biology*, vol. 234, no. 1, pp. 140–155, 1993.

[86] G. U. Nienhaus, J. R. Mourant and H. Frauenfelder, 'Spectroscopic evidence for conformational relaxation in myoglobin.' *Proceedings of the National Academy of Sciences*, vol. 89, no. 7, pp. 2902–2906, 1992.

[87] R. J. Woods and M. B. Tessier, 'Computational glycoscience: Characterizing the spatial and temporal properties of glycans and glycan–protein complexes', en, *Current Opinion in Structural Biology*, vol. 20, no. 5, pp. 575–583, 2010.

[88] C. David and D. Jacobs, 'Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins', *Methods in molecular biology (Clifton, N.J.)*, vol. 1084, pp. 193–226, 2014.

[89] R. C. Bernardi, M. C. R. Melo and K. Schulten, 'Enhanced sampling techniques in molecular dynamics simulations of biological systems', *Biochimica et Biophysica Acta (BBA) - General Subjects*, Recent developments of molecular dynamics, vol. 1850, no. 5, pp. 872–877, 2015.

[90] J. A. McCammon, ‘Protein dynamics’, en, *Reports on Progress in Physics*, vol. 47, no. 1, p. 1, 1984.

[91] G. S. Buchner, R. D. Murphy, N.-V. Buchete and J. Kubelka, ‘Dynamics of protein folding: Probing the kinetic network of folding–unfolding transitions with experiment and theory’, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, Protein Dynamics: Experimental and Computational Approaches, vol. 1814, no. 8, pp. 1001–1020, 2011.

[92] I. Luque and E. Freire, ‘Structural stability of binding sites: Consequences for binding affinity and allosteric effects’, en, *Proteins: Structure, Function, and Bioinformatics*, vol. 41, no. S4, pp. 63–71, 2000, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0134%282000%2941%3A4%2B%3C63%3A%3AAID-PROT60%3E3.0.CO%3B2-6.

[93] N. Sinha and R. Nussinov, ‘Point mutations and sequence variability in proteins: Redistributions of preexisting populations’, eng, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 6, pp. 3139–3144, 2001.

[94] K. Teilum, J. Olsen and B. B. Kragelund, ‘Functional aspects of protein flexibility’, *Cellular and molecular life sciences : CMLS*, vol. 66, pp. 2231–47, 2009.

[95] K. Teilum, J. G. Olsen and B. B. Kragelund, ‘Protein stability, flexibility and function’, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, Protein Dynamics: Experimental and Computational Approaches, vol. 1814, no. 8, pp. 969–976, 2011.

[96] P. H. Hünenberger, A. E. Mark and W. F. van Gunsteren, ‘Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations’, *Journal of Molecular Biology*, vol. 252, no. 4, pp. 492–503, 1995.

[97] A. Bornot, C. Etchebest and A. G. De Brevern, ‘Predicting protein flexibility through the prediction of local structures’, *Proteins*, vol. 79, no. 3, pp. 839–852, 2011.

[98] M. Berjanskii and D. S. Wishart, ‘NMR: Prediction of protein flexibility’, en, *Nature Protocols*, vol. 1, no. 2, pp. 683–688, 2006, Number: 2 Publisher: Nature Publishing Group.

[99] E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts and W. F. Vranken, ‘The DynaMine webserver: Predicting protein dynamics from sequence’, *Nucleic Acids Research*, vol. 42, no. Web Server issue, W264–W270, 2014.

[100] L. P. Kagami *et al.*, ‘b2bTools: Online predictions for protein biophysical features and their conservation’, *Nucleic Acids Research*, vol. 49, no. W1, W52–W59, 2021.

[101] G. Orlando, D. Raimondi, L. P. Kagami and W. F. Vranken, ‘ShiftCrypt: A web server to understand and biophysically align proteins through their NMR chemical shift values’, *Nucleic Acids Research*, vol. 48, no. W1, W36–W40, 2020.

[102] R. Apweiler, H. Hermjakob and N. Sharon, ‘On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database11Dedicated to Prof. Akira Kobata and Prof. Harry Schachter on the occasion of their 65th birthdays.’ en, *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1473, no. 1, pp. 4–8, 1999.

[103] H. Lis and N. Sharon, ‘Protein glycosylation’, en, *European Journal of Biochemistry*, vol. 218, no. 1, pp. 1–27, 1993, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1432-1033.1993.tb18347.x.

[104] S. Neelamegham *et al.*, ‘Updates to the Symbol Nomenclature for Glycans guidelines’, eng, *Glycobiology*, vol. 29, no. 9, pp. 620–624, 2019.

[105] C. R. Bertozzi and D. Rabuka, ‘Structural Basis of Glycan Diversity’, eng, in *Essentials of Glycobiology*, A. Varki *et al.*, Eds., 2nd, Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2009.

[106] R. U. Lemieux and S. Koto, ‘The conformational properties of glycosidic linkages’, *Tetrahedron*, vol. 30, no. 13, pp. 1933–1944, 1974.

[107] R. Kornfeld and S. Kornfeld, ‘Structure of Glycoproteins and Their Oligosaccharide Units’, en, in *The Biochemistry of Glycoproteins and Proteoglycans*, W. J. Lennarz, Ed., Boston, MA: Springer US, 1980, pp. 1–34.

[108] *24.8: Disaccharides and Glycosidic Bonds*, en, 2016.

[109] 'Ionsource’, *N- and O-linked Protein Glycosylation.*

[110] F. Higel, A. Seidl, F. Sörgel and W. Friess, ‘N-glycosylation heterogeneity and the influence on structure, function and pharmacokinetics of monoclonal antibodies and Fc fusion proteins’, *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 100, pp. 94–100, 2016.

[111] K. W. Moremen, M. Tiemeyer and A. V. Nairn, ‘Vertebrate protein glycosylation: Diversity, synthesis and function’, en, *Nature Reviews Molecular Cell Biology*, vol. 13, no. 7, pp. 448–462, 2012, Publisher: Nature Publishing Group.

[112] S. M. Muthana, C. Campbell and J. C. Gildersleeve, ‘Modifications of Glycans: Biological Significance and Therapeutic Opportunities’, *ACS Chemical Biology*, vol. 7, no. 1, pp. 31–43, 2012.

[113] T. Horvat, V. Zoldoš and G. Lauc, ‘Evolutional and clinical implications of the epigenetic regulation of protein glycosylation’, eng, *Clinical Epigenetics*, vol. 2, no. 2, pp. 425–432, 2011.

[114] J. W. Dennis, M. Granovsky and C. E. Warren, ‘Protein glycosylation in development and disease’, en, *BioEssays*, vol. 21, no. 5, pp. 412–421, 1999, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291521-1878%28199905%2921%3A5%3C412%3A%3AAID-BIES8%3E3.0.CO%3B2-5.

[115] Y. Mazola, G. Chinea and A. Musacchio, ‘Glycosylation and Bioinformatics: Current status for glycosylation prediction tools’, EN, *Biotecnología Aplicada*, vol. 28, no. 1, pp. 6–12, 2011.

[116] A. Varki *et al.*, Eds., ‘Essentials of Glycobiology’, eng, 2nd. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2009.

[117] W. B. Struwe and C. V. Robinson, ‘Relating glycoprotein structural heterogeneity to function – insights from native mass spectrometry’, *Current opinion in structural biology*, vol. 58, pp. 241–248, 2019.

[118] A. Dell, A. Galadari, F. Sastre and P. Hitchen, ‘Similarities and Differences in the Glycosylation Mechanisms in Prokaryotes and Eukaryotes’, *International Journal of Microbiology*, vol. 2010, p. 148 178, 2010.

[119] J. Lombard, ‘The multiple evolutionary origins of the eukaryotic N-glycosylation pathway’, en, *Biology Direct*, vol. 11, no. 1, p. 36, 2016.

[120] R. Kornfeld and S. Kornfeld, ‘Assembly of asparagine-linked oligosaccharides’, eng, *Annual Review of Biochemistry*, vol. 54, pp. 631–664, 1985.

[121] H. J. An *et al.*, ‘Extensive determination of glycan heterogeneity reveals an unusual abundance of high mannose glycans in enriched plasma membranes of human embryonic stem cells’, eng, *Molecular & cellular proteomics: MCP*, vol. 11, no. 4, p. M111.010660, 2012.

[122] P. Stanley, ‘Golgi Glycosylation’, en, *Cold Spring Harbor Perspectives in Biology*, vol. 3, no. 4, a005199, 2011, Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[123] D. J. Kelleher, D. Karaoglu, E. C. Mandon and R. Gilmore, ‘Oligosaccharyltransferase isoforms that contain different catalytic STT3 subunits have distinct enzymatic properties’, eng, *Molecular Cell*, vol. 12, no. 1, pp. 101–111, 2003.

[124] A. Helenius and M. Aebi, ‘Roles of N-linked glycans in the endoplasmic reticulum’, eng, *Annual Review of Biochemistry*, vol. 73, pp. 1019–1049, 2004.

[125] K. W. Moremen and M. Molinari, ‘N-linked glycan recognition and processing: The molecular basis of endoplasmic reticulum quality control’, *Current opinion in structural biology*, vol. 16, no. 5, pp. 592–599, 2006.

[126] I. Gudelj and G. Lauc, ‘Protein N-Glycosylation in Cardiovascular Diseases and Related Risk Factors’, en, *Current Cardiovascular Risk Reports*, vol. 12, no. 6, p. 16, 2018.

[127] L. Jia, J. Zhang, T. Ma, Y. Guo, Y. Yu and J. Cui, ‘The Function of Fucosylation in Progression of Lung Cancer’, English, *Frontiers in Oncology*, vol. 8, 2018, Publisher: Frontiers.

[128] T. S. Keeley, S. Yang and E. Lau, ‘The Diverse Contributions of Fucose Linkages in Cancer’, *Cancers*, vol. 11, no. 9, p. 1241, 2019.

[129] S. Yang, J. Xia, Z. Yang, M. Xu and S. Li, ‘Lung cancer molecular mutations and abnormal glycosylation as biomarkers for early diagnosis’, *Cancer Treatment and Research Communications*, vol. 27, p. 100 311, 2021.

[130] M. Wang, J. Zhu, D. M. Lubman and C. Gao, 'Aberrant glycosylation and cancer biomarker discovery: A promising and thorny journey', en, *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 57, no. 4, pp. 407–416, 2019, Publisher: De Gruyter.

[131] R. Francisco *et al.*, 'Congenital disorders of glycosylation (CDG): State of the art in 2022', *Orphanet Journal of Rare Diseases*, vol. 18, no. 1, p. 329, 2023.

[132] J. Munkley and D. J. Elliott, 'Hallmarks of glycosylation in cancer', *Oncotarget*, vol. 7, no. 23, pp. 35 478–35 489, 2016.

[133] A. Kirwan, M. Utratna, M. E. O'Dwyer, L. Joshi and M. Kilcoyne, 'Glycosylation-Based Serum Biomarkers for Cancer Diagnostics and Prognostics', *BioMed Research International*, vol. 2015, p. 490 531, 2015.

[134] S. Perez and O. Makshakova, 'Multifaceted Computational Modeling in Glycoscience', *Chemical Reviews*, vol. 122, no. 20, pp. 15 914–15 970, 2022, Publisher: American Chemical Society.

[135] I. Tvaroŝka and T. Bleha, 'Anomeric and Exo-Anomeric Effects in Carbohydrate Chemistry', in *Advances in Carbohydrate Chemistry and Biochemistry*, R. S. Tipson and D. Horton, Eds., vol. 47, Academic Press, 1989, pp. 45–123.

[136] M. Wormald, A. J. Petrescu, Y.-L. Pao, A. Fowler, T. Elliott and R. Dwek, 'Conformational Studies of Oligosaccharides and Glycopeptides: Complementarity of NMR, X-ray Crystallography, and Molecular Modelling', *Chemical reviews*, vol. 102, pp. 371–86, 2002.

[137] V. Palivec, P. Michal, J. Kapitán, H. Martinez-Seara and P. Bouř, 'Raman Optical Activity of Glucose and Sorbose in Extended Wavenumber Range', *ChemPhysChem*, vol. 21, 2020.

[138] A. M. Salisburg, A. L. Deline, K. W. Lexa, G. C. Shields and K. N. Kirschner, 'Ramachandran-type plots for glycosidic linkages: Examples from molecular dynamic simulations using the Glycam06 force field', eng, *Journal of Computational Chemistry*, vol. 30, no. 6, pp. 910–921, 2009.

[139] L. Perić-Hassler, H. S. Hansen, R. Baron and P. H. Hünenberger, 'Conformational properties of glucose-based disaccharides investigated using molecular dynamics simulations with local elevation umbrella sampling', eng, *Carbohydrate Research*, vol. 345, no. 12, pp. 1781–1801, 2010.

[140] B. Mulloy, G. W. Hart and P. Stanley, ‘Structural Analysis of Glycans’, en, in *Essentials of Glycobiology. 2nd edition*, Cold Spring Harbor Laboratory Press, 2009.

[141] M. Nagae and Y. Yamaguchi, ‘Function and 3D structure of the N-glycans on glycoproteins’, eng, *International Journal of Molecular Sciences*, vol. 13, no. 7, pp. 8398–8429, 2012.

[142] J. H. Prestegard, ‘A perspective on the PDB’s impact on the field of glycobiology’, *The Journal of Biological Chemistry*, vol. 296, p. 100 556, 2021.

[143] S. Pérez and D. de Sanctis, ‘Glycoscience@Synchrotron: Synchrotron radiation applied to structural glycoscience’, eng, *Beilstein Journal of Organic Chemistry*, vol. 13, pp. 1145–1167, 2017.

[144] S. Jo, T. Kim, V. G. Iyer and W. Im, ‘CHARMM-GUI: A web-based graphical user interface for CHARMM’, eng, *Journal of Computational Chemistry*, vol. 29, no. 11, pp. 1859–1865, 2008.

[145] J. Huang *et al.*, ‘CHARMM36m: An improved force field for folded and intrinsically disordered proteins’, en, *Nature Methods*, vol. 14, no. 1, pp. 71–73, 2017, Number: 1 Publisher: Nature Publishing Group.

[146] K. N. Kirschner *et al.*, ‘GLYCAM06: A generalizable biomolecular force field. Carbohydrates’, en, *Journal of Computational Chemistry*, vol. 29, no. 4, pp. 622–655, 2008, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20820.

[147] D. Kony, W. Damm, S. Stoll and W. F. Van Gunsteren, ‘An improved OPLS-AA force field for carbohydrates’, eng, *Journal of Computational Chemistry*, vol. 23, no. 15, pp. 1416–1429, 2002.

[148] L. Pol-Fachin, C. L. Fernandes and H. Verli, ‘GROMOS96 43a1 performance on the characterization of glycoprotein conformational ensembles through molecular dynamics simulations’, en, *Carbohydrate Research*, vol. 344, no. 4, pp. 491–500, 2009.

[149] R. J. Woods, A. Pathiaseril, M. R. Wormald, C. J. Edge and R. A. Dwek, ‘The high degree of internal flexibility observed for an oligomannose oligosaccharide does not alter the overall topology of the molecule’, *European journal of biochemistry*, vol. 258, no. 2, pp. 372–386, 1998.

[150] J. Gonzalez-Outeiriño, K. N. Kirschner, S. Thobhani and R. J. Woods, 'Reconciling solvent effects on rotamer populations in carbohydrates — A joint MD and NMR analysis', *Canadian journal of chemistry*, vol. 84, no. 4, pp. 569–579, 2006.

[151] V. Gervais, A. Zerial and H. Oschkinat, 'NMR investigations of the role of the sugar moiety in glycosylated recombinant human granulocyte-colony-stimulating factor', eng, *European Journal of Biochemistry*, vol. 247, no. 1, pp. 386–395, 1997.

[152] S.-T. Jiang, G.-H. Chen, S.-J. Tang and C.-S. Chen, 'Effect of glycosylation modification (N-Q-(108)I –> N-Q-(108)T) on the freezing stability of recombinant chicken Cystatin overexpressed in Pichia pastoris X-33', eng, *Journal of Agricultural and Food Chemistry*, vol. 50, no. 19, pp. 5313–5317, 2002.

[153] H. C. Joao, I. G. Scragg and R. A. Dwek, 'Effects of glycosylation on protein conformation and amide proton exchange rates in RNase B', eng, *FEBS letters*, vol. 307, no. 3, pp. 343–346, 1992.

[154] P. M. Rudd *et al.*, 'Glycoforms modify the dynamic stability and functional activity of an enzyme', eng, *Biochemistry*, vol. 33, no. 1, pp. 17–22, 1994.

[155] Y. Hashimoto *et al.*, 'Effects of glycosylation on the structure and dynamics of eel calcitonin in micelles and lipid bilayers determined by nuclear magnetic resonance spectroscopy', eng, *Biochemistry*, vol. 38, no. 26, pp. 8377–8384, 1999.

[156] G. Gupta, S. Sinha, N. Mitra and A. Surolia, 'Probing into the role of conserved N-glycosylation sites in the Tyrosinase glycoprotein family', eng, *Glycoconjugate Journal*, vol. 26, no. 6, pp. 691–695, 2009.

[157] B. J. Alder and T. E. Wainwright, 'Studies in molecular dynamics. i. general method', *The Journal of Chemical Physics*, vol. 31, no. 2, pp. 459–466, 1959.

[158] R. G. Schmidt and J. Brickmann, 'Molecular dynamics simulation of the proton transport in water', *Berichte der Bunsengesellschaft für physikalische Chemie*, vol. 101, no. 12, pp. 1816–1827, 1997.

[159] A. Laaksonen and Y. Tu, 'Methods of incorporating quantum mechanical calculations into molecular dynamics simulations', *Balbuena, Seminario [25]*, pp. 1–29, 1999.

[160] D. Frenkel and B. Smit, *Molecular simulation: From algorithms to applications*, 2000.

[161] J. Meller *et al.*, ‘Molecular dynamics’, *Encyclopedia of life sciences*, vol. 18, 2001.

[162] Q. Spreiter and M. Walter, ‘Classical molecular dynamics simulation with the velocity verlet algorithm at strong external magnetic fields’, *Journal of Computational Physics*, vol. 152, no. 1, pp. 102–119, 1999.

[163] E. Por, M. van Kooten and V. Sarkovic, ‘Nyquist–shannon sampling theorem’, *Leiden University*, vol. 1, no. 1, pp. 1–2, 2019.

[164] S. A. Hollingsworth and R. O. Dror, ‘Molecular dynamics simulation for all’, *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018.

[165] T. A. Wassenaar and A. E. Mark, ‘The effect of box shape on the dynamic properties of proteins simulated under periodic boundary conditions’, *Journal of computational chemistry*, vol. 27, no. 3, pp. 316–325, 2006.

[166] W. W. Wood, J. J. Erpenbeck, G. A. Baker Jr and J. Johnson, ‘Molecular dynamics ensemble, equation of state, and ergodicity’, *Physical Review E*, vol. 63, no. 1, p. 011 106, 2000.

[167] P. H. Hünenberger and J. A. McCammon, ‘Effect of artificial periodicity in simulations of biomolecules under ewald boundary conditions: A continuum electrostatics study’, *Biophysical chemistry*, vol. 78, no. 1-2, pp. 69–88, 1999.

[168] D. J. Evans and D. J. Searles, ‘Causality, response theory, and the second law of thermodynamics’, *Physical Review E*, vol. 53, no. 6, p. 5808, 1996.

[169] J. Wereszczynski and J. A. McCammon, ‘Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition’, *Quarterly reviews of biophysics*, vol. 45, no. 1, pp. 1–25, 2012.

[170] D. M. Zuckerman, ‘Statistical physics of biomolecules: an introduction’. CRC press, 2010.

[171] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. a. Swaminathan and M. Karplus, ‘Charmm: A program for macromolecular energy, minimization, and dynamics calculations’, *Journal of computational chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[172] A. L. Justin, ‘From proteins to perturbed hamiltonians: A suite of tutorials for the gromacs-2018 molecular simulation package [article v1. 0]’, *Living Journal of Computational Molecular Science*, vol. 1, no. 1, p. 5068, 2018.

[173] D. L. Theobald, ‘Rapid calculation of rmsds using a quaternion-based characteristic polynomial’, *Foundations of Crystallography*, vol. 61, no. 4, pp. 478–480, 2005.

[174] K. Sargsyan, C. Grauffel and C. Lim, ‘How molecular size impacts rmsd applications in molecular dynamics simulations’, *Journal of chemical theory and computation*, vol. 13, no. 4, pp. 1518–1524, 2017.

[175] R. S. Macomber, ‘A complete introduction to modern NMR spectroscopy’. John Wiley & Sons, 1997.

[176] S. Curry, ‘Structural biology: A century-long journey into an unseen world’, *Interdisciplinary Science Reviews*, vol. 40, no. 3, pp. 308–328, 2015.

[177] I. R. Kleckner and M. P. Foster, ‘An introduction to nmr-based approaches for measuring protein dynamics’, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1814, no. 8, pp. 942–968, 2011.

[178] G. Orlando, D. Raimondi and W. F. Vranken, ‘Auto-encoding nmr chemical shifts from their native vector space to a residue-level biophysical index’, *Nature communications*, vol. 10, no. 1, p. 2511, 2019.

[179] E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts and W. F. Vranken, ‘From protein sequence to dynamics and disorder with dynamine’, *Nature communications*, vol. 4, no. 1, p. 2741, 2013.

[180] M. V. Berjanskii and D. S. Wishart, ‘Application of the random coil index to studying protein flexibility’, *Journal of biomolecular NMR*, vol. 40, pp. 31–48, 2008.

[181] M. Berjanskii, *Protein nmr dynamics: Random coil index*, Lecture presentation, accessed from SlideShare, 2013.

[182] A. Bundi and K. Wüthrich, ‘1h-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides h-gly-gly-x-l-ala-oh’, *Biopolymers: Original Research on Biomolecules*, vol. 18, no. 2, pp. 285–297, 1979.

[183] J. A. Vila, D. R. Ripoll, H. A. Baldoni and H. A. Scheraga, ‘Unblocked statistical-coil tetrapeptides and pentapeptides in aqueous solution: A theoretical study’, *Journal of Biomolecular NMR*, vol. 24, pp. 245–262, 2002.

[184] M. V. Berjanskii and D. S. Wishart, ‘A simple method to predict protein flexibility using secondary chemical shifts’, *Journal of the American Chemical Society*, vol. 127, no. 43, pp. 14 970–14 971, 2005.

[185] J. L. Goodman, M. D. Pagel and M. J. Stone, ‘Relationships between protein structure and dynamics from a database of nmr-derived backbone order parameters’, *Journal of molecular biology*, vol. 295, no. 4, pp. 963–978, 2000.

[186] S. Hayward and B. L. De Groot, ‘Normal modes and essential dynamics’, *Molecular Modeling of Proteins*, pp. 89–106, 2008.

[187] I. Kolossváry, ‘A fresh look at the normal mode analysis of proteins: Introducing allosteric co-vibrational modes’, *JACS Au*, vol. 4, no. 4, pp. 1303–1309, 2024.

[188] H. Wako and S. Endo, ‘Normal mode analysis as a method to derive protein dynamics information from the protein data bank’, *Biophysical reviews*, vol. 9, no. 6, pp. 877–893, 2017.

[189] A. Ghysels, B. T. Miller, F. C. Pickard IV and B. R. Brooks, ‘Comparing normal modes across different models and scales: Hessian reduction versus coarse-graining’, en, *Journal of Computational Chemistry*, vol. 33, no. 28, pp. 2250–2275, 2012, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.23076.

[190] M. M. Tirion, ‘Large amplitude elastic motions in proteins from a single-parameter, atomic analysis’, *Physical review letters*, vol. 77, no. 9, p. 1905, 1996.

[191] X. Qiang *et al.*, ‘Large-scale silicon quantum photonics implementing arbitrary two-qubit processing’, *Nature photonics*, vol. 12, no. 9, pp. 534–539, 2018.

[192] W. Wriggers, *Normal mode and principal component analysis*, Lecture presentation, accessed from Biomachina.

[193] J. A. Bauer, J. Pavlović and V. Bauerová-Hlinková, ‘Normal mode analysis as a routine part of a structural investigation’, *Molecules*, vol. 24, no. 18, p. 3293, 2019.

[194] S. P. Tiwari *et al.*, ‘Webnm@ v2. 0: Web server and services for comparing protein flexibility’, *BMC bioinformatics*, vol. 15, no. 1, pp. 1–12, 2014.

[195] K. Hinsen, A.-J. Petrescu, S. Dellerue, M.-C. Bellissent-Funel and G. R. Kneller, ‘Harmonicity in slow protein dynamics’, *Chemical Physics*, vol. 261, no. 1-2, pp. 25–37, 2000.

[196] S. M. Hollup, G. Salensminde and N. Reuter, ‘Webnm@: A web application for normal mode analyses of proteins’, *BMC bioinformatics*, vol. 6, pp. 1–8, 2005.

[197] N. J. Fowler and M. P. Williamson, ‘The accuracy of protein structures in solution determined by alphafold and nmr’, *Structure*, vol. 30, no. 7, pp. 925–933, 2022.

[198] L. Unione *et al.*, ‘Glycoprofile analysis of an intact glycoprotein as inferred by nmr spectroscopy’, *ACS central science*, vol. 5, no. 9, pp. 1554–1561, 2019.

Appendices

Supporting Information

Conformational dynamics of α-1 acid glycoprotein (AGP) in cancer: A comparative study of glycosylated and unglycosylated AGP

Bhawna Dixit,[abc] *Wim Vranken,*[bc] *An Ghysels* [a,*]

a IBiTech–BioMMeda Group, Ghent University, Belgium

b Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Belgium

c Structural Biology Brussels, Vrije Universiteit Brussel, Belgium

Contents

1. B2BTools: prediction of backbone dynamics propensity with DynaMine 179
2. Selection out of 60 mutants 179
3. Overview of MD systems 181
4. RMSD 182
5. Radius of gyration 185
6. RMSF 186
7. Average values for RMSD, RMSF, *R*g 190
8. Visualization of MD snapshots 191
9. Contact analysis 194
10. SASA 195
10.1 SASA of protein atoms 195
10.2 SASA of glycans 196
11 H-bonds (protein) 199
12 Principal component analysis (PCA) 200
13 Asparagine side-chain conformations 202
14 Glycan torsion angles 205
(a) Distribution of β-GlcNAc(1→)N linkages 206
(b) Distribution of glycosidic linkages 207
15 H-bonds between protein and glycans 211

1. B2BTools: prediction of backbone dynamics propensity with DynaMine

DynaMine (from the b2bTools software) is a sequence-based approach that predicts the flexibility of the residues: flexible, rigid, or context-dependent, which indicates a capability of being either rigid or flexible. This flexibility score is called the backbone dynamics propensity.

The wild-type sequence is the variant ORM1*F1, as collected from the UniProt database with entry P02763 (accessed September 2021). The full wild-type sequence (denoted 'AGP') of 201 residues, including the 18 residues signal peptide of the protein precursor, was analysed with DynaMine. The backbone dynamics propensity is given in Figure S1. Next, the 60 missense mutations of AGP were collected from COSMIC. The AGP sequence was mutated and the ".fasta" files of the point missense mutations were run with DynaMine to predict their backbone dynamics propensity (Figure S1).

Variants

AGP	ORM1*F1 variant (wild-type AGP), often referred to as R38Q
Q38R	ORM1*S variant
V174M	ORM1*F2 variant

The ORM1*S variant (with an arginine at position 38) can undergo a somatic mutation to the ORM1*F1 variant (AGP). This is often referred to as R38Q.

R167C has been listed as a natural variant in older versions of the UniProt entry P02763. In 2022, it is listed as a missense mutation.

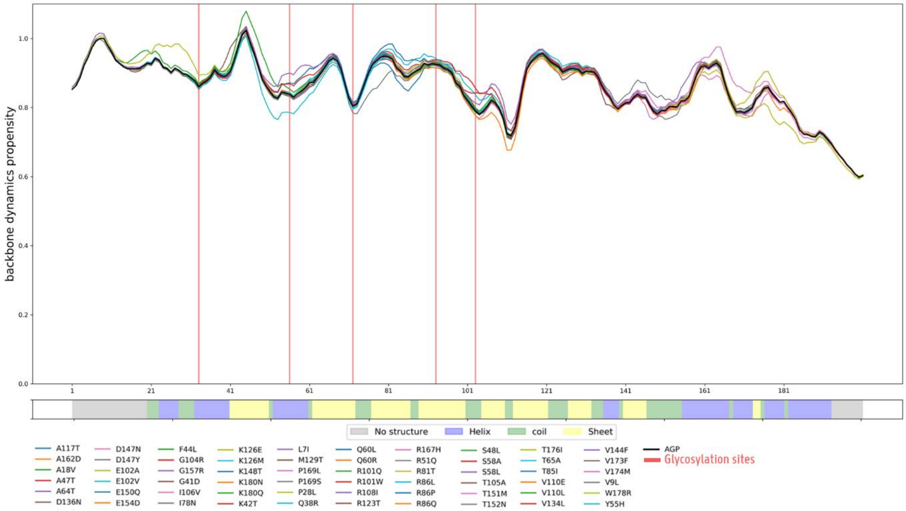


Figure S 1. Backbone dynamics propensity predicted by DynaMine (B2BTools) for all missense mutations of AGP. Wild-type in black. Glycosylation sites are marked with red vertical lines. Secondary structure elements, from the PDB code 3kq0, are given at the bottom: helix (blue), sheet (yellow), loop (green), no structure information (grey).

2. Selection out of 60 mutants

We first selected AGP variants Q38R (ORM1*S) and V174M (ORM1*F2). The mutation R167C was also selected because it was categorized as a variant in earlier versions of UniProt. Out of all the B2BTools predictions of COSMIC dataset, we selected five mutations. Point mutations at residues that did not have a well-defined crystal structure in X-ray crystal structure of AGP (PDB: 3kq0) were discarded. For instance, L7I is a point mutation at a residue in the signal peptide, for which the structure information is missing in the 3kq0.pdb file. Mutations like L7I have therefore not been considered. Next, we calculated the root-mean-square-error (RMSE) between the predicted backbone dynamics propensity of AGP and mutants (Figure S 2). A high RMSE value for a certain point mutation indicates that this point mutation has a high impact on the predicted backbone dynamics propensity. Therefore, the mutations with high RMSE were considered. Moreover, we mapped the mutations on the structure of AGP (PDB: 3kq0) and chose the mutations that are structurally close to one or more of the five glycosylation sites (Figure S 3).

The final selection included (1) AGP, (2) two variants from UniProt: Q38R (ORM1*S), V174M (ORM1*F2), and (3) 6 mutations: P28L, Q60L, I78N, R101W, R167C, and P169L.

As R167C was listed as a natural variant in earlier versions of UniProt, it is coloured as a natural variant (green) in Figure S 2 and Figure S 3. In the current version of UniProt, this mutation is identified as a missense mutation.

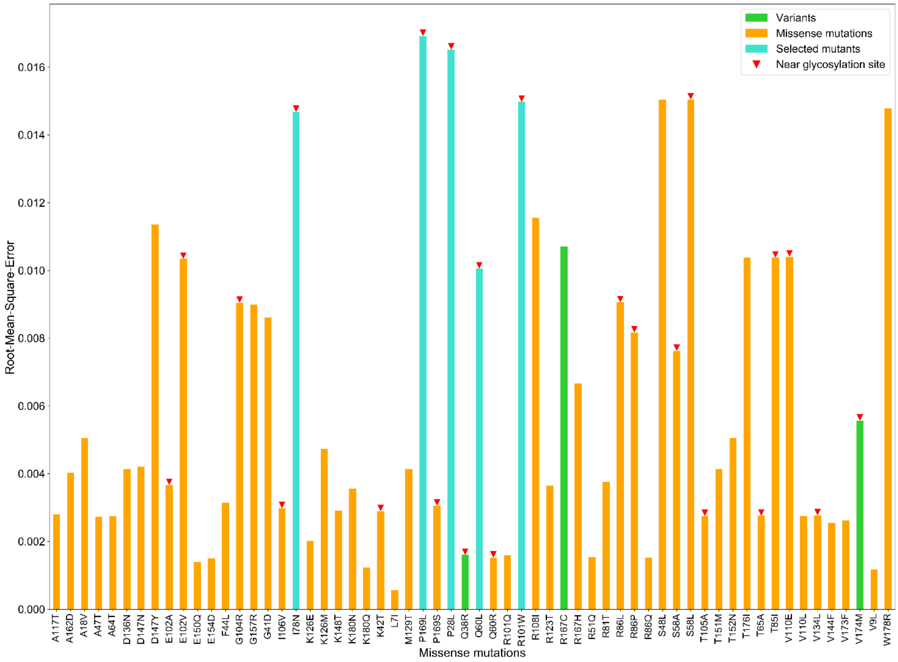


Figure S 2. Root-mean-square-error (RMSE) between backbone dynamics propensity of wild-type AGP and mutations. Structural proximity to glycosylation sites is marked with red markers. Selected mutants for the current study have high RMSE and are structurally close (cyan), or they are natural variants (green).

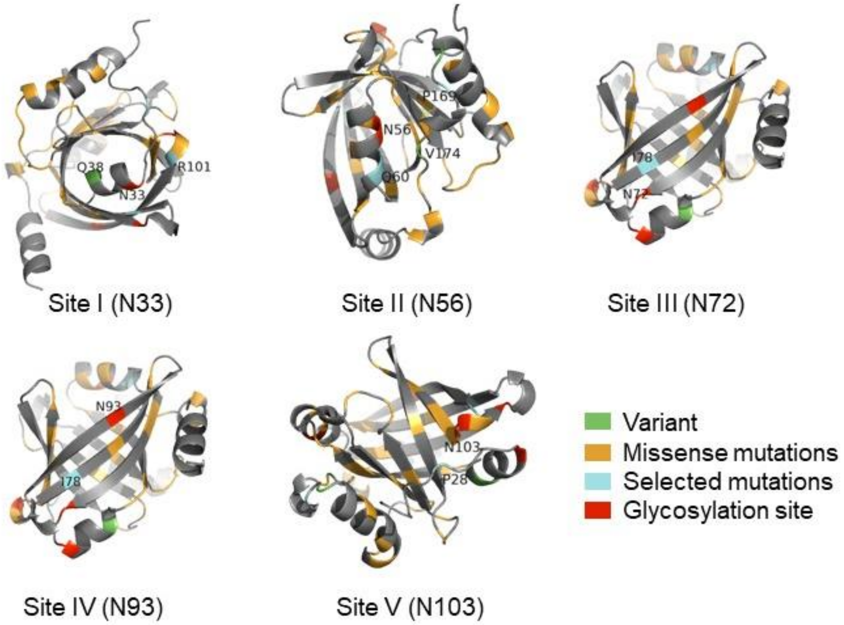


Figure S 3 The location of the single point mutations is coloured on the crystal structure of AGP (PDB: 3kq0): selected missense mutations (cyan), selected variant mutations (green), and other missense mutations (orange). Each view focuses on 1 glycosylation site, and close-by selected mutation sites are labelled with the residue number.

3. Overview of MD systems

The simulation box in the MD simulations contained the (un)glycosylated protein, water atoms, and salt ions.

Table S 1 Overview of systems studied with MD simulations.

Systems	Na^+ ions	Cl^- ions	Number of water molecules	Number of atoms
AGP	48	41	15741	46164
P28L	48	41	15741	46181
Q38R	47	41	15741	46179
Q60L	48	41	15741	46157
I78N	48	41	15741	46159
R101W	49	41	15741	46135
R167C	47	39	15056	44129
P169L	48	41	15741	46220
V174M	48	41	15741	46168
gly-AGP	108	92	34871	102209
gly-P28L	114	98	36914	108331
gly-Q38R	99	84	31859	93286
gly-Q60L	111	95	35837	105142
gly-I78N	114	98	36914	108285
gly-R101W	115	98	36914	108282
gly-R167C	118	101	38089	111863

gly-P169L	114	98	36914	108253
gly-V174M	117	101	38089	111777

4. RMSD

The RMSD based on the protein's C_α atoms can be used to detect large conformational changes with respect to the reference structure. We used the X-ray crystal structure of AGP as well as initial equilibrated structure as the reference structure. The results are calculated over the 100 ns per replica per system.

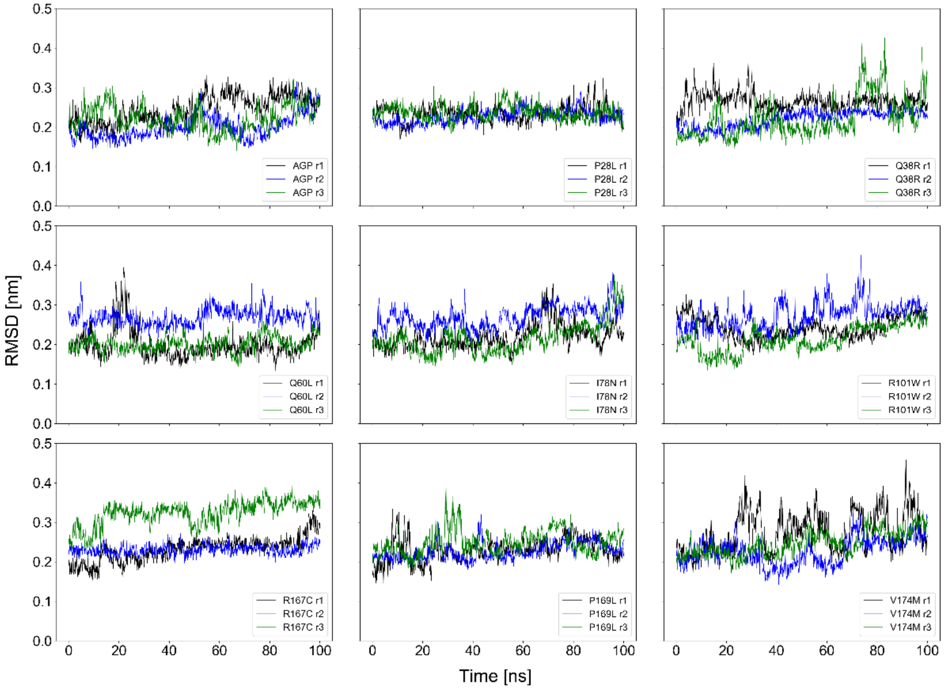


Figure S 4 C_α RMSD of unglycosylated structures of AGP and its mutants with respect to X-ray crystal structure of AGP during the NPT production run of 100 ns for replica 1 (black), replica 2 (blue), and replica 3 (green).

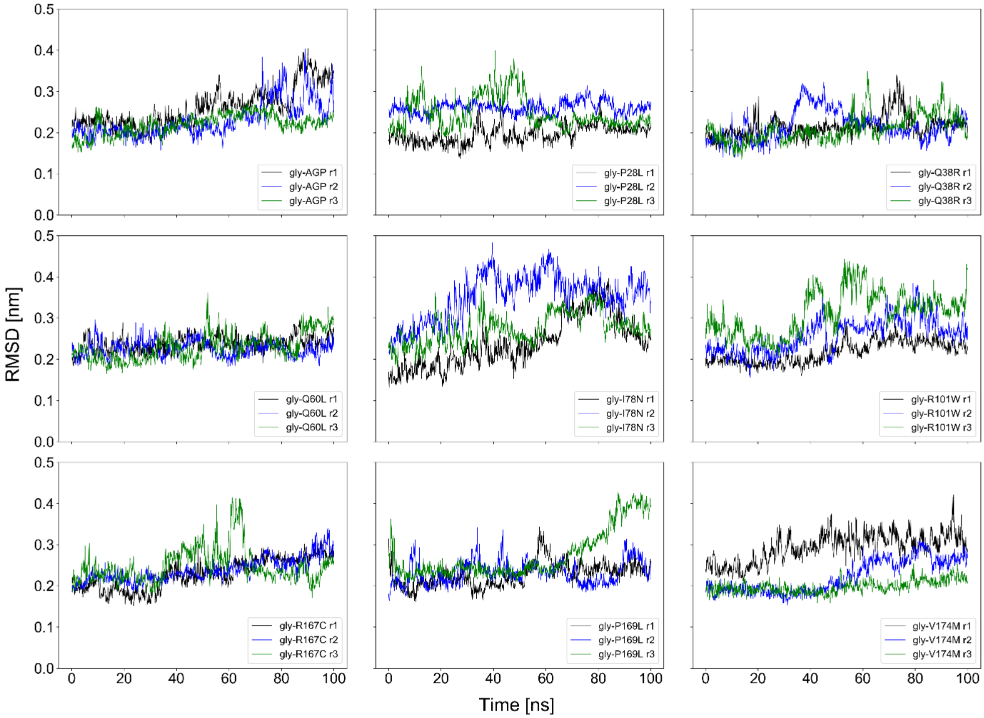


Figure S 5 C$_\alpha$ RMSD of glycosylated structures of AGP and its mutants with respect to X-ray crystal structure of AGP during the NPT production run of 100 ns for replica 1 (r1 as black), replica 2 (r2 as blue), and replica 3 (r3 as green).

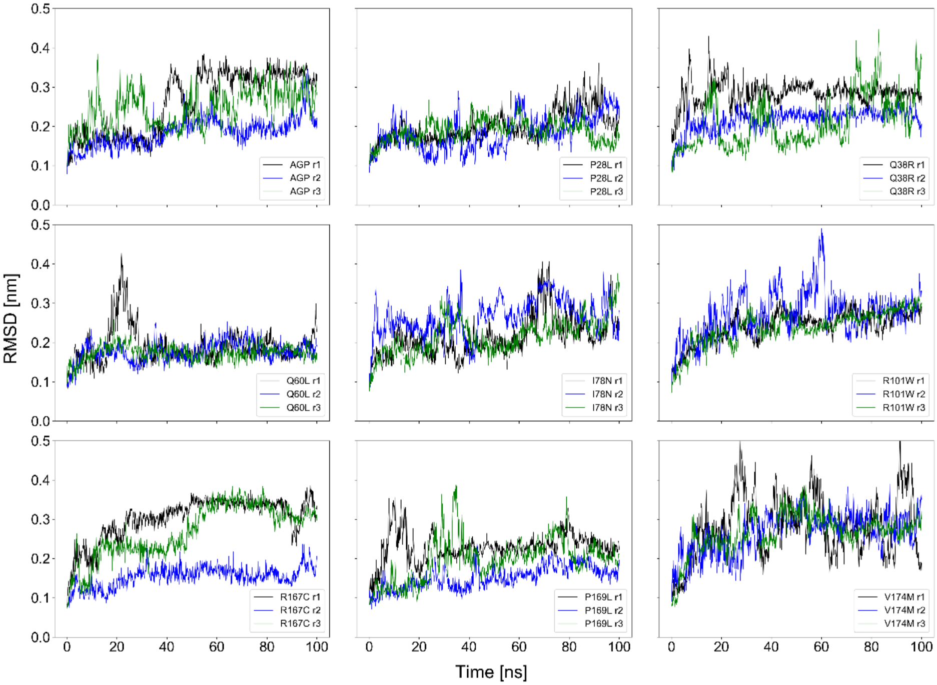


Figure S 6 C$_\alpha$ RMSD of unglycosylated structures of AGP and its mutants with respect to their initial equilibrated structure during the NPT production run of 100 ns for replica 1 (black), replica 2 (blue), and replica 3 (green).

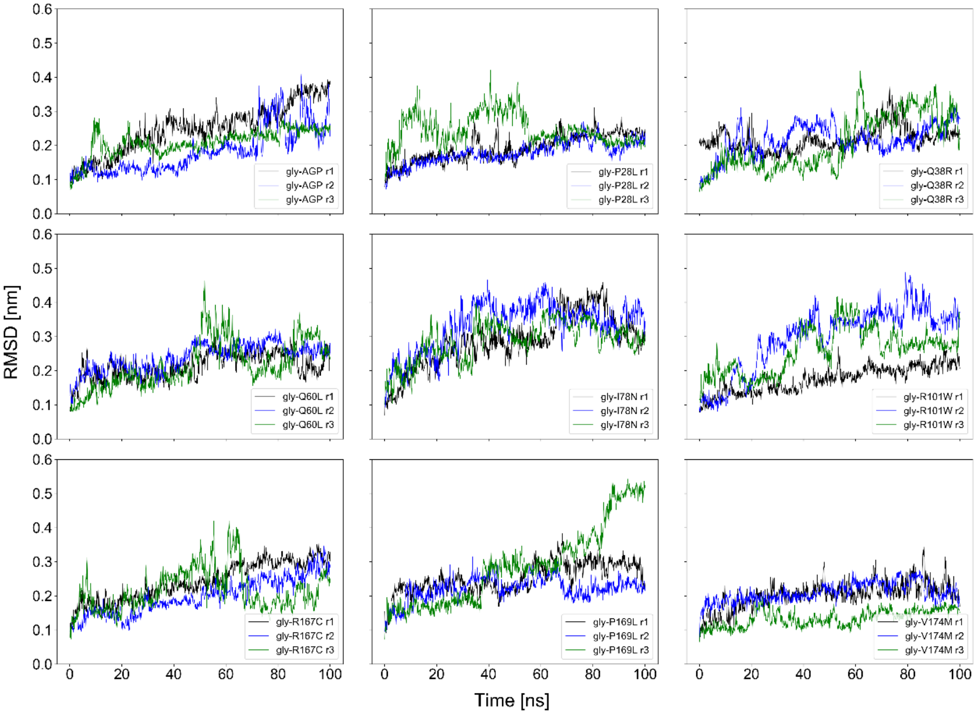


Figure S 7 C_α RMSD of glycosylated structures of AGP and its mutants with respect to their initial equilibrated structure during the NPT production run of 100 ns for replica 1 (black), replica 2 (blue), and replica 3 (green).

5. Radius of gyration

The radius of gyration (R_g) of the protein based on the protein's C_α atoms measures the compactness of the protein's backbone.

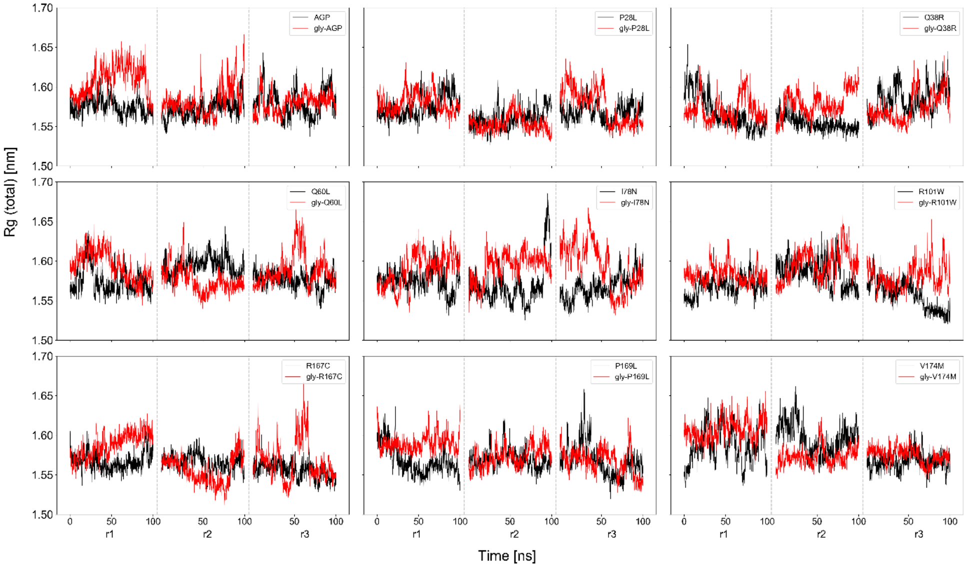


Figure S 8 Radius of gyration (R_g) of AGP and its mutants: unglycosylated (black), glycosylated (red). The radius of gyration is computed based on the protein's C_α atoms during the NPT production run of 100 ns from replica 1 to replica 3 (referred to as r1, r2, and r3).

6. RMSF

The RMSF of the protein's C_α atoms measures the fluctuations (in Cartesian coordinates) of the protein's backbone and is thus a metric for the backbone flexibility. The RMSF values in Figure S 10 are used to compute the RMSF of the fragments in the LBE, LBS, and hPPI. The RMSF for a fragment is obtained by summing over the RMSF values of the residues in the fragment.

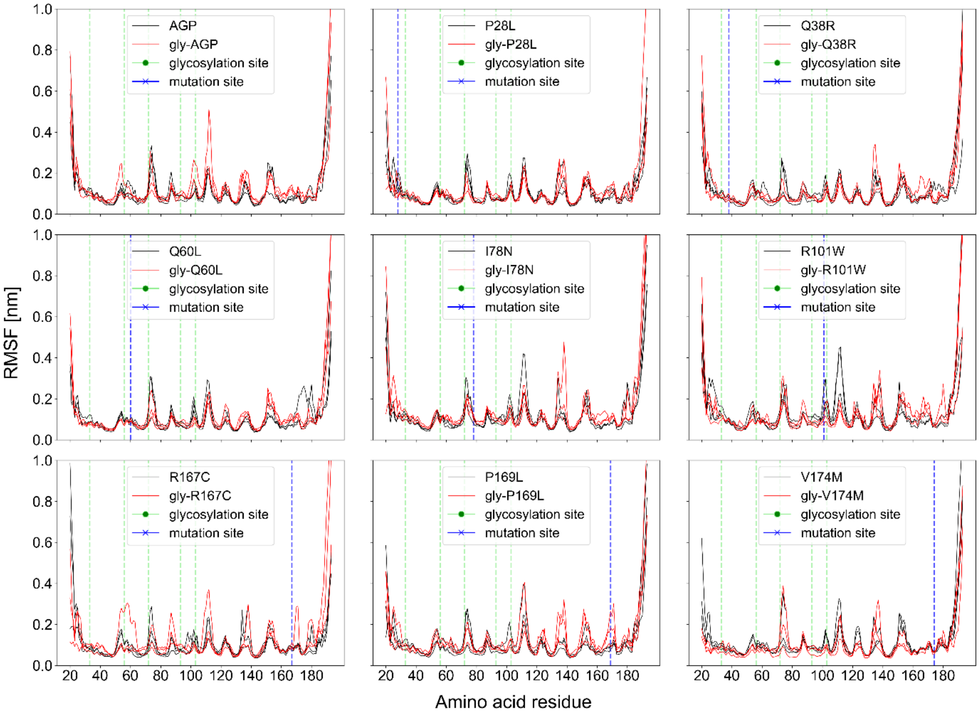


Figure S 9 RMSF of AGP and mutants as a function of residue number. The black curves represent replicas of unglycosylated systems and red curves represent replicas of glycosylated systems. RMSF was based on the protein's C_α positions during the NPT production run of 100 ns per replica. Glycosylation sites (green) and mutation sites (blue) are also indicated.

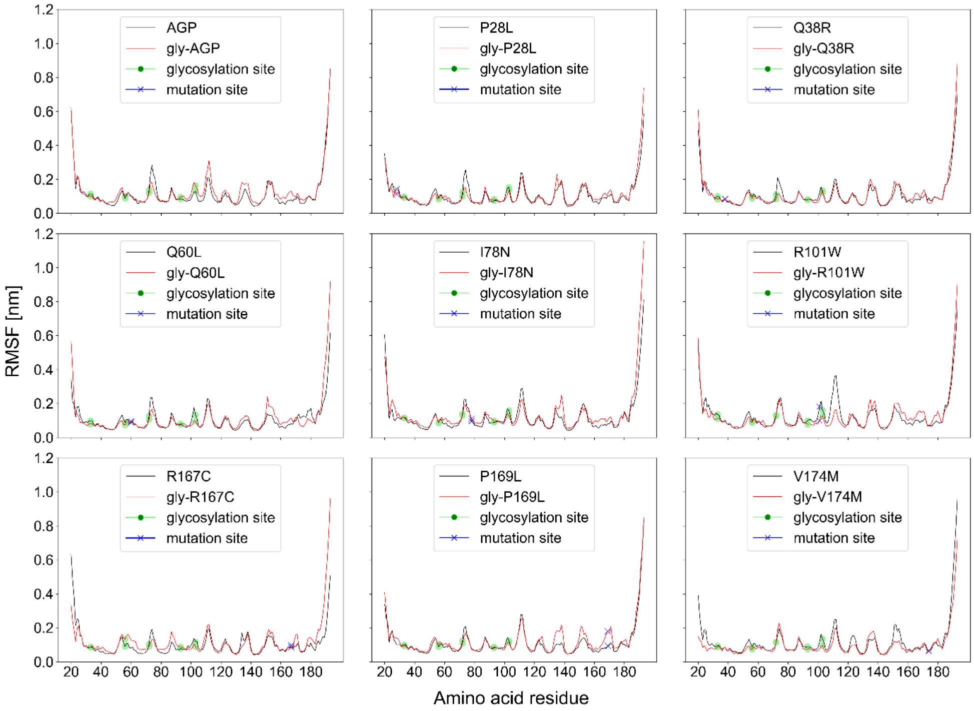


Figure S 10 RMSF (averaged over the RMSF of the 3 replicas as shown in Figure S9) of AGP and mutants as a function of residue number. Unglycosylated systems (black) and glycosylated systems (red). Glycosylation sites (green) and mutation sites (blue) are also indicated.

The data of Figure S 11 are summed for each line, giving four values per subplot of Figure S 12.

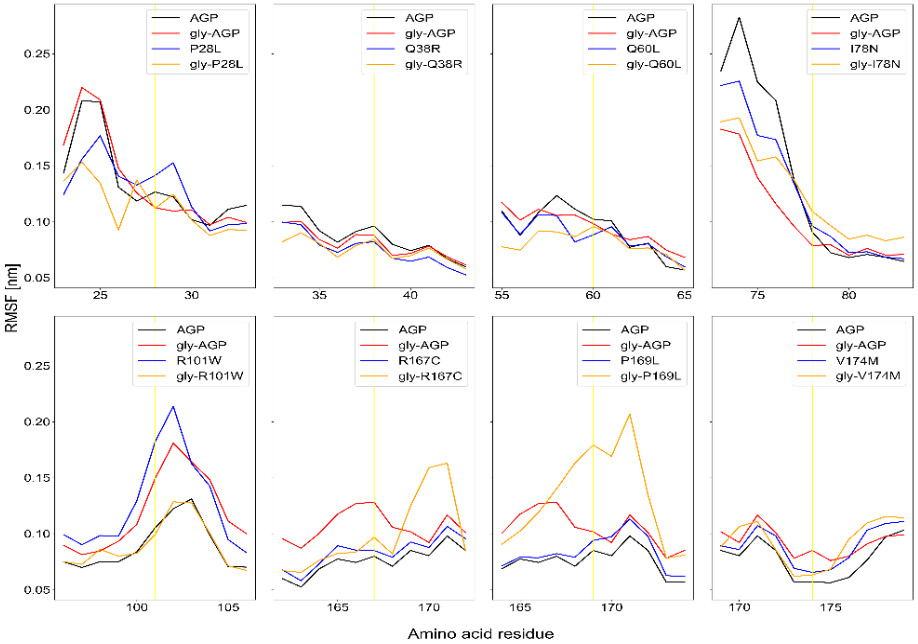


Figure S 11 RMSF (averaged over the 3 replicas) of each mutant compared to AGP, unglycosylated or glycosylated. Zooming in on 11 residues: 5 residues left and right to the mutation site. The first 5 mutations are sequentially close to glycosylation sites, the last 3 are not sequentially close to a glycosylation site.

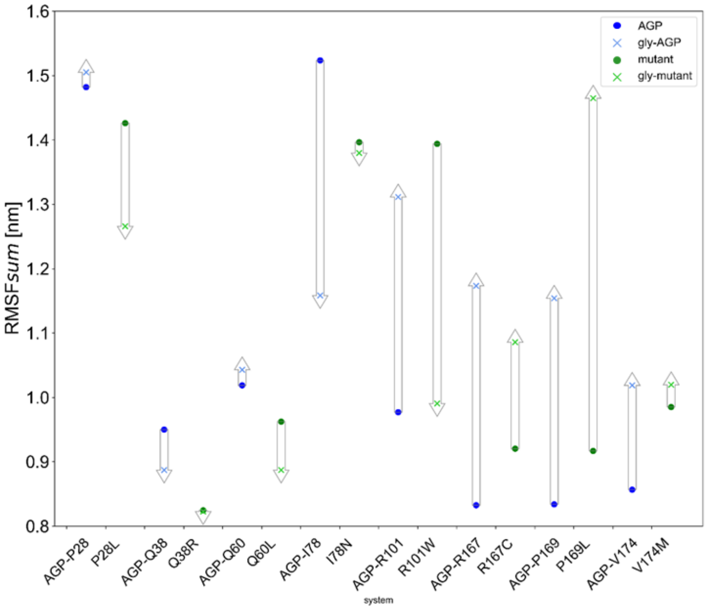


Figure S 12 Effect of glycosylation on the local flexibility of AGP and its mutants around the site of mutation.

The dark blue line lies above the light blue line. The RMSF is therefore increased locally around the mutation sites.

7. Average values for RMSD, RMSF, *R*g

Table S 2 Summary of parameters with mean and standard deviation for all systems of AGP calculated over 100 ns per replica.

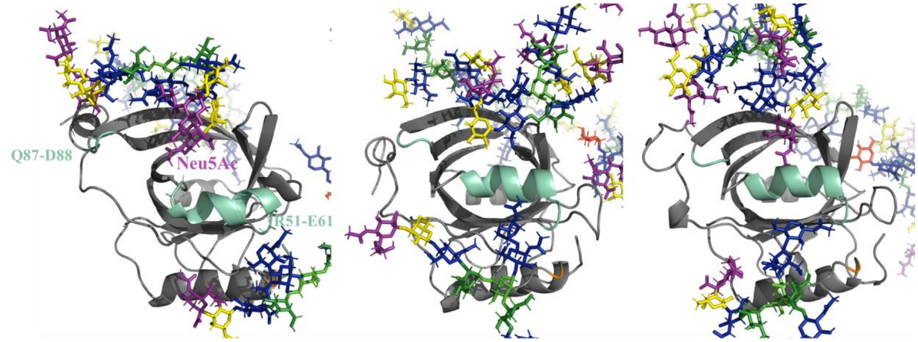
systems	**RMSF [nm]**			**RMSD$_{pdb}$ [nm]**			**RMSD$_{init}$ [nm]**			***R*g [nm]**		
	r1	r2	r3	r1	r2	r3	r1	r2	r3	r1	r2	r3
AGP	0.13 ± 0.11	0.10 ± 0.08	0.12 ± 0.13	0.25 ± 0.04	0.20 ± 0.03	0.22 ± 0.03	0.26 ± 0.08	0.19 ± 0.04	0.24 ± 0.05	1.58 ± 0.01	1.57 ± 0.01	1.58 ± 0.02
P28L	0.10 ± 0.08	0.11 ± 0.07	0.11 ± 0.07	0.23 ± 0.02	0.23 ± 0.02	0.24 ± 0.02	0.20 ± 0.04	0.18 ± 0.04	0.19 ± 0.03	1.58 ± 0.01	1.56 ± 0.01	1.57 ± 0.01
Q38R	0.12 ± 0.09	0.09 ± 0.06	0.12 ± 0.12	0.27 ± 0.02	0.22 ± 0.02	0.22 ± 0.05	0.28 ± 0.03	0.22 ± 0.03	0.21 ± 0.07	1.57 ± 0.02	1.55 ± 0.01	1.58 ± 0.02
Q60L	0.11 ± 0.09	0.10 ± 0.07	0.10 ± 0.07	0.20 ± 0.04	0.27 ± 0.02	0.20 ± 0.02	0.19 ± 0.05	0.18 ± 0.03	0.17 ± 0.02	1.57 ± 0.02	1.59 ± 0.01	1.58 ± 0.01
I78N	0.12 ± 0.09	0.14 ± 0.13	0.11 ± 0.10	0.22 ± 0.03	0.27 ± 0.03	0.21 ± 0.04	0.21 ± 0.05	0.27 ± 0.05	0.22 ± 0.05	1.58 ± 0.01	1.57 ± 0.03	1.57 ± 0.01
R101W	0.13 ± 0.09	0.14 ± 0.14	0.12 ± 0.09	0.24 ± 0.03	0.27 ± 0.03	0.21 ± 0.03	0.24 ± 0.04	0.27 ± 0.06	0.24 ± 0.04	1.57 ± 0.01	1.59 ± 0.02	1.56 ± 0.02
R167C	0.11 ± 0.09	0.09 ± 0.06	0.11 ± 0.11	0.23 ± 0.03	0.23 ± 0.01	0.32 ± 0.03	0.30 ± 0.05	0.15 ± 0.03	0.27 ± 0.07	1.57 ± 0.01	1.57 ± 0.01	1.56 ± 0.01
P169L	0.12 ± 0.11	0.09 ± 0.07	0.12 ± 0.12	0.23 ± 0.03	0.23 ± 0.02	0.25 ± 0.03	0.23 ± 0.04	0.15 ± 0.03	0.20 ± 0.05	1.57 ± 0.02	1.57 ± 0.01	1.57 ± 0.02
V174M	0.13 ± 0.14	0.11 ± 0.11	0.12 ± 0.09	0.28 ± 0.05	0.23 ± 0.03	0.25 ± 0.03	0.28 ± 0.08	0.26 ± 0.05	0.27 ± 0.05	1.59 ± 0.02	1.60 ± 0.02	1.57 ± 0.01
gly-AGP	0.14 ± 0.14	0.12 ± 0.11	0.12 ± 0.08	0.26 ± 0.05	0.23 ± 0.04	0.22 ± 0.02	0.25 ± 0.07	0.18 ± 0.07	0.21 ± 0.04	1.61 ± 0.02	1.58 ± 0.02	1.58 ± 0.01
gly-P28L	0.11 ± 0.08	0.1 ± 0.05	0.12 ± 0.13	0.20 ± 0.02	0.26 ± 0.02	0.25 ± 0.04	0.19 ± 0.04	0.18 ± 0.03	0.25 ± 0.05	1.58 ± 0.01	1.55 ± 0.01	1.58 ± 0.02
gly-Q38R	0.11 ± 0.10	0.13 ± 0.11	0.11 ± 0.11	0.21 ± 0.03	0.22 ± 0.04	0.21 ± 0.03	0.22 ± 0.03	0.22 ± 0.05	0.21 ± 0.08	1.57 ± 0.02	1.58 ± 0.02	1.57 ± 0.02
gly-Q60L	0.11 ± 0.12	0.10 ± 0.09	0.12 ± 0.12	0.24 ± 0.02	0.23 ± 0.02	0.23 ± 0.03	0.21 ± 0.04	0.24 ± 0.04	0.22 ± 0.07	1.59 ± 0.02	1.58 ± 0.02	1.59 ± 0.02
gly-I78N	0.13 ± 0.13	0.13 ± 0.13	0.15 ± 0.16	0.24 ± 0.06	0.35 ± 0.06	0.27 ± 0.04	0.29 ± 0.08	0.33 ± 0.07	0.29 ± 0.06	1.59 ± 0.02	1.59 ± 0.02	1.59 ± 0.03
gly-R101W	0.11 ± 0.07	0.14 ± 0.13	0.14 ± 0.13	0.22 ± 0.03	0.26 ± 0.04	0.31 ± 0.05	0.17 ± 0.04	0.30 ± 0.09	0.26 ± 0.07	1.58 ± 0.01	1.60 ± 0.02	1.58 ± 0.02
gly-R167C	0.12 ± 0.09	0.12 ± 0.11	0.12 ± 0.15	0.23 ± 0.04	0.24 ± 0.03	0.25 ± 0.04	0.24 ± 0.05	0.20 ± 0.05	0.22 ± 0.06	1.59 ± 0.02	1.56 ± 0.02	1.57 ± 0.03
gly-P169L	0.10 ± 0.08	0.12 ± 0.09	0.15 ± 0.13	0.23 ± 0.03	0.23 ± 0.03	0.27 ± 0.06	0.25 ± 0.04	0.22 ± 0.03	0.29 ± 0.11	1.59 ± 0.01	1.57 ± 0.01	1.57 ± 0.02
gly-V174M	0.11 ± 0.1	0.11 ± 0.09	0.08 ± 0.06	0.29 ± 0.03	0.22 ± 0.04	0.20 ± 0.02	0.20 ± 0.04	0.21 ± 0.03	0.13 ± 0.02	1.61 ± 0.01	1.57 ± 0.01	1.57 ± 0.01

Table S 3 Mean and standard deviation of RMSF$_{region}$ for all systems of AGP calculated from RMSF averaged over three replicas.

systems	RMSF$_{region}$		
	LBE [nm]	LBS [nm]	hPPI [nm]
AGP	0.12 ± 0.04	0.08 ± 0.03	0.10 ± 0.02
P28L	0.13 ± 0.05	0.08 ± 0.03	0.10 ± 0.02
Q38R	0.12 ± 0.03	0.08 ± 0.03	0.09 ± 0.02
Q60L	0.11 ± 0.04	0.07 ± 0.03	0.09 ± 0.02
I78N	0.12 ± 0.06	0.09 ± 0.04	0.10± 0.02
R101W	0.15 ± 0.08	0.10 ± 0.05	0.12 ± 0.02
R167C	0.12 ± 0.04	0.07 ± 0.03	0.09 ± 0.02
P169L	0.12 ± 0.06	0.08 ± 0.03	0.10 ± 0.01
V174M	0.12 ± 0.05	0.08 ± 0.03	0.10 ± 0.03
gly-AGP	0.14 ± 0.06	0.09 ± 0.04	0.10 ± 0.02

gly-P28L	0.13 ± 0.05	0.08 ± 0.02	0.09 ± 0.01
gly-Q38R	0.12 ± 0.04	0.07 ± 0.02	0.09 ± 0.02
gly-Q60L	0.11 ± 0.03	0.07 ± 0.02	0.08 ± 0.02
gly-I78N	0.14 ± 0.05	0.09 ± 0.03	0.11 ± 0.01
gly-R101W	0.13 ± 0.04	0.08 ± 0.03	0.11 ± 0.02
gly-R167C	0.14 ± 0.04	0.08 ± 0.03	0.09 ± 0.01
gly-P169L	0.15 ± 0.05	0.09 ± 0.03	0.10 ± 0.03
gly-V174M	0.11 ± 0.05	0.07 ± 0.03	0.08 ± 0.01

8. Visualization of MD snapshots



Replica 1 Replica 2 Replica 3

Figure S 13 MD snapshot of replica 1, 2, and 3 of gly-R167C around 60 ns showing the structural geometries of Q87-D88 and R51-E61 fragments (cyan) and α-Neu5Ac (purple) of glycan chain IV.

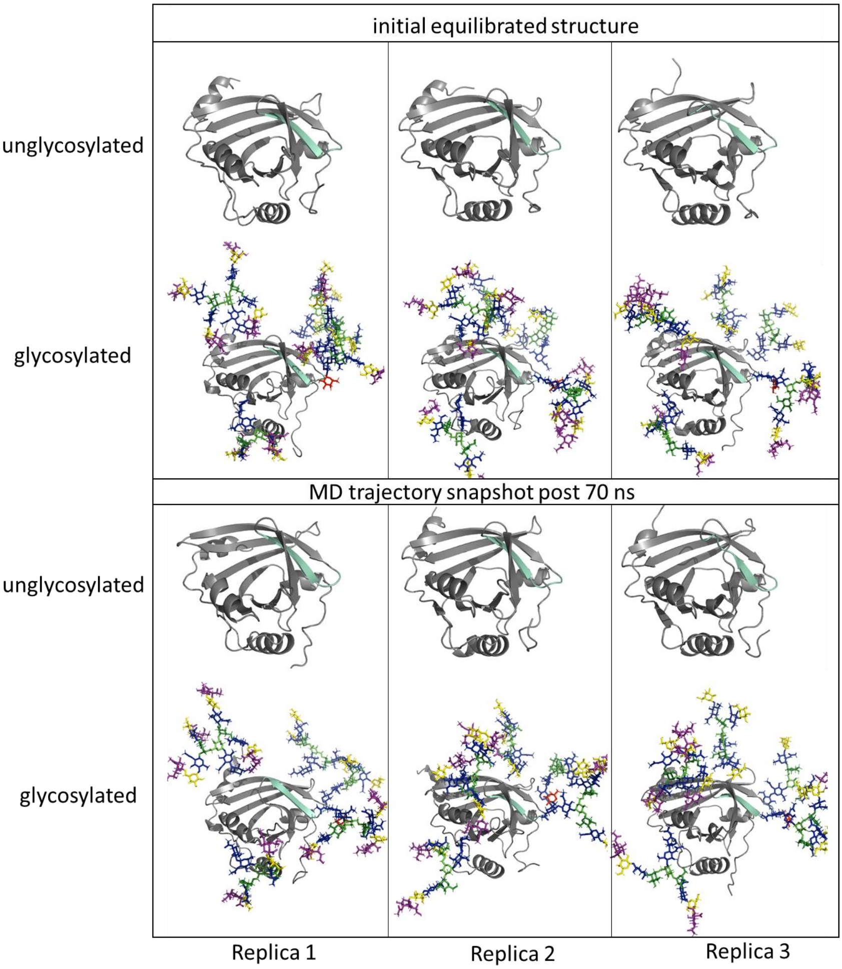


Figure S 14 MD snapshot of replica 1, 2, and 3 of AGP and gly-AGP showing the structural geometries of fragment E102-V110 (cyan) before (initial equilibrated structure) and during MD run (post 70 ns).

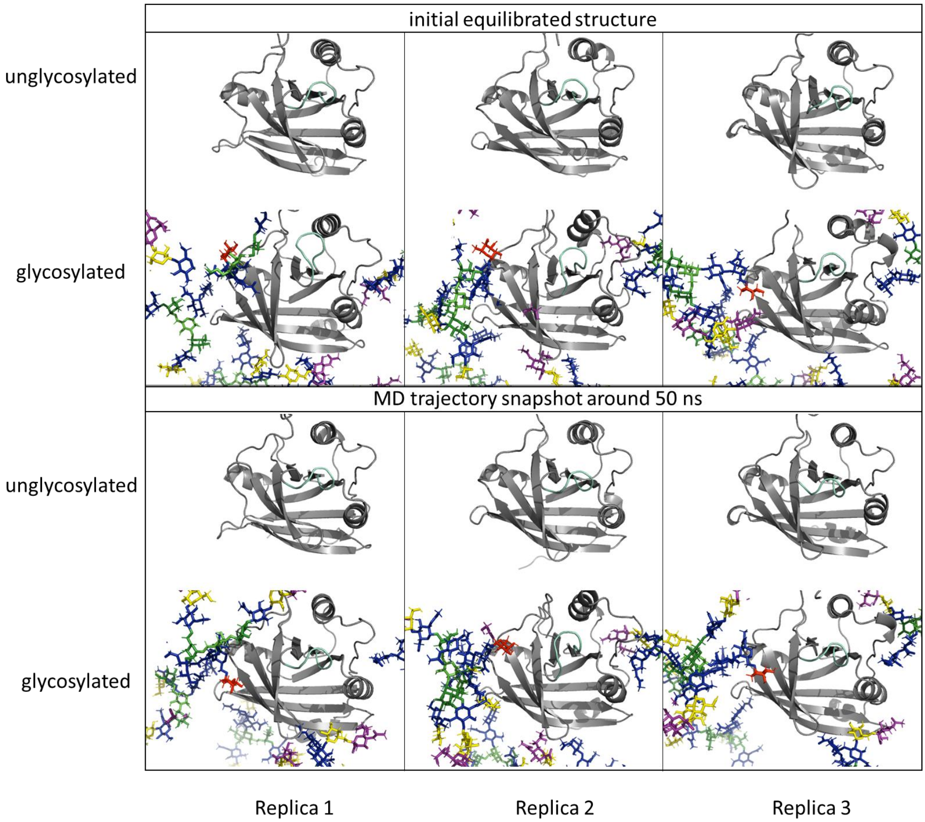


Figure S 15 MD snapshot of replica 1, 2, and 3 of unglycosylated and glycosylated I78N showing the structural geometries of N135-W140 fragment (cyan) before (initial equilibrated structure) and during MD run (around 50 ns).

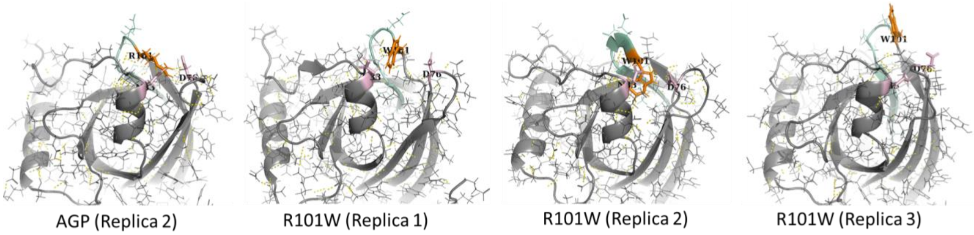


Figure S 16 Sidechain H-bond interactions between R101 (orange) and D76 and/or N33 (pink), in AGP (replica 2 snapshot as a reference), and W101 (orange) in replica 1, 2 and 3 of mutant R101W. The fragment E102-V110 is shown in cyan, and H-bonds are shown in yellow dotted lines.

9. Contact analysis

In Figure S 17, the fraction of C$_{\alpha}$ native contacts is shown with X-ray crystal structure as a reference with a cut-off distance of 0.45 nm. The total number of contacts in X-ray crystal structure is 558. In Figure S 18, number of C$_{\alpha}$ contacts within a 0.45 nm radius in each system is shown as a function of time.

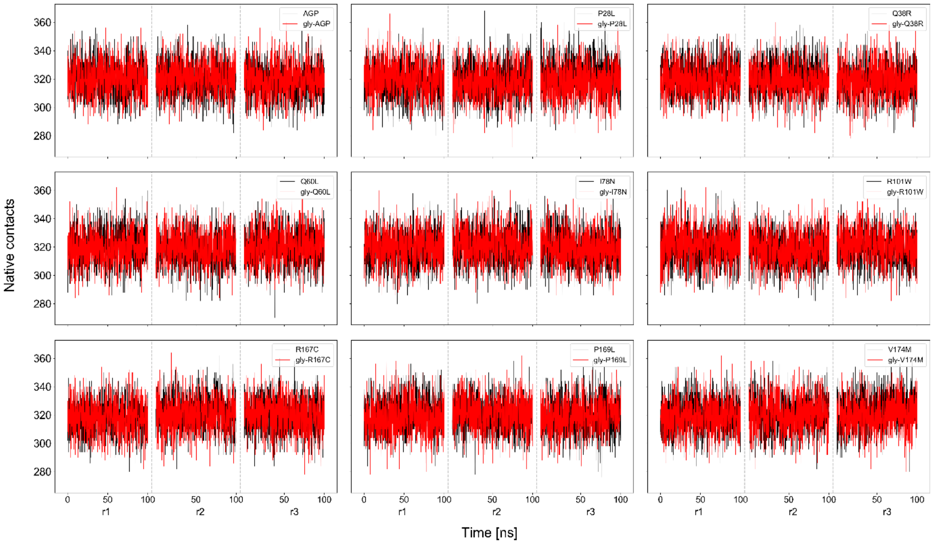


Figure S 17 Number of C$_{\alpha}$ native contacts calculated over 100 ns trajectory of systems with X-ray structure as reference and a cut-off distance of 0.45 nm. The unglycosylated systems are shown in grey, and the glycosylated systems in red. The native contacts are computed based on the protein's C$_{\alpha}$ atoms during the NPT production run of 100 ns from replica 1, replica 2, and replica 3 (referred to as r1, r2, and r3).

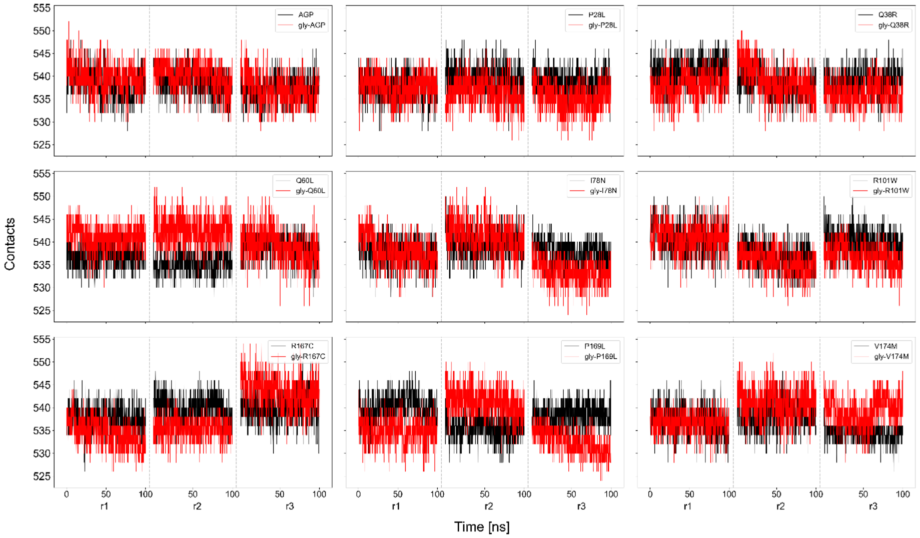


Figure S 18 Number of C$_{\alpha}$ contacts calculated over 100 ns trajectory of systems with initial equilibrated structure as reference and a cut-off radius of 0.45 nm. The unglycosylated systems are shown in grey, and the glycosylated systems in red. The native contacts are computed based on the protein's C$_\alpha$ atoms during the NPT production run of 100 ns from replica 1, replica 2, and replica 3 (referred to as r1, r2, and r3).

Table S 4 Mean and standard deviation of contacts of all systems of AGP per replica r1, r2, and r3 calculated over 100 ns. Native contacts are calculated with X-ray crystal structure of AGP as a reference with a hard cut-off distance of 0.45 nm, while contacts are calculated with their corresponding NPT equilibrated structure as reference with a radius of 0.45 nm.

	Native contacts (X-ray structure)			Overall contacts (Equilibrated structure)		
Systems	r1	r2	r3	r1	r2	r3
AGP	320 ± 12	320 ± 12	318 ± 12	539 ± 3	539 ± 3	537 ± 3
P28L	318 ± 12	319 ± 12	322 ± 12	538 ± 3	539 ± 3	539 ± 3
Q38R	322 ± 11	321 ± 11	318 ± 11	541 ± 3	538 ± 3	539 ± 3
Q60L	320 ± 12	318 ± 11	320 ± 12	537 ± 3	535 ± 2	539 ± 3
I78N	318 ± 11	320 ± 12	319 ± 12	538 ± 3	540 ± 3	538 ± 3
R101W	319 ± 12	318 ± 11	319 ± 11	541 ± 3	537 ± 3	540 ± 3
R167C	318 ± 11	322 ± 12	320 ± 12	538 ± 3	540 ± 3	540 ± 3
P169L	320 ± 11	319 ± 11	319 ± 12	540 ± 3	536 ± 3	538 ± 3
V174M	320 ± 12	321 ± 12	323 ± 12	537 ± 3	538 ± 3	534 ± 2
gly-AGP	321 ± 11	321 ± 11	319 ± 11	540 ± 3	540 ± 3	537 ± 3
gly-P28L	320 ± 12	318 ± 11	318 ± 12	538 ± 3	536 ± 3	535 ± 3
gly-Q38R	320 ± 11	319 ± 11	319 ± 12	538 ± 3	539 ± 4	536 ± 3
gly-Q60L	319 ± 11	320 ± 11	320 ± 12	541 ± 3	543 ± 3	539 ± 4
gly-I78N	320 ± 11	321 ± 11	319 ± 11	538 ± 4	541 ± 4	533 ± 3
gly-R101W	321 ± 12	319 ± 11	321 ± 11	541 ± 3	536 ± 3	537 ± 3
gly-R167C	319 ± 12	320 ± 12	320 ± 12	535 ± 3	535 ± 2	544 ± 4
gly-P169L	318 ± 12	321 ± 11	320 ± 11	535 ± 3	541 ± 3	532 ± 3
gly-V174M	318 ± 11	321 ± 12	320 ± 11	537 ± 3	541 ± 4	539 ± 3

10. SASA

The values of protein and glycan SASA are reported in the sections below.

10.1 SASA of protein atoms

In Table S 5, the mean and standard deviation of SASA$_{\text{region}}$ (as explained in methods) is shown for the LBE, LBS, and hPPI regions for all systems.

Table S 5 Mean and standard deviation of SASA$_{\text{region}}$ of LBE, LBS, and hPPI for all systems of AGP calculated for three different probe sizes 0.14 nm, 0.5 nm, and 1.0 nm over 300 ns.

systems	**0.14 nm probe**			**0.5 nm probe**			**1.0 nm probe**		
	LBE [nm]^2	LBS [nm]^2	hPPI [nm]^2	LBE [nm]^2	LBS [nm]^2	hPPI [nm]^2	LBE [nm]^2	LBS [nm]^2	hPPI [nm]^2
AGP	18.13 ± 1.04	34.87 ± 1.89	11.93 ± 0.72	20.72 ± 1.66	27.44 ± 1.71	12.52 ± 0.98	28.61 ± 2.84	32.57 ± 2.43	16.04 ± 1.60
P28L	17.79 ± 0.72	33.96 ± 1.53	12.08 ± 0.77	20.32 ± 1.14	26.89 ± 1.83	12.58 ± 0.91	27.86 ± 1.89	32.08 ± 2.90	16.11 ± 1.71
Q38R	17.27 ± 0.74	33.64 ± 1.78	11.73 ± 0.69	19.30 ± 1.04	27.11 ± 2.10	12.42 ± 0.85	26.61 ± 1.60	32.24 ± 3.00	15.95 ± 1.49
Q60L	18.01 ± 0.78	34.31 ± 1.46	11.58 ± 0.60	19.84 ± 1.07	27.41 ± 1.56	12.24 ± 0.81	27.30 ± 1.78	32.38 ± 2.42	15.66 ± 1.33
I78N	17.43 ± 1.08	33.84 ± 1.84	11.57 ± 0.79	19.44 ± 1.64	27.63 ± 2.11	12.16 ± 1.00	26.73 ± 2.38	32.96 ± 3.26	15.42 ± 1.70
R101W	17.23 ± 0.93	34.64 ± 1.71	11.69 ± 0.63	19.21 ± 1.37	28.46 ± 1.73	11.95 ± 0.87	26.15 ± 2.08	34.72 ± 3.36	15.37 ± 1.61
R167C	17.18 ± 1.15	33.42 ± 1.58	11.31 ± 0.64	19.46 ± 1.52	26.96 ± 2.04	11.87 ± 0.93	26.78 ± 2.10	32.28 ± 2.98	15.04 ± 1.67
P169L	17.39 ± 0.95	33.64 ± 1.70	11.48 ± 0.67	19.43 ± 1.16	27.06 ± 1.85	12.13 ± 0.92	26.49 ± 1.73	31.96 ± 2.93	15.40 ± 1.53

V174M	17.27 ± 1.06	32.88 ± 1.21	11.10 ± 0.72	19.55 ± 1.27	27.21 ± 1.49	11.86 ± 0.93	27.19 ± 1.77	32.45 ± 2.69	15.21 ± 1.53
gly-AGP	15.11 ± 1.47	27.75 ± 1.83	10.76 ± 0.67	12.15 ± 2.34	13.15 ± 2.12	10.26 ± 0.86	11.67 ± 3.16	9.57 ± 2.32	11.79 ± 1.47
gly-P28L	15.21 ± 1.18	26.97 ± 1.84	11.07 ± 0.62	11.46 ± 2.22	12.70 ± 1.86	10.80 ± 0.80	9.88 ± 3.19	9.42 ± 1.92	12.84 ± 1.42
gly-Q38R	15.37 ± 1.19	26.12 ± 1.50	10.76 ± 0.59	12.15 ± 2.20	11.77 ± 1.89	10.82 ± 0.78	10.63 ± 3.81	8.57 ± 2.14	12.63 ± 1.40
gly-Q60L	15.01 ± 1.45	28.11 ± 1.56	10.32 ± 0.81	11.79 ± 1.94	12.90 ± 2.22	9.52 ± 1.26	10.10 ± 2.77	8.64 ± 2.16	10.55 ± 2.06
gly-I78N	16.18 ± 0.97	28.42 ± 1.63	10.64 ± 0.64	13.99 ± 1.85	12.90 ± 1.77	10.06 ± 0.99	13.60 ± 3.00	9.08 ± 2.11	11.68 ± 1.70
gly-R101W	15.28 ± 1.13	27.10 ± 1.86	10.82 ± 0.74	12.04 ± 1.79	12.05 ± 1.93	9.69 ± 1.19	11.53 ± 2.79	9.65 ± 2.10	10.80 ± 1.92
gly-R167C	15.41 ± 1.22	26.92 ± 2.20	10.88 ± 0.69	13.67 ± 2.01	11.87 ± 2.21	10.42 ± 1.02	14.27 ± 3.47	8.52 ± 2.11	12.21 ± 1.75
gly-P169L	15.08 ± 1.45	27.62 ± 1.67	10.74 ± 0.84	11.61 ± 3.08	12.39 ± 1.84	10.22 ± 1.41	10.53 ± 4.72	8.79 ± 2.07	11.79 ± 2.24
gly-V174M	14.72 ± 1.37	26.56 ± 1.72	10.57 ± 0.63	12.24 ± 1.75	12.03 ± 1.48	10.12 ± 0.84	11.81 ± 2.19	8.95 ± 1.81	11.67 ± 1.46

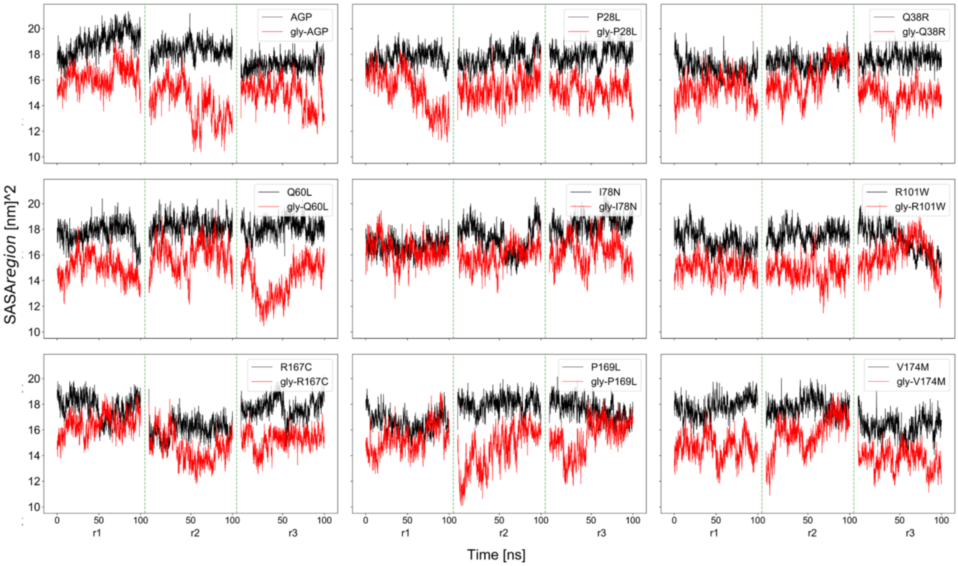


Figure S 19 SASA$_{region}$ of LBE calculated during the NPT production run of 100 ns from replica 1, replica 2, and replica 3 (referred to as r1, r2, and r3) for all systems with 0.14 nm probe. The unglycosylated systems are shown in black, and the glycosylated systems in red.

10.2 SASA of glycans

Glycan shielding is shown in Figure S 20. The number of glycan atoms of chain IV within 0.26 nm around LBE atoms is computed for every snapshot. The higher this number, the more glycan chain IV shields access to the LBE.

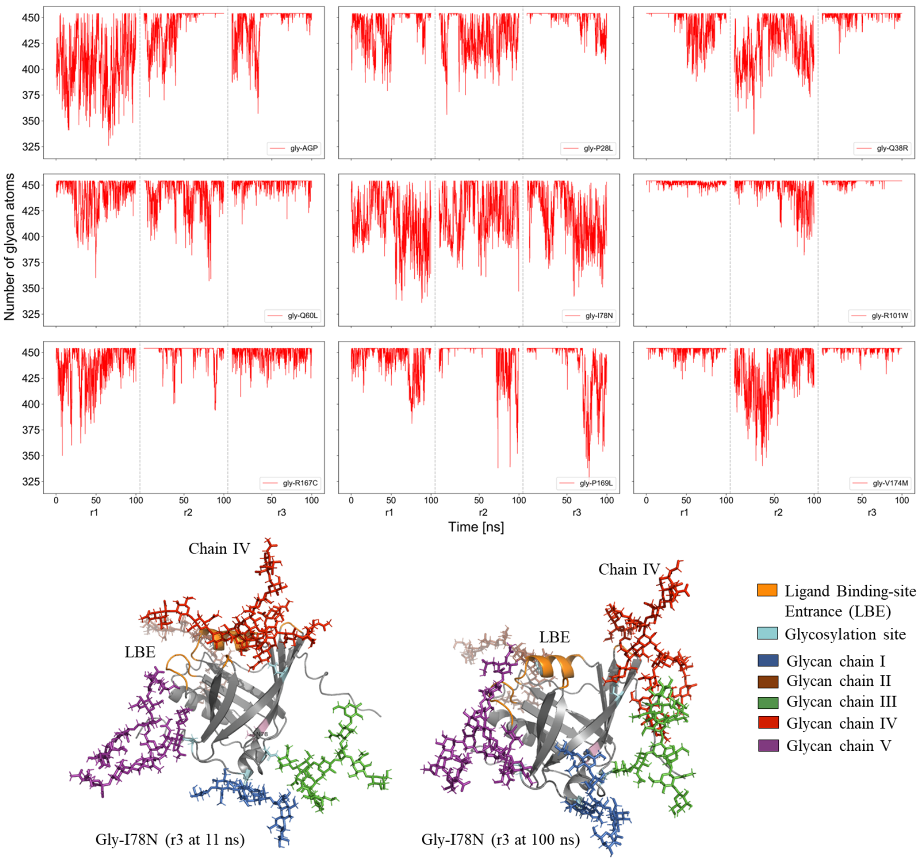
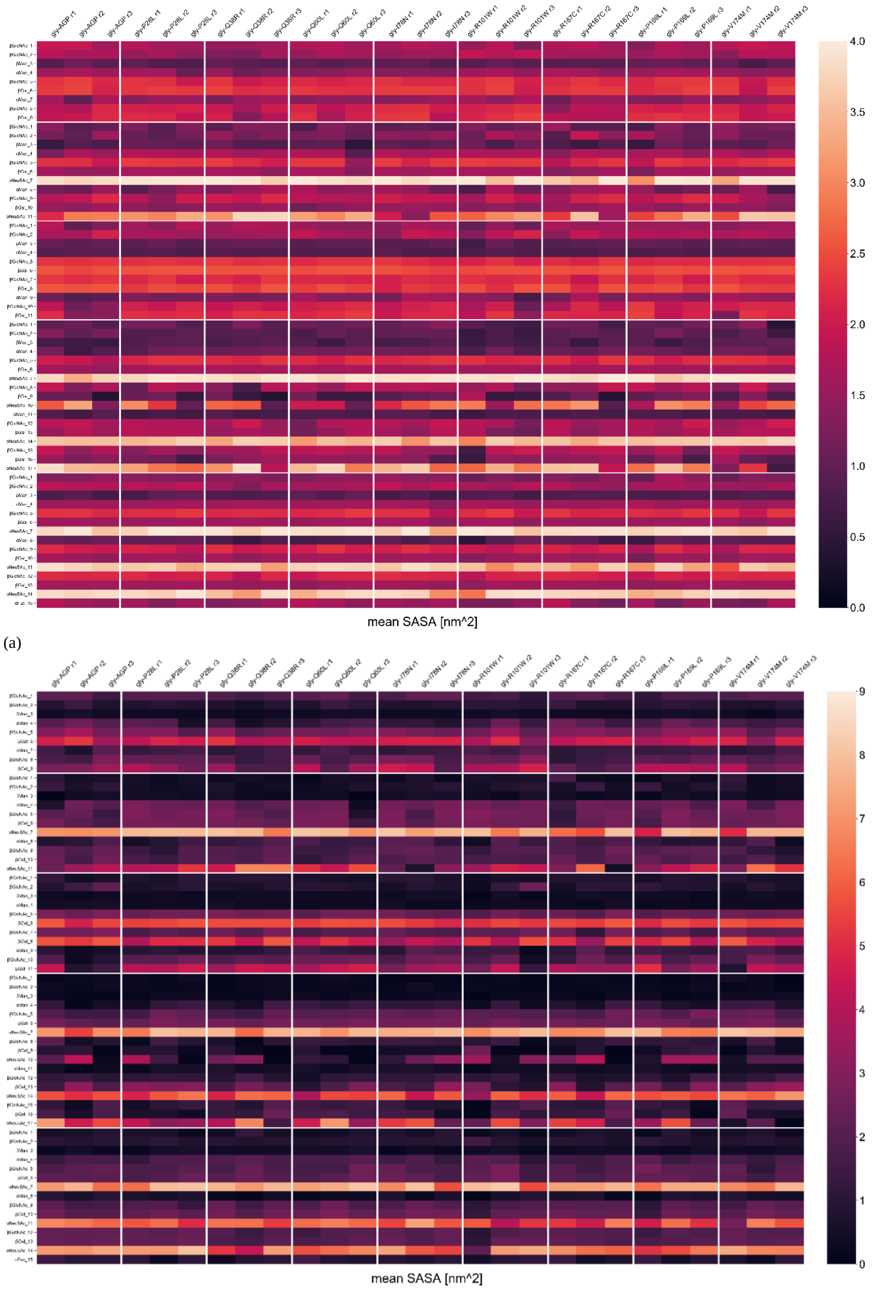


Figure S 20 Number of glycan atoms calculated during the NPT production run of 100 ns from replica 1 to replica 3 (referred to as r1, r2, and r3) for all systems with 0.14 nm probe. The glycosylated systems are shown in red. In the bottom MD snapshots of gly-I78N (replica 3) are shown at 11 ns (left) and at 100 ns (right).

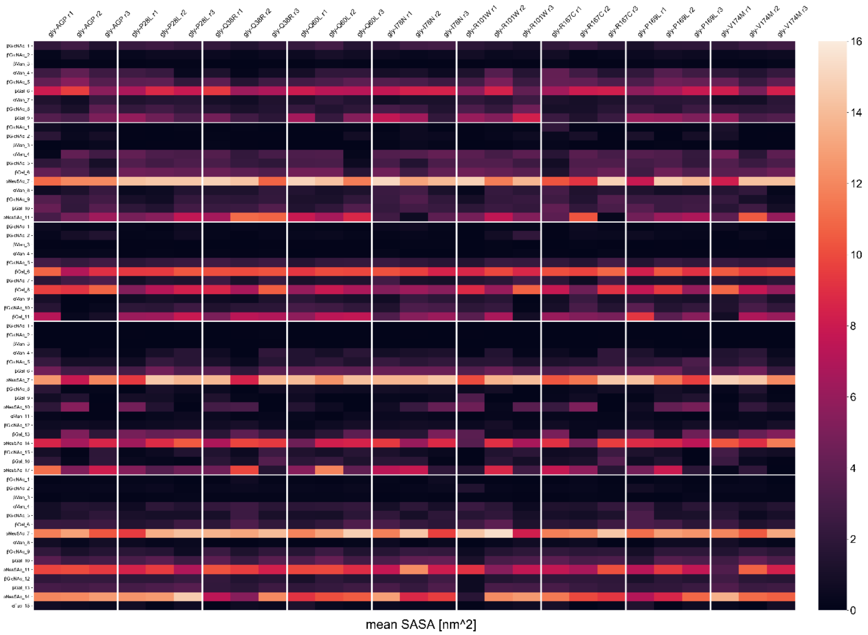
The SASA of glycans for all glycan chains for all glycosylated systems are shown in Figure S 21 for 0.14 nm, 0.5 nm, and 1.0 nm probes.


gly-AGP r1
gly-AGP r2
gly-AGP r3
gly-P28L r1
gly-P28L r2
gly-P28L r3
gly-Q38R r1
gly-Q38R r2
gly-Q38R r3
gly-Q60L r1
gly-Q60L r2
gly-Q60L r3
gly-I78N r1
gly-I78N r2
gly-I78N r3
gly-R101W r1
gly-R101W r2
gly-R101W r3
gly-R167C r1
gly-R167C r2
gly-R167C r3
gly-P169L r1
gly-P169L r2
gly-P169L r3
gly-V174M r1
gly-V174M r2
gly-V174M r3
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
αMan_7
βGlcNAc_8
βGal_9
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
αNeu5Ac_7
αMan_8
βGlcNAc_9
βGal_10
αNeu5Ac_11
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
βGlcNAc_7
βGal_8
αMan_9
βGlcNAc_10
βGal_11
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
αNeu5Ac_7
βGlcNAc_8
βGal_9
αNeu5Ac_10
αMan_11
βGlcNAc_12
βGal_13
αNeu5Ac_14
βGlcNAc_15
βGal_16
αNeu5Ac_17
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
αNeu5Ac_7
αMan_8
βGlcNAc_9
βGal_10
αNeu5Ac_11
βGlcNAc_12
βGal_13
αNeu5Ac_14
αFuc_15
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
mean SASA [nm^2]

(a)

gly-AGP r1
gly-AGP r2
gly-AGP r3
gly-P28L r1
gly-P28L r2
gly-P28L r3
gly-Q38R r1
gly-Q38R r2
gly-Q38R r3
gly-Q60L r1
gly-Q60L r2
gly-Q60L r3
gly-I78N r1
gly-I78N r2
gly-I78N r3
gly-R101W r1
gly-R101W r2
gly-R101W r3
gly-R167C r1
gly-R167C r2
gly-R167C r3
gly-P169L r1
gly-P169L r2
gly-P169L r3
gly-V174M r1
gly-V174M r2
gly-V174M r3
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
αMan_7
βGlcNAc_8
βGal_9
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
αNeu5Ac_7
αMan_8
βGlcNAc_9
βGal_10
αNeu5Ac_11
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
βGlcNAc_7
βGal_8
αMan_9
βGlcNAc_10
βGal_11
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
αNeu5Ac_7
βGlcNAc_8
βGal_9
αNeu5Ac_10
αMan_11
βGlcNAc_12
βGal_13
αNeu5Ac_14
βGlcNAc_15
βGal_16
αNeu5Ac_17
βGlcNAc_1
βGlcNAc_2
βMan_3
αMan_4
βGlcNAc_5
βGal_6
αNeu5Ac_7
αMan_8
βGlcNAc_9
βGal_10
αNeu5Ac_11
βGlcNAc_12
βGal_13
αNeu5Ac_14
αFuc_15
9
8
7
6
5
4
3
2
1
0
mean SASA [nm^2]

(b)



(c)

Figure S 21 SASA of glycan chains calculated during the NPT production run of 100 ns from replica 1, replica 2, and replica 3 (referred to as r1, r2, and r3) for all glycosylated systems for a) 0.14 nm probe b) 0.5 nm probe, and c) 1.0 nm probe. The shown mean SASA is the average value over the 100 ns time interval.

11 H-bonds (protein)

In Figure S 23 (a) and (b), the number of H-bonds are calculated between sidechains of Y68 and E187 referred to as HB1, Y83 and H190 (HB2), respectively, which are observed in X-ray crystal structure of AGP (Figure S 22).

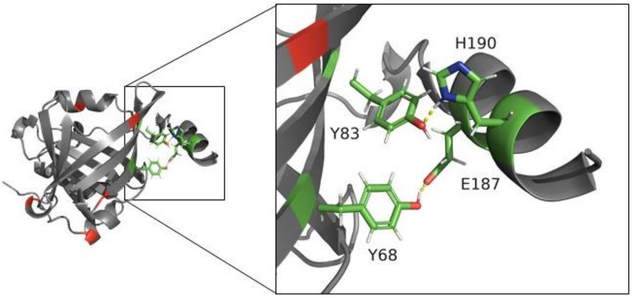
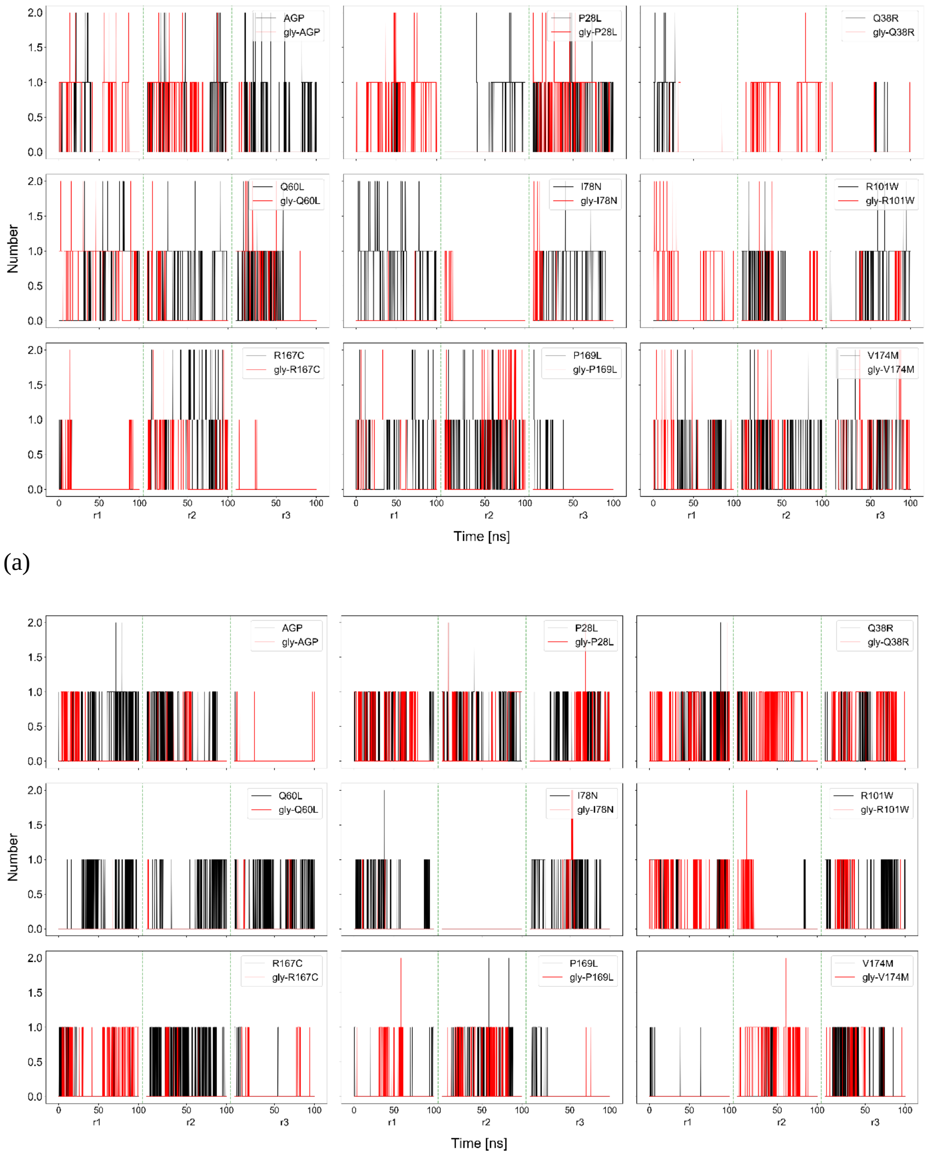


Figure S 22 H-bonds in the X-ray crystal structure of AGP: between Y68 and E187, and Y83 and H190.



(a)

(b)

Figure S 23 Number of H-bonds between the sidechains of (a) Y68 and E187, and (b) Y83 and H190 over 100 ns per replica for AGP and its mutants: glycosylated (red) and unglycosylated (black). The replica 1 to replica 3 are referred to as r1, r2, and r3.

12 Principal component analysis (PCA)

In Figure S 25, the cartesian coordinates of C_α atoms is projected on the first three principal modes (eigenvectors with highest eigenvalues) which capture approximately 50% to 60% motions in all unglycosylated and glycosylated systems demonstrated by explained variance ratio (Figure S 24). The explained variance ratio describes the variance explained by each PC in the data and is useful for finding the number of PCs to retain in PCA. The projected coordinates show the fluctuations around the mean position in conformational space of all systems.

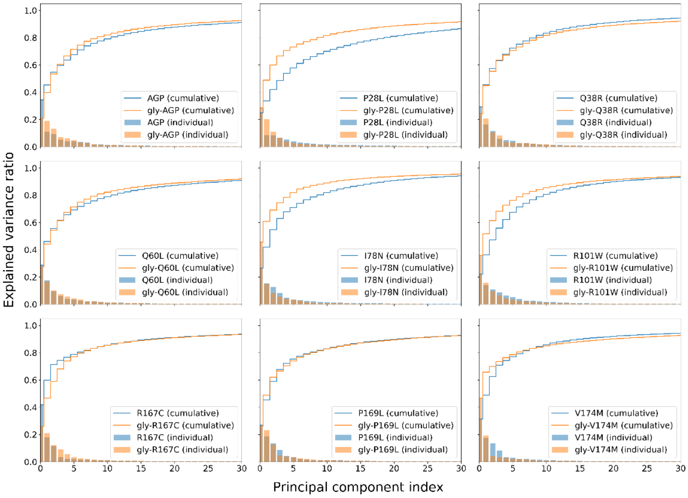


Figure S 24 Scree plots demonstrating explained variance and eigenvalues for 30 principal components (PCs) from PCA for AGP and its mutants: unglycosylated (blue), and glycosylated (red) calculated over 300 ns. Principal Component is shown on the x-axis and explained variance ration is shown on the y-axis. The bars represent individual explained variance ration, and the curves represent cumulative explained variance ration.

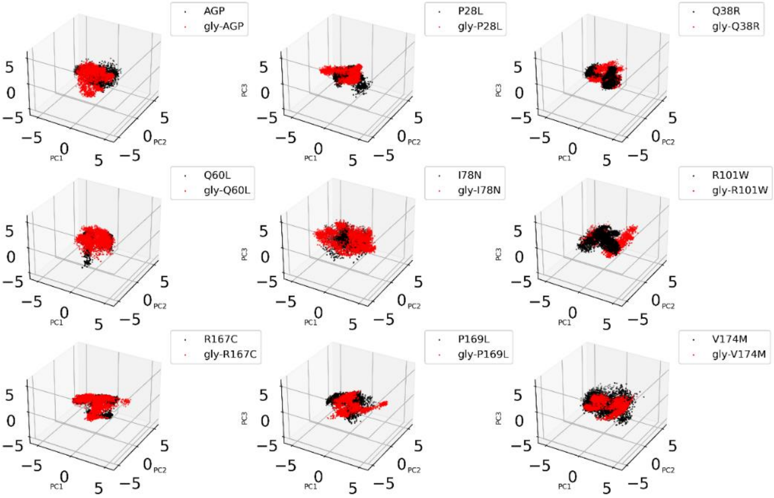


Figure S 25 The PCA plot generated by projecting the 300 ns MD trajectory of C$_\alpha$ atoms on the first three PCs (eigenvectors) of AGP and its mutants: unglycosylated (black) and glycosylated (red). The units of PC are in nm.

In addition, the root-mean-square-inner product (RMSIP as explained in Methods section) is computed for various subspaces, to show the overlap between those subspaces. The subspace covers a single PC (A=B=1 in Eq. 5 of main document) up to a total of 10 PCs (A=B=10 in Eq. 5). For instance, the RMSIP matrix for a subspace with 3 PCs (A=B=3 in Eq.6) is shown in Figure 5 of the main document.

These RMSIP matrices are divided in 3 blocks: RMSIP values among 8 unglycosylated mutants (left upper block), among 8 glycosylated mutants (right lower block), and between 8 unglycosylated and 8 glycosylated mutants (right upper block). In Figure S 26, the distribution of the RMSIP values of each of those three subblocks (excluding trivial 1.0 overlap values) is shown: between unglycosylated and glycosylated mutants (green), unglycosylated mutants (black) and glycosylated mutants (red).

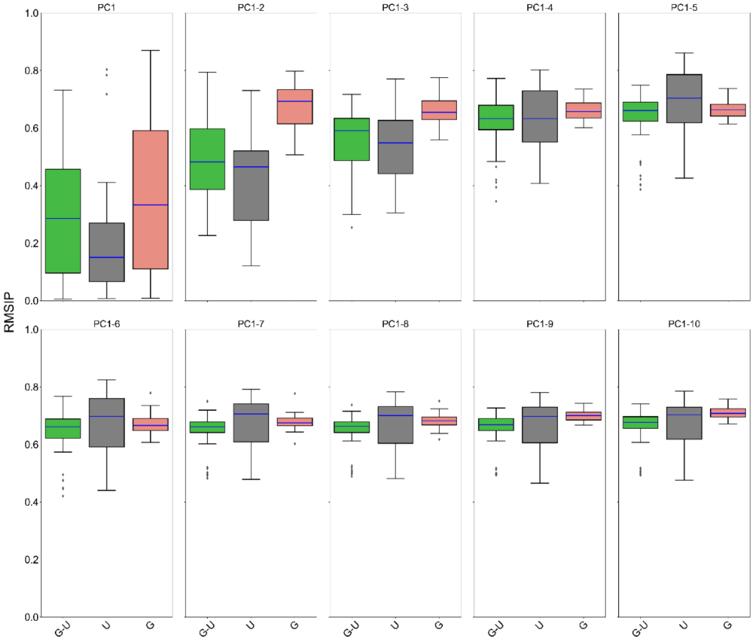


Figure S 26 Distribution of RMSIP scores between subspaces spanning 1 single PC (i.e., PC1) up to 10 PCs (PC1 to PC10): between unglycosylated and glycosylated mutants (‘GU’ as green), for unglycosylated mutants (‘U’ as black) and glycosylated mutants (‘G’ as red). Blue line indicates the median, the boxes represent the interquartile range (IQR), and whiskers represent the range of data within 1.5 times the IQR from the first and third quartiles. For instance, PC1-3 represents the distribution of RMSIP scores which are calculated over first three PCs amongst different mutants (corresponds to the RMSIP matrix in Fig. 6.)

13 Asparagine side-chain conformations

In all systems of AGP, there were five glycan chains within each glycosylated system attached at N33, N56, N72, N93, and N103 in protein via β-GlcNAc. For the sidechain conformations of asparagine before and after glycosylation, we calculated χ (chi) angles (Cγ-Cβ-Cα-N) at each glycosylation site which are presented for all unglycosylated and glycosylated systems (Figure S 27, Figure S 28, Figure S 29, Figure S 30, Figure S 31).

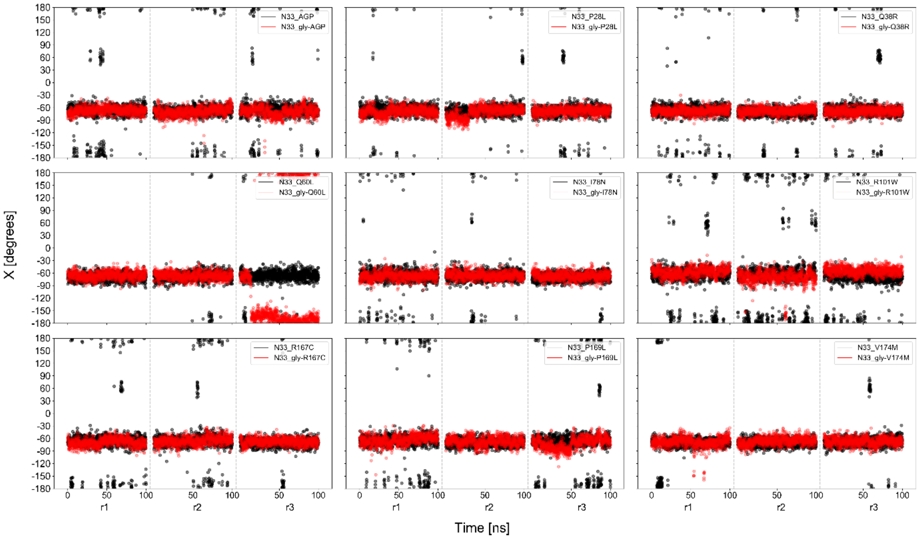


Figure S 27 χ(chi) angle of N33 as a function of time in all systems of AGP and its mutants: unglycosylated (black) and glycosylated (red) computed based on the protein's C_α atoms during the NPT production run of 100 ns from replica 1 to replica 3 (referred to as r1, r2, and r3).

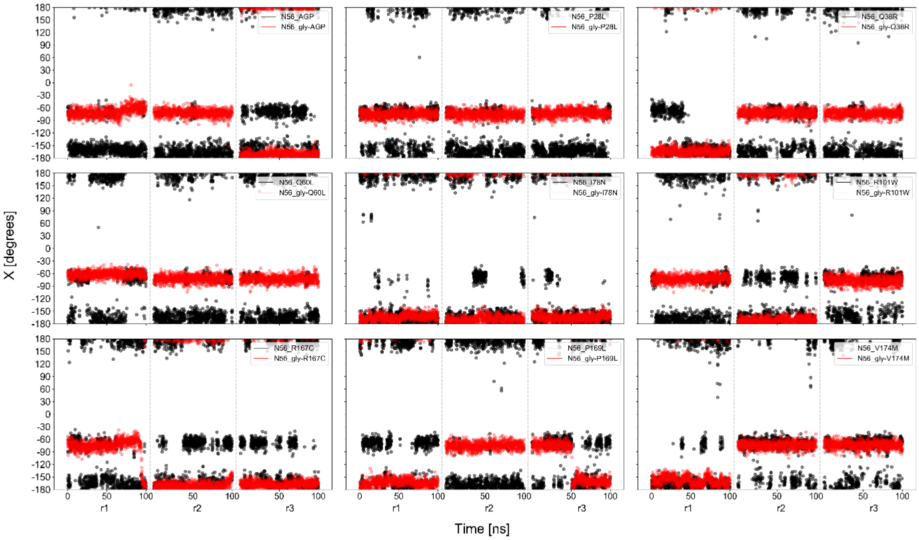


Figure S 28 χ(chi) angle of N56 as a function of time in all systems of AGP and its mutants: unglycosylated (black) and glycosylated (red) computed based on the protein's C_α atoms during the NPT production run of 100 ns from replica 1 to replica 3 (referred to as r1, r2, and r3).

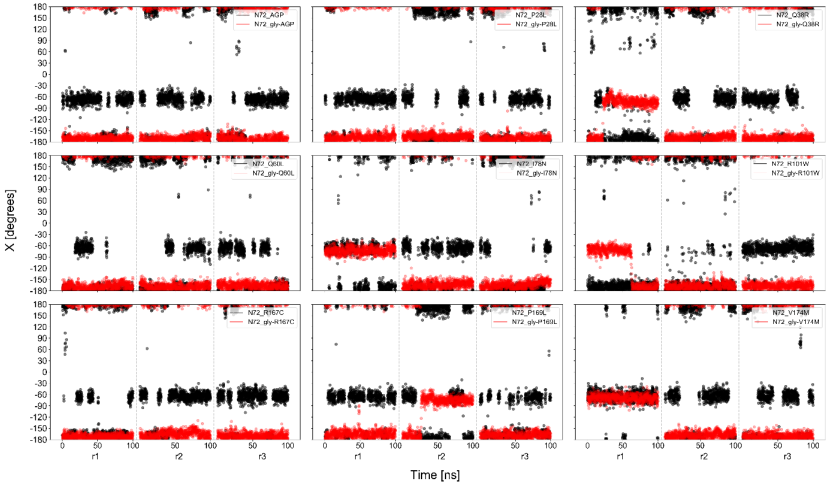


Figure S 29 χ(chi) angle of N72 as a function of time in all systems of AGP and its mutants: unglycosylated (black) and glycosylated (red) computed based on the protein's C_α atoms during the NPT production run of 100 ns from replica 1 to replica 3 (referred to as r1, r2, and r3).

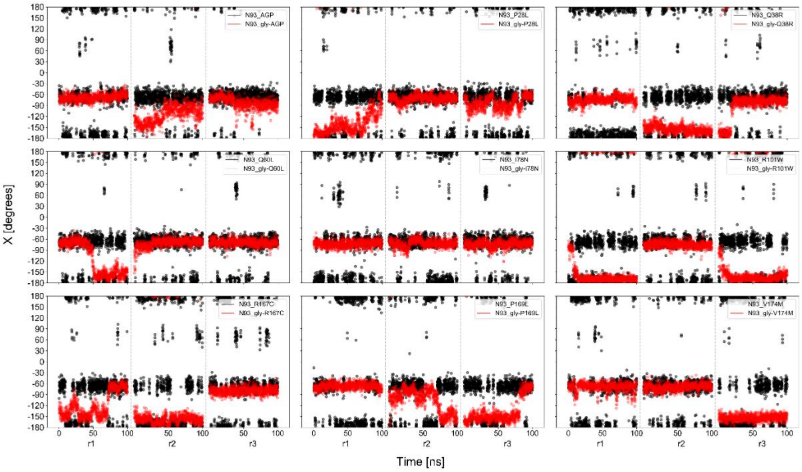


Figure S 30 χ(chi) angle of N93 as a function of time in all systems of AGP and its mutants: unglycosylated (black) and glycosylated (red) computed based on the protein's C_α atoms during the NPT production run of 100 ns from replica 1 to replica 3 (referred to as r1, r2, and r3).

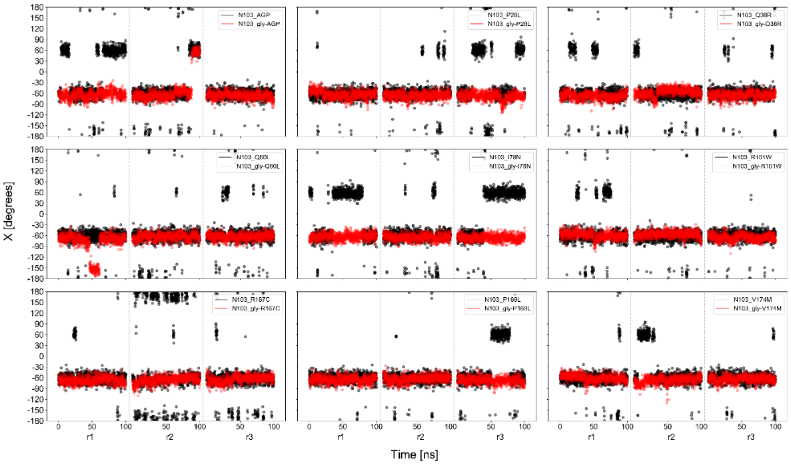


Figure S 31 χ(chi) angle of N103 as a function of time in all systems of AGP and its mutants: unglycosylated (black) and glycosylated (red) computed based on the protein’s C_α atoms during the NPT production run of 100 ns from replica 1 to replica 3 (referred to as r1, r2, and r3).

14 Glycan torsion angles

The following conventional definitions were used for glycosidic angles (Figure S 32):

(1) for 1→n (n = 2, 3, 4) linkages: ϕ = O5-C1-O’x-C'x and ψ = C1-O’x-C’x-C'x+1.

(2) for 2→3 linkages: ϕ = O6-C2-O’3-C'3 and ψ = C2-O’3-C’3-C'x+1.

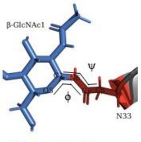
(3) for 1→6 and 2→6 glycosidic linkages: ϕ = O5-C1-O-C'6, ψ = C1-O-C’6-C’5, ω = O-C’6-C’5-O'5,

(4) for the N-glycosidic linkage: ϕ = O5-C1-Nδ-Cγ, ψ = C1-Nδ-Cγ-Cβ, ω = Nδ-Cγ-Cβ-Cα where Nδ, Cγ, Cβ, Cα belong to the corresponding sidechain atoms of N.

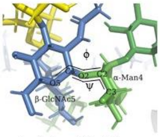
In this section, the following linkages are reported below for all glycosylated systems:

(a) β-GlcNAc(1→)N linkage

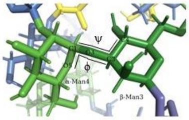
(b) Glycosidic linkages of all five chains



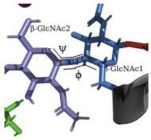
N-glycosidic linkage



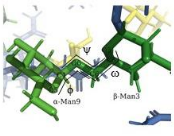
1→2 glycosidic linkage



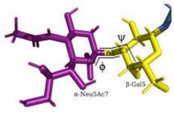
1→3 glycosidic linkage



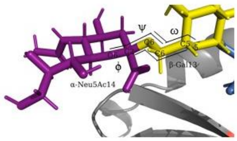
1→4 glycosidic linkage



1→6 glycosidic linkage



2→3 glycosidic linkage



2→6 glycosidic linkage

Figure S 32 Examples of ϕ, ψ, and ω torsion angles for all N-glycosidic linkages (protein-glycan) and all possible glycosidic linkages (glycan-glycan) in all glycosylated systems.

(a) Distribution of β-GlcNAc(1→)N linkages

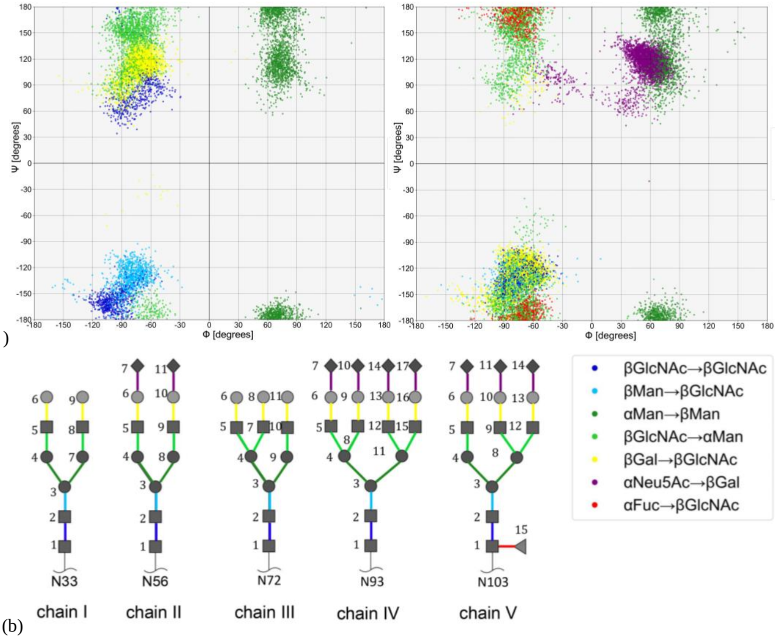
Table S 6 Circular standard deviation of ϕ, ψ and, ω of N-glycosidic linkages over 300 ns of all five chains at the glycosylation sites N33, N56, N72, N93 and N103 for all glycosylated systems.

glycan chain	system	mean csd ϕ [degrees]	mean csd ψ [degrees]	mean csd ω [degrees]
N33	gly-AGP	38.38	9.33	32.05
	gly-P28L	72.57	9.17	30.58
	gly-Q38R	9.67	8.12	11.62
	gly-Q60L	75.08	9.17	37.45
	gly-I78N	49.19	8.76	23.07
	gly-R101W	73.42	8.98	77.24
	gly-R167C	67.84	10.04	23.96
	gly-P169L	60.49	10.34	53.12
	gly-V174M	78.30	9.58	24.23
N56	gly-AGP	63.59	9.87	22.95
	gly-P28L	9.96	8.31	10.06
	gly-Q38R	12.93	10.74	27.42
	gly-Q60L	9.75	8.30	16.17
	gly-I78N	63.74	11.64	22.08
	gly-R101W	64.54	9.26	21.97
	gly-R167C	67.53	12.58	25.67
	gly-P169L	15.08	11.22	31.18
	gly-V174M	39.78	9.04	33.09
	gly-AGP	37.97	8.69	27.33
N72	gly-P28L	15.10	10.50	19.19
	gly-Q38R	49.95	9.68	40.55
	gly-Q60L	74.89	9.67	15.25
	gly-I78N	70.84	9.75	53.89
	gly-R101W	14.72	9.63	67.11
	gly-R167C	16.05	9.57	17.81
	gly-P169L	47.44	9.51	45.59
	gly-V174M	57.24	9.28	71.07
N93	gly-AGP	58.33	12.14	107.54
	gly-P28L	48.99	11.81	82.80
	gly-Q38R	74.63	12.41	73.65
	gly-Q60L	35.56	9.79	71.35
	gly-I78N	15.44	10.23	16.20
	gly-R101W	27.09	8.86	28.23
	gly-R167C	80.72	10.17	59.76
	gly-P169L	79.85	10.13	113.84
	gly-V174M	66.83	9.28	66.88
N103	gly-AGP	66.53	9.15	87.79
	gly-P28L	12.39	8.67	49.42
	gly-Q38R	60.91	8.53	83.74
	gly-Q60L	83.11	9.79	65.41
	gly-I78N	83.38	9.15	69.48
	gly-R101W	67.15	8.41	72.85
	gly-R167C	86.89	9.79	72.14
	gly-P169L	70.34	10.14	51.40

	gly-V174M	80.29	9.57	84.09

(b) Distribution of glycosidic linkages

The five chains were distinct in terms of branching such as bi-antennary, tri-antennary, and tetra-antennary chains. They were composed of β-N-Acetyl-D-glucosamine (β-GlcNAc), β-D-Galactose (β-Gal), α- and β-D-Mannose (α-Man, β-Man), terminal α-N-Acetylneuraminic acid (α-Neu5Ac), and α-L-Fucose (α-Fuc). Out of the five glycan chains, the terminal α-Neu5Ac was present in chain II, IV, and V, while α-Fuc was present only on chain V. The five glycan chains in glycosylated mutants were identical to the glycan chains in gly-AGP. As an example, carb-Rama plot of chain I and chain V in gly-AGP is shown for ϕ and ψ distribution of distinct glycosidic linkages (Figure S 33(a)). The glycans in the Figure S 33 (b) are numbered to show mean circular standard deviation of ϕ, ψ, and ω (in specific cases) angles of all glycosylated systems for each linkage in a glycan chain (Table S 7).



(b)

Figure S 33 (a) Carb-Rama plot of glycan chain I (left) and glycan chain V (right) of gly-AGP showing ϕ/ψ distributions for all glycosidic linkages between two glycans. The colours represent type of glycosidic linkage between two glycans in the chain. (b) Chain diagram of glycan chains from I to V, showing coloured glycosidic linkages, the glycans are numbered.

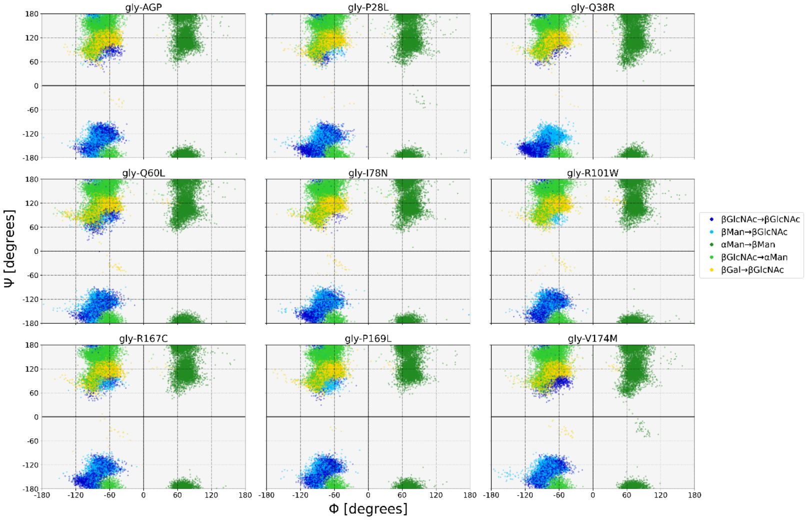


Figure S 34 ϕ and ψ distributions of glycan chain I of gly-AGP and its glycosylated mutants over 300 ns. The colours represent the different type of glycosidic linkages in chain I.

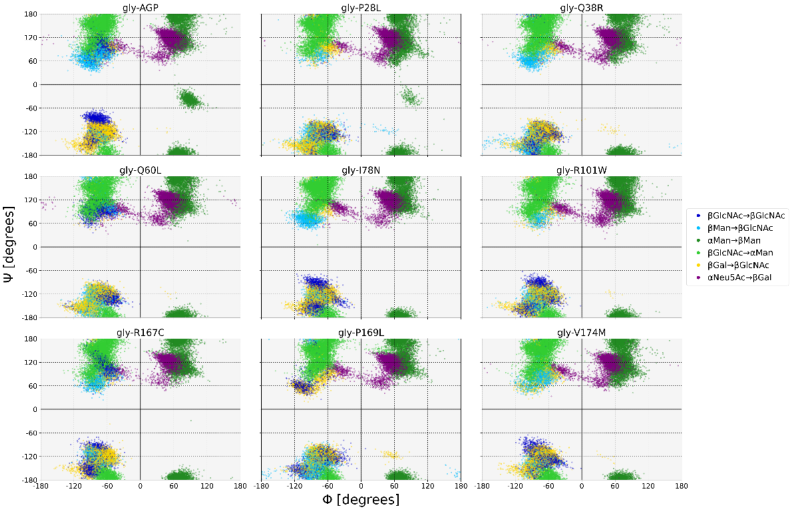


Figure S 35 ϕ and ψ distributions of glycan chain II of gly-AGP and its glycosylated mutants over 300 ns. The colours represent the different type of glycosidic linkages in chain II.

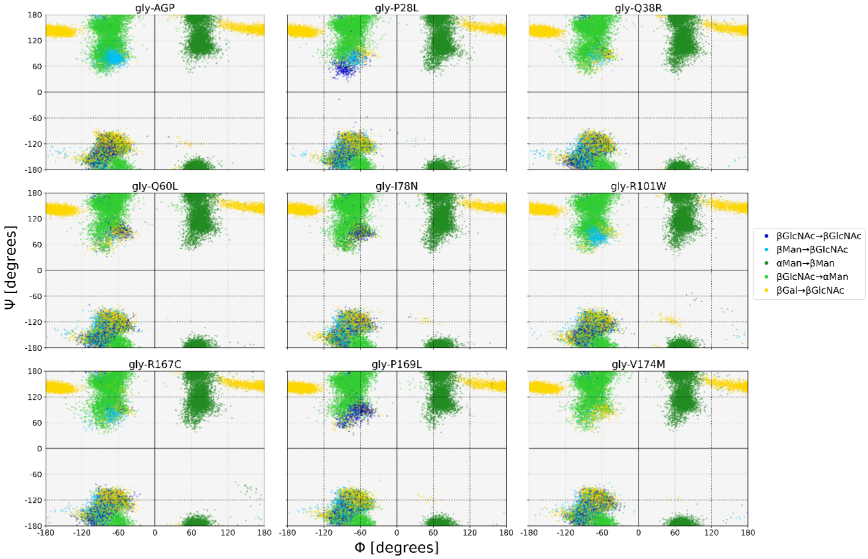


Figure S 36 ϕ and ψ distributions of glycan chain III of gly-AGP and its glycosylated mutants over 300 ns. The colours represent the different type of glycosidic linkages in chain III.

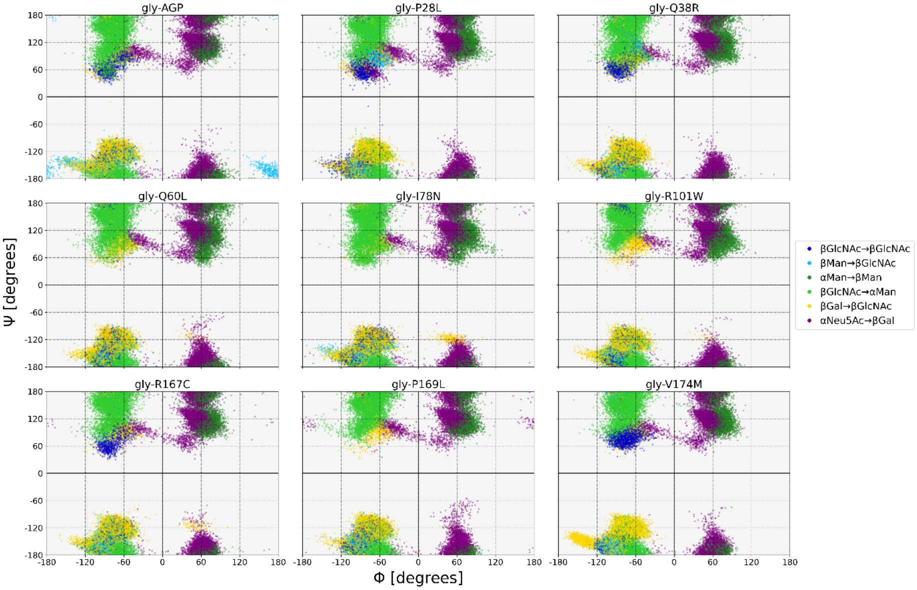


Figure S 37 ϕ and ψ distributions of glycan chain IV of gly-AGP and its glycosylated mutants over 300 ns. The colours represent the different type of glycosidic linkages in chain IV.

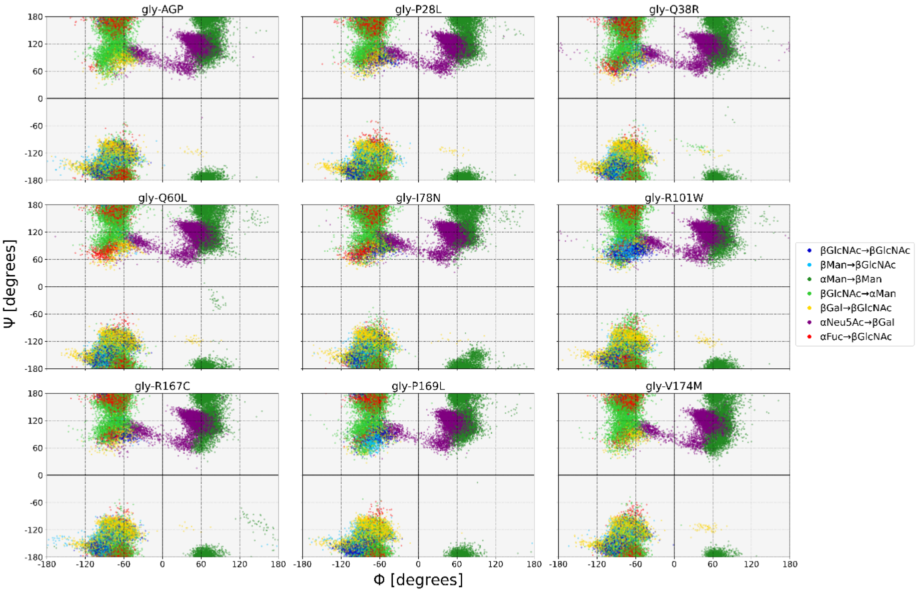


Figure S 38 ϕ and ψ distributions of glycan chain V of gly-AGP and its glycosylated mutants over 300 ns. The colours represent the different type of glycosidic linkages in chain V.

Table S 7 Mean of circular standard deviation (csd) of ϕ, ψ and, ω angles of all glycosylated systems for each linkage in a glycan chain over 300 ns.

glycan chain	glycan linkages	mean csd ϕ [degrees]	mean csd ψ [degrees]	mean csd ω [degrees]
chain I	βGlcNAc1→N33	58.33	9.27	34.81
	βGlcNAc2→βGlcNAc1	15.80	36.77	
	βMan3→βGlcNAc2	14.40	33.15	
	αMan4→βMan3	11.36	22.71	
	βGlcNAc5→αMan4	16.70	31.72	
	βGal6→βGlcNAc5	14.32	16.35	
	αMan7→βMan3	10.41	34.11	66.74
	βGlcNAc8→αMan7	15.93	30.96	
	βGal9→βGlcNAc8	13.57	15.86	
chain II	βGlcNAc1→N56	38.55	10.11	23.40
	βGlcNAc2→βGlcNAc1	14.87	42.53	
	βMan3→βGlcNAc2	15.36	64.20	
	αMan4→βMan3	11.80	30.99	
	βGlcNAc5→αMan4	19.03	32.45	
	βGal6→βGlcNAc5	15.55	36.99	
	αNeu5Ac7→βGal6	22.38	15.75	
	αMan8→βMan3	10.42	28.83	56.26
	βGlcNAc9→αMan8	15.57	30.50	
	βGal10→βGlcNAc9	16.02	35.76	
	αNeu5Ac11→βGal10	22.99	15.76	
chain III	βGlcNAc1→N72	42.69	9.59	39.76
	βGlcNAc2→βGlcNAc1	14.18	31.98	
	βMan3→βGlcNAc2	14.40	43.78	
	αMan4→βMan3	11.14	21.69	
	βGlcNAc5→αMan4	17.44	28.81	
	βGal6→βGlcNAc5	15.77	30.69	

	βGlcNAc7→αMan4	15.02	44.11	
	βGal8→βGlcNAc7	15.15	33.82	
	αMan9→βMan3	10.65	32.50	49.83
	βGlcNAc10→αMan9	14.62	31.77	
	βGal11→βGlcNAc10	28.64	7.19	
chain IV	βGlcNAc1→N93	54.16	10.54	68.92
	βGlcNAc2→βGlcNAc1	14.15	53.62	
	βMan3→βGlcNAc2	16.35	23.29	
	αMan4→βMan3	10.54	14.87	
	βGlcNAc5→αMan4	17.45	30.33	
	βGal6→βGlcNAc5	15.51	34.43	
	αNeu5Ac7→βGal6	21.89	16.20	
	βGlcNAc8→αMan4	16.58	17.42	15.35
	βGal9→βGlcNAc8	14.85	30.70	
	αNeu5Ac10→βGal9	13.50	20.77	85.38
	αMan11→βMan3	10.46	30.35	41.53
	βGlcNAc12→αMan11	15.79	28.55	
	βGal13→βGlcNAc12	17.51	27.25	
	αNeu5Ac14→βGal13	12.86	19.04	46.30
	βGlcNAc15→αMan11	13.57	43.16	
	βGal16→βGlcNAc15	19.70	28.64	
	αNeu5Ac17→βGal16	27.31	16.61	
chain V	βGlcNAc1→N103	67.89	9.24	70.70
	βGlcNAc2→βGlcNAc1	15.72	37.07	
	βMan3→βGlcNAc2	13.90	33.48	
	αMan4→βMan3	11.75	22.84	
	βGlcNAc5→αMan4	14.93	34.47	
	βGal6→βGlcNAc5	16.08	31.85	
	αNeu5Ac7→βGal6	25.18	16.30	
	αMan8→βMan3	10.96	37.94	41.61
	βGlcNAc9→αMan8	16.55	30.28	
	βGal10→βGlcNAc9	15.93	30.67	
	αNeu5Ac11→βGal10	22.44	16.03	
	βGlcNAc12→αMan8	12.48	30.88	29.70
	βGal13→βGlcNAc12	15.78	28.47	
	αNeu5Ac14→βGal13	23.48	16.22	
	αFuc15→βGlcNAc1	11.46	32.71	52.62

15 H-bonds between protein and glycans

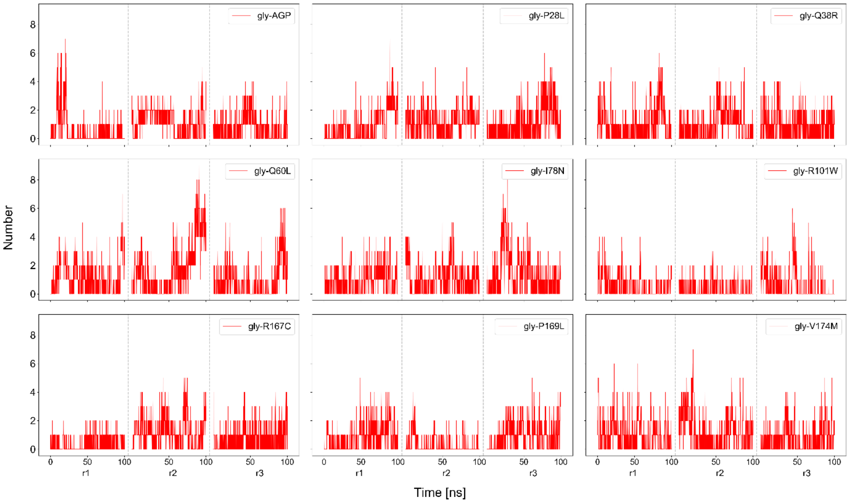


Figure S 39 H-bond between glycan chain I and protein of gly-AGP and glycosylated mutants.

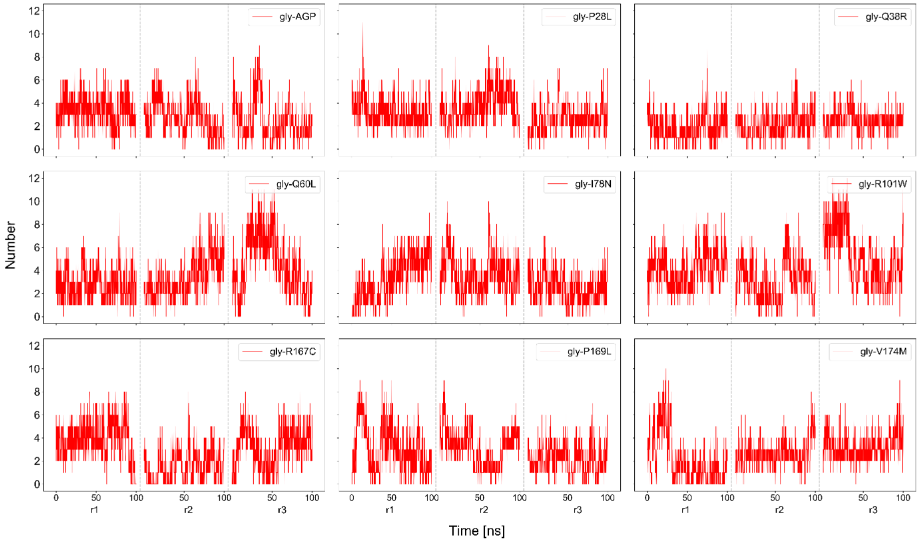


Figure S 40 H-bond between glycan chain II and protein of gly-AGP and glycosylated mutants.

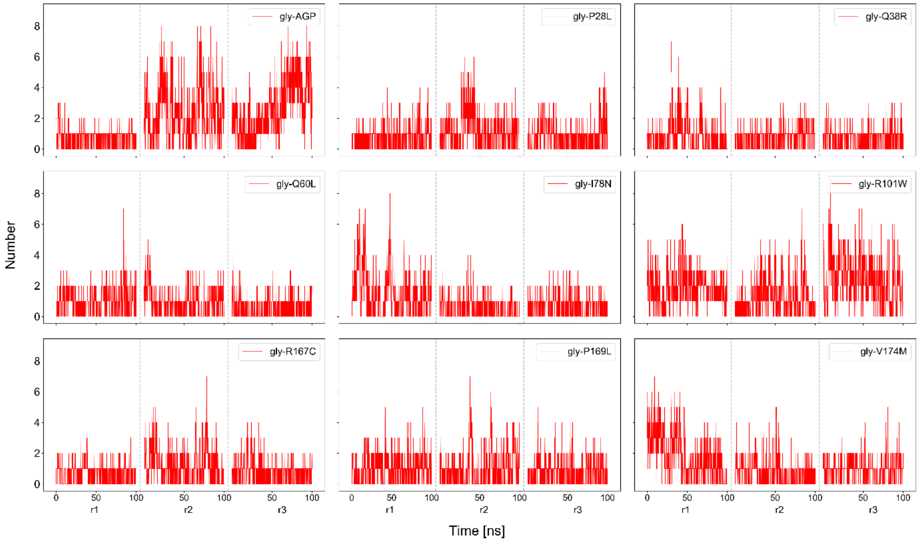


Figure S 41 H-bond between glycan chain III and protein of gly-AGP and glycosylated mutants.

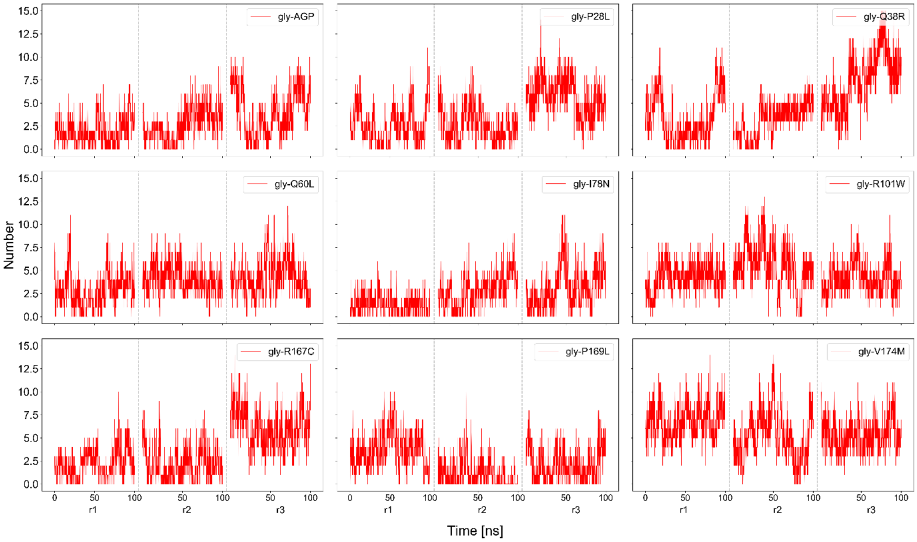


Figure S 42 H-bond between glycan chain IV and protein of gly-AGP and glycosylated mutants.

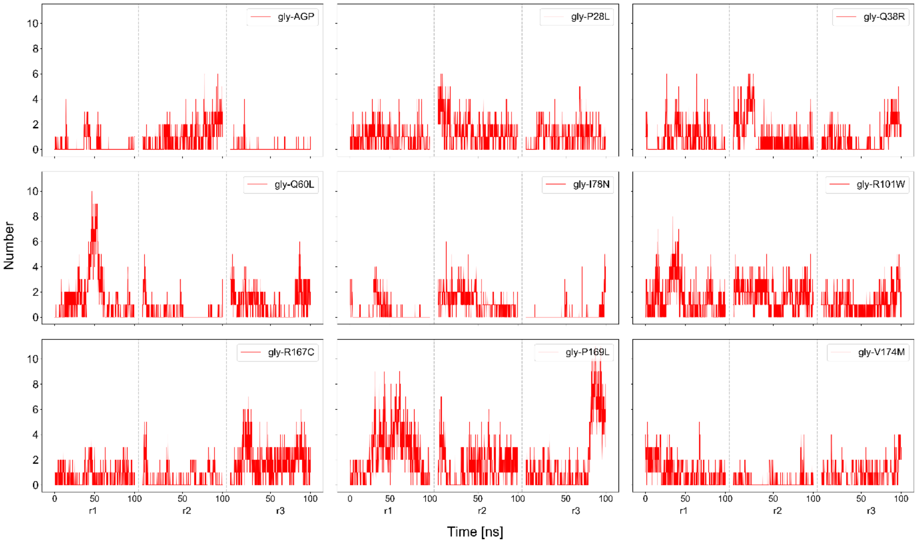


Figure S 43 H-bond between glycan chain V and protein of gly-AGP and glycosylated mutants.

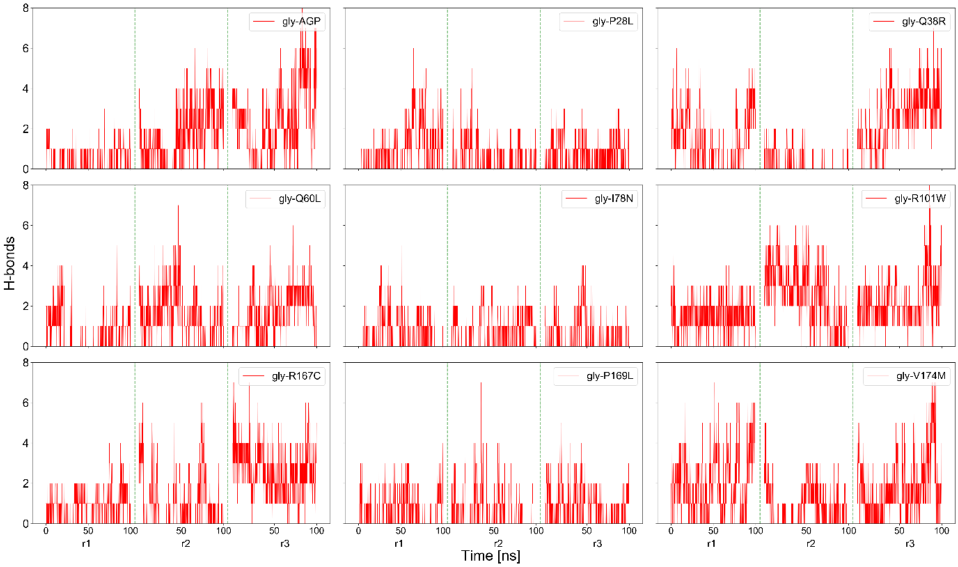


Figure S 44 H-bond between glycan chain IV and LBE of gly-AGP and glycosylated mutants.

Supplementary information for "Gradations in protein dynamics captured by experimental NMR are not well represented by AlphaFold2 models and other computational metrics"

Jose Gavalda-Garcia [1,2,†], Bhawna Dixit [1,2,3,†], Adrián Díaz [1,2], An Ghysels [3], and Wim Vranken [1,2,4,5,6,*]

[1]Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Brussels, Belgium
[2]Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium
[3]IBiTech - BioMMedA group, Ghent University, Belgium
[4]AI lab, Vrije Universiteit Brussel, Brussels, Belgium
[5]Chemistry department, Vrije Universiteit Brussel, Brussels, Belgium
[6]Biomedical sciences, Vrije Universiteit Brussel, Brussels, Belgium

[†]Authors contributed equally to this work.
[*]To whom correspondence should be addressed: wim.vranken@vub.be

February 10, 2025

Contents

1 Supplementary figures and tables in support of the main text 217

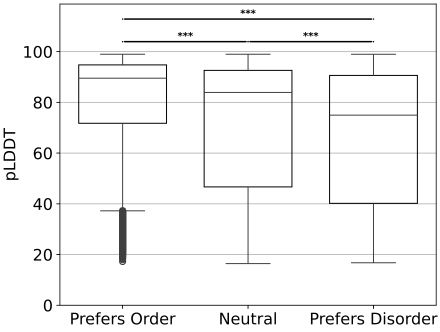
2 Additional information related to NMA 228
2.1 Generation of NMA data . . . 228
2.2 Truncation criterion . . . 230
2.3 Additional analysis of RMSF and correlation with pLDDT or S^2_{RCI} . . . 231
2.4 Examples of S^2_{RCI} and RMSF correlation for Alphafold2 and NMR models . . . 237

3 Conflicting secondary structure elements between AlphaFold2 and NMR models 242

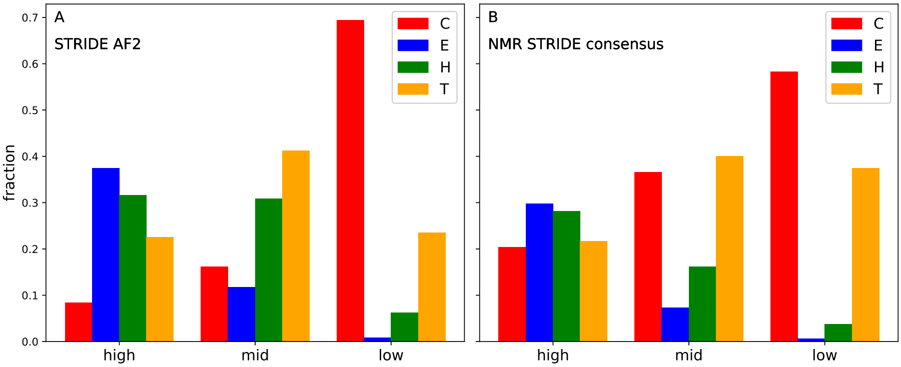
4 Correlation assessment between MD and NMA-derived RMSF in the Constava dataset 244

5 Secondary structure content variation 251

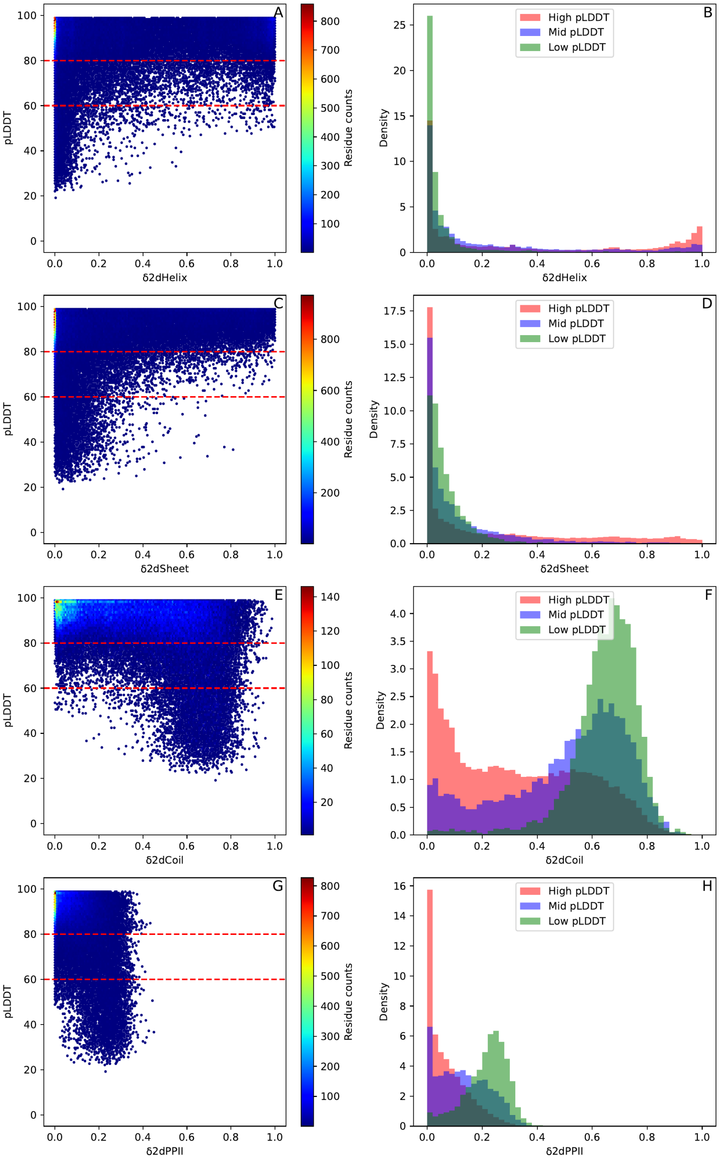
1 Supplementary figures and tables in support of the main text



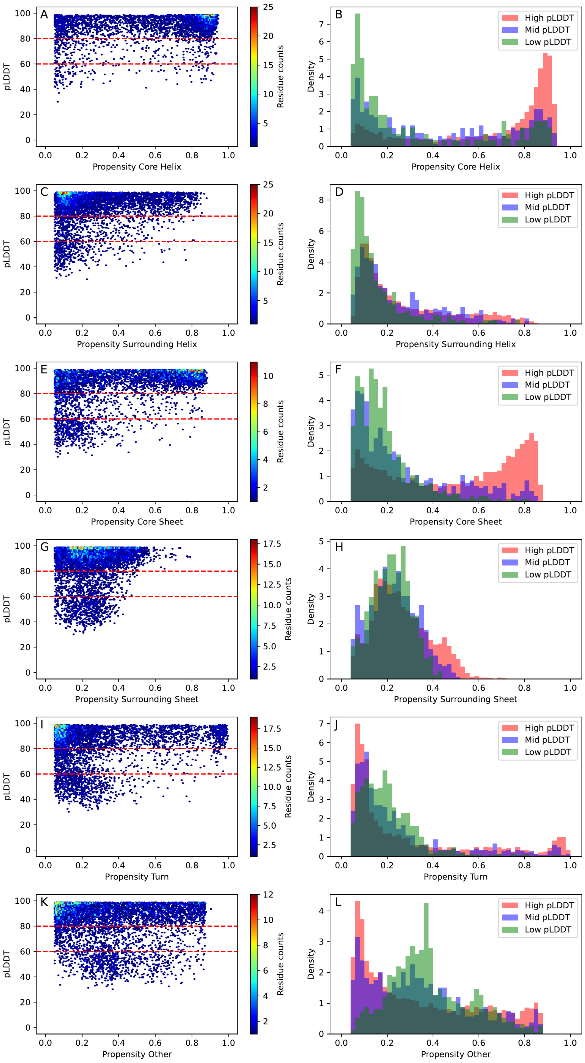
Supplementary Fig. 1: **pLDDT values per amino acid order classes.** The amino acids from the S^2_{RCI} dataset were clustered in 3 classes according to their order preference: preferentially ordered (cysteine, phenylalanine, isoleucine, leucine, valine, tryptophan & tyrosine; N=106,363), neutral (alanine, glutamic acid, lysine, methionine, glutamine, arginine, threonine & histidine; N=155,050) and preferentially disordered (aspartic acid, glycine, asparagine, proline & serine; N=112,945). The distributions were tested with a Mann–Whitney U test, which resulted in p-values < 0.001 in all tests.



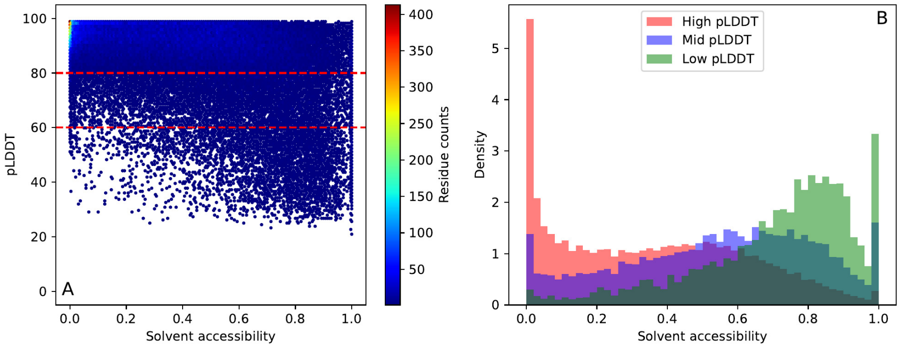
Supplementary Fig. 2: **Abundance of secondary structures by pLDDT ranges.** A) The STRIDE secondary structure assignment from the structure that AlphaFold2 produces. B) The STRIDE consensus secondary structure assignment here provided is the most abundantly assigned secondary structure in the ensemble of NMR structures available. The STRIDE consensus assignment (panel B) produces a decrease in abundance for helix (H) and sheet (E) fractions as the pLDDT decreases, as well as an increase in coil (C) and turn (T) conformations. These tendencies are maintained for AlphaFold2's single structure assignment (panel A) in sheet and coil fractions, but not for helix and turn.



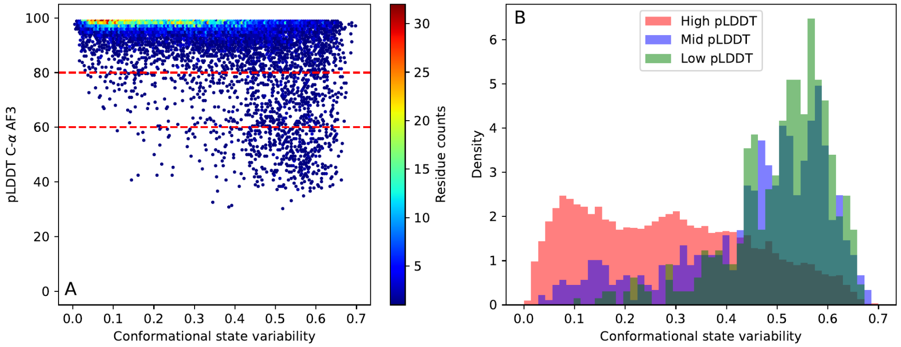
Supplementary Fig. 3: **Comparison of pLDDT and $\delta 2D$ populations.** For all available residues in the S^2_{RCI} dataset, the populations of conformational states were calculated with δ2D method. High populations values of a conformation indicate high presence of such conformation in the ensemble, and vice-versa. A-D: For both Helix and Sheet conformations, higher δ2D populations are obtained for residues featuring high pLDDT ($\geq$ 80, N=62,014), gradually adopting lower populations for mid (80 > pLDDT $\geq$ 60, N=8,539) and low (< 60, N=4,773) pLDDT values. E-H: Coil and PPII populations are prominently higher for higher for low pLDDT ranges than for mid and high ranges. Mann-Whitney two-sided U tests p-values between each pLDDT-stratified distribution for each $\delta 2D$ conformation confirmed difference at p-values < 0.001 (table 2).



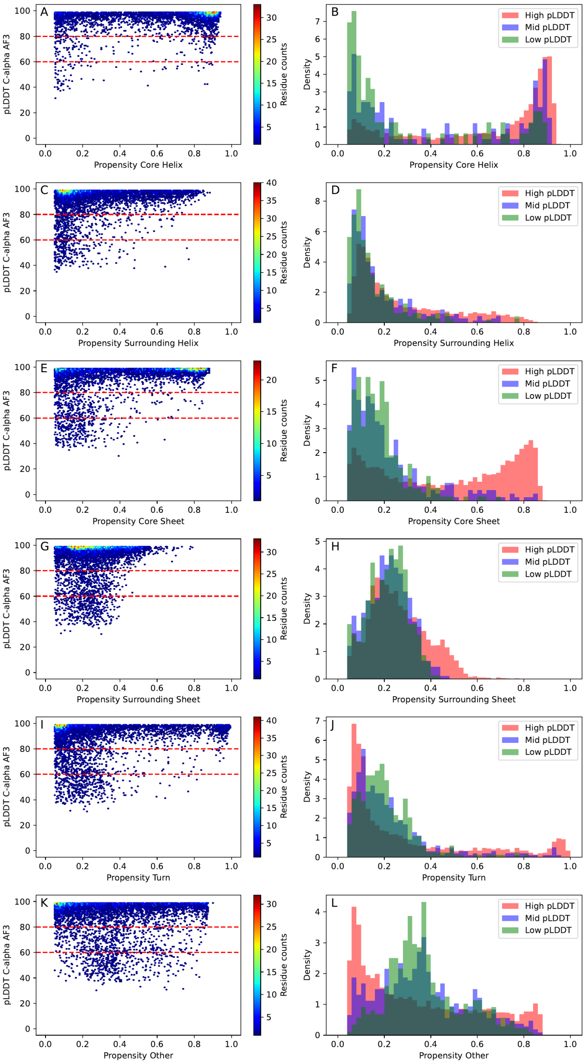
Supplementary Fig. 4: **Comparison between AlphaFold2 pLDDT and Constava's conformational state propensities.** The pLDDT of each residue was plotted against the propensity for each of the 6 conformational states calculated in Constava. Any propensity below 0.05 was deemed non-informative and discarded to facilitate interpretation. All residues in the MD dataset were stratified in high (N=9,523), mid (N=1,038) and low (N=809) AlphaFold2 ranges. A & B) Hexagonal binning of AlphaFold3 C-α pLDDT vs Core Helix propensity and its associated pLDDT-stratified distributions. C & D) Hexagonal binning of AlphaFold3 C-α pLDDT vs Surrounding Helix propensity and its associated pLDDT-stratified distributions. E & F) Hexagonal binning of AlphaFold3 C-α pLDDT vs Core Sheet propensity and its associated pLDDT-stratified distributions. G & H) Hexagonal binning of AlphaFold3 C-α pLDDT vs Surrounding Sheet propensity and its associated pLDDT-stratified distributions. I & J) Hexagonal binning of AlphaFold3 C-α pLDDT vs Turn propensity and its associated pLDDT-stratified distributions. K & L) Hexagonal binning of AlphaFold3 C-α pLDDT vs Other propensity and its associated pLDDT-stratified distributions. The p-values for the Mann-Whitney two-sided U tests between each pLDDT-stratified distribution for every conformational state can be found in supplementary table 1.



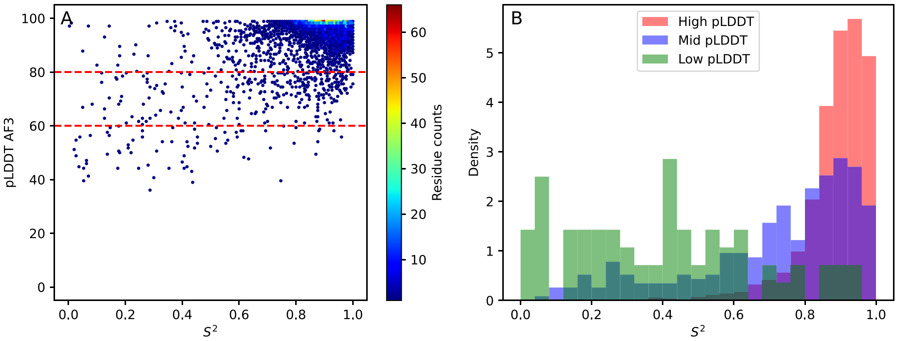
Supplementary Fig. 5: **Comparison of pLDDT and solvent accessibility.** A) The AlphaFold2 pLDDT in the S^2_{RCI} dataset were plotted against the solvent accessibility score for every residue with available values, in an hexagonal binning plot. B) Distributions of solvent accessibility scores, stratified in high ($\geq$ 80, N=62,319), mid (80 $>$ pLDDT $\geq$ 60, N=8,708) and low ($<$ 60, N=4,842) AlphaFold2 pLDDT values. Mann-Whitney two-sided U test confirmed significant differences between all distribution pairs with a p-value $<$ 0.001 (table 3).



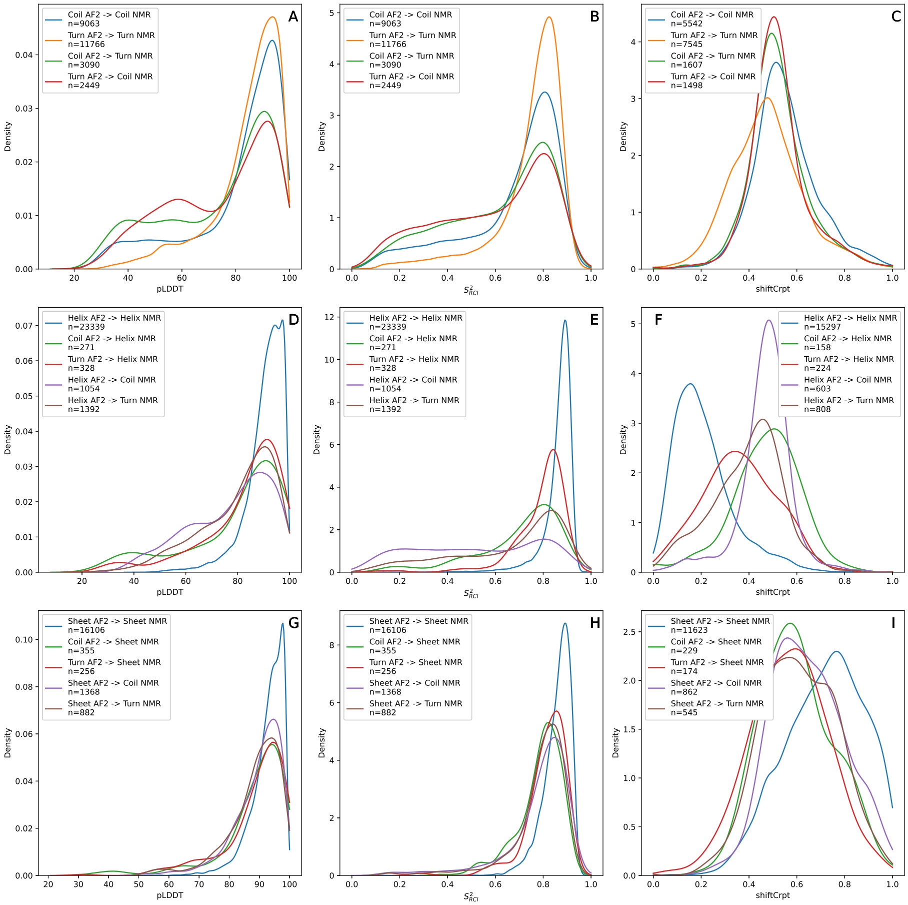
Supplementary Fig. 6: **AlphaFold3 C-α pLDDT vs. conformational state variability.** A) High pLDDT values ($\geq$ 80, N = 10,272) concentrate in areas with lower conformational state variability. Low pLDDT $<$ 60, N = 463) usually correspond to residues with high conformational state variability. B) This tendency is more clearly observed with pLDDT-stratified distributions, which shows that low pLDDT residues correspond to residues with high conformational state variability, therefore with high potential to exist in multiple conformations, and vice-versa for high pLDDT and low variability residues. Mid pLDDT residues 80 $>$ pLDDT $\geq$ 60, N = 634) exhibit an intermediate distribution. Mann-Whitney two-sided U test yielded a p-value $<$ 0.001 between all distributions (table 3). The associated distributions per propensity can be found in supplementary fig. 7.



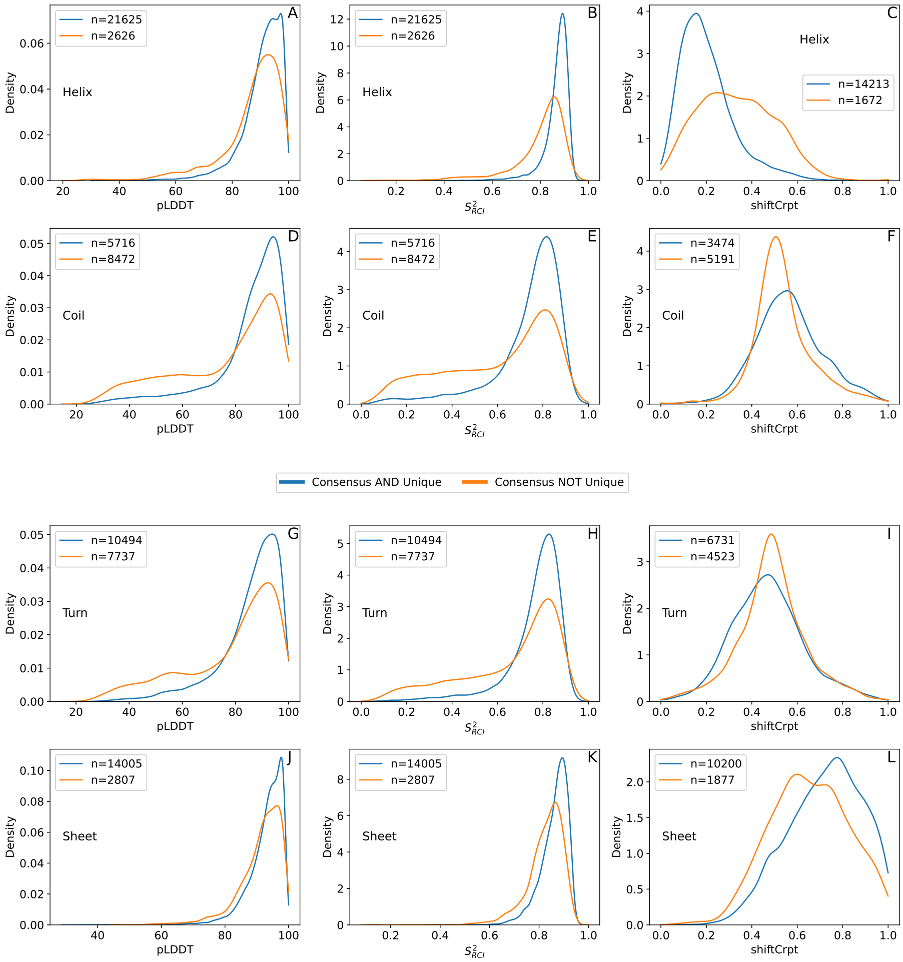
Supplementary Fig. 7: **Comparison between AlphaFold3 C-α pLDDT and Constava's conformational state propensities.** The pLDDT of each residue was plotted against the propensity for each of the 6 conformational states calculated in Constava. Any propensity below 0.05 was deemed non-informative and discarded to facilitate interpretation. All residues in the MD dataset were stratified in high (N=10,272), mid (N=634) and low (N=463) AlphaFold3 C-α ranges. A & B) Hexagonal binning of AlphaFold3 C-α pLDDT vs Core Helix propensity and its associated pLDDT-stratified distributions. C & D) Hexagonal binning of AlphaFold3 C-α pLDDT vs Surrounding Helix propensity and its associated pLDDT-stratified distributions. E & F) Hexagonal binning of AlphaFold3 C-α pLDDT vs Core Sheet propensity and its associated pLDDT-stratified distributions. G & H) Hexagonal binning of AlphaFold3 C-α pLDDT vs Surrounding Sheet propensity and its associated pLDDT-stratified distributions. I & J) Hexagonal binning of AlphaFold3 C-α pLDDT vs Turn propensity and its associated pLDDT-stratified distributions. K & L) Hexagonal binning of AlphaFold3 C-α pLDDT vs Other propensity and its associated pLDDT-stratified distributions. The p-values for the Mann-Whitney two-sided U tests between each pLDDT-stratified distribution for every conformational state can be found in supplementary table 1



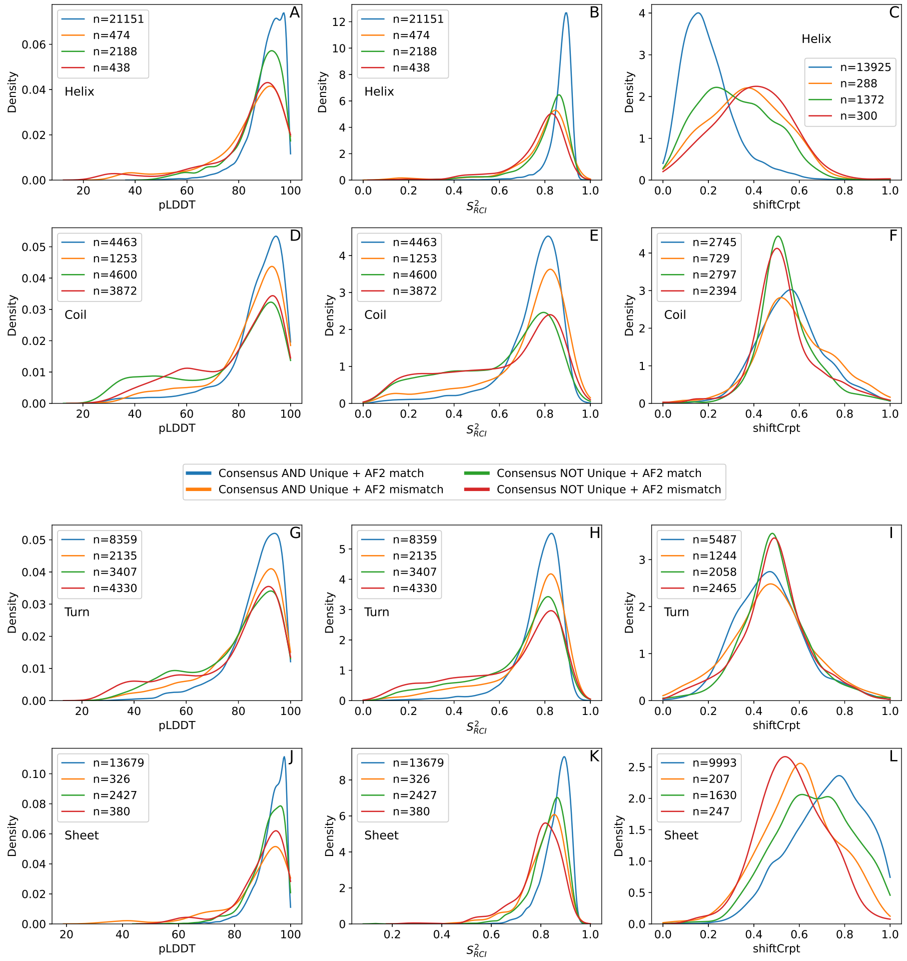
Supplementary Fig. 8: **Comparison between AlphaFold3 C-α pLDDT and S^2.** Most residues in this dataset exhibit high pLDDT and high S^2 values, represented by warmer hues in panel A. Due to the limited and uneven residues sample sizes (high pLDDT, ≥ 80, N = 4,125; mid pLDDT, $80 > \text{pLDDT} \geq 60$, N = 287; low pLDDT, < 60, N = 70), it is challenging to make definitive conclusions about the sparsely populated mid and low pLDDT regions. Panel B illustrates the distributions of S^2 values of each pLDDT range, for which Mann-Whitney two-sided U test yielded a p-value < 0.001 between all distributions (table 3).



Supplementary Fig. 9: **Distribution of STRIDE fold assignment matches and mismatches between AlphaFold2 structures and consensus NMR ensemble assignments, for diverse metrics.** Those residues in the S^2_{RCI} dataset with STRIDE consensus were stratified according to their secondary assignment pairs, derived from AlphaFold2 structures and the NMR models consensus assignment. A, D & G) pLDDT distributions for residues whose consensus STRIDE assignment from NMR ensembles and from AlphaFold2 models match or mismatch. B, E & H) S^2_{RCI} distributions for residues whose consensus STRIDE assignment from NMR ensembles and from AlphaFold2 models match or mismatch. C, F & I) shiftCrypt distributions for residues whose consensus STRIDE assignment from NMR ensembles and from AlphaFold2 models match or mismatch.



Supplementary Fig. 10: **Distribution of residues with NMR consensus STRIDE assignment with and without unique STRIDE assignment, for diverse metrics.** Those residues in the S^2_{RCI} dataset with STRIDE consensus were stratified according to whether they featured a unique STRIDE assignment across all the models in their corresponding NMR ensemble. A, D, G & J) pLDDT distributions for α-helix, coil, turn and β-sheet respectively. B, E, H & K) S^2_{RCI} distributions for α-helix, coil, turn and β-sheet respectively. C, F, I & L) shiftCrypt distributions for α-helix, coil, turn and β-sheet respectively.



Supplementary Fig. 11: **Distribution of residues with NMR consensus STRIDE assignment with and without unique STRIDE assignment, with matching and mismatching NMR and AlphaFold2 fold assignments, for diverse metrics.** Those residues in the S^2_{RCI} dataset with STRIDE consensus were stratified according to whether they featured a unique STRIDE assignment across all the models in their corresponding NMR ensemble. Then, they were further stratified on whether or not their NMR consensus STRIDE assignment matched the AlphaFold2 STRIDE assignment. A, D, G & J) pLDDT distributions for α-helix, coil, turn and β-sheet respectively. B, E, H & K) S^2_{RCI} distributions for α-helix, coil, turn and β-sheet respectively. C, F, I & L) shiftCrypt distributions for α-helix, coil, turn and β-sheet respectively.

Supplementary Table 1: **Results of the Mann-Whitney two-sided U tests between pLDDT-stratified subsets per Constava's conformational states on the MD dataset.** For each of Constava's conformational states, the pLDDT-stratified subsets were tested against each other with a Mann-Whitney two-sided U test to assess differential distributions.

Conformational state	AlphaFold version	pLDDT ranges	Difference between means	p-value
Core Helix	2	High-Mid	0.2161	6.67×10^{-34}
Core Helix	2	High-Low	0.3427	2.52×10^{-31}
Core Helix	2	Mid-Low	0.1266	1.53×10^{-5}
Surr. Helix	2	High-Mid	0.0264	0.01955
Surr. Helix	2	High-Low	0.1168	5.17×10^{-43}
Surr. Helix	2	Mid-Low	0.0903	1.24×10^{-20}
Core Sheet	2	High-Mid	0.2254	1.60×10^{-60}
Core Sheet	2	High-Low	0.3074	2.12×10^{-104}
Core Sheet	2	Mid-Low	0.0821	1.58×10^{-5}
Surr. Sheet	2	High-Mid	0.0339	1.96×10^{-8}
Surr. Sheet	2	High-Low	0.0455	3.02×10^{-15}
Surr. Sheet	2	Mid-Low	0.0117	0.08711
Turn	2	High-Mid	0.0394	0.44796
Turn	2	High-Low	0.0826	0.22141
Turn	2	Mid-Low	0.0432	0.45743
Other	2	High-Mid	0.0082	0.28323
Other	2	High-Low	-0.0397	6.19×10^{-15}
Other	2	Mid-Low	-0.0479	1.85×10^{-9}
Core Helix	3	High-Mid	0.1990	4.07×10^{-14}
Core Helix	3	High-Low	0.3498	4.63×10^{-20}
Core Helix	3	Mid-Low	0.1507	1.85×10^{-4}
Surr. Helix	3	High-Mid	0.0861	6.18×10^{-18}
Surr. Helix	3	High-Low	0.1203	7.55×10^{-27}
Surr. Helix	3	Mid-Low	0.0343	0.00256
Core Sheet	3	High-Mid	0.2657	1.94×10^{-58}
Core Sheet	3	High-Low	0.3019	6.68×10^{-64}
Core Sheet	3	Mid-Low	0.0362	0.29869
Surr. Sheet	3	High-Mid	0.0423	6.90×10^{-10}
Surr. Sheet	3	High-Low	0.0462	1.51×10^{-10}
Surr. Sheet	3	Mid-Low	0.0039	0.66095
Turn	3	High-Mid	0.0628	0.48281
Turn	3	High-Low	0.0779	0.13434
Turn	3	Mid-Low	0.0152	0.33197
Other	3	High-Mid	-0.0178	0.000078
Other	3	High-Low	-0.0288	2.72×10^{-8}
Other	3	Mid-Low	-0.0110	0.15908

Supplementary Table 2: **Results of the Mann-Whitney two-sided U tests between pLDDT-stratified subsets for S^2_{RCI} $\delta 2D$ conformations in AlphaFold2.** For each conformation type, the pLDDT-stratified subsets were tested against each other with a Mann-Whitney two-sided U test to assess differential distributions. *Note: p-values marked with * were too low for Scipy to differentiate from 0.*

Conformation	pLDDT ranges	Difference between means	p-value
$\delta 2D$ Helix	High-Mid	0.1185	3.38×10^{-55}
$\delta 2D$ Helix	High-Low	0.2896	0.0*
$\delta 2D$ Helix	Mid-Low	0.1711	1.23×10^{-300}
$\delta 2D$ Sheet	High-Mid	0.1181	5.03×10^{-49}
$\delta 2D$ Sheet	High-Low	0.1723	5.44×10^{-53}
$\delta 2D$ Sheet	Mid-Low	0.0542	2.63×10^{-7}
$\delta 2D$ Coil	High-Mid	-0.1775	0.0*
$\delta 2D$ Coil	High-Low	-0.3184	0.0*
$\delta 2D$ Coil	Mid-Low	-0.1410	3.76×10^{-305}
$\delta 2D$ PPII	High-Mid	-0.0592	0.0*
$\delta 2D$ PPII	High-Low	-0.1435	0.0*
$\delta 2D$ PPII	Mid-Low	-0.0843	0.0*

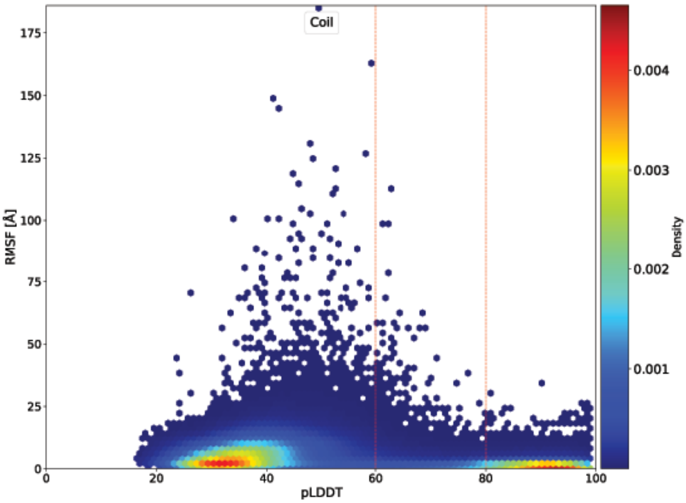
Supplementary Table 3: **Results of the Mann-Whitney two-sided U tests between pLDDT-stratified subsets.** For each dataset, the pLDDT-stratified subsets were tested against each other with a Mann-Whitney two-sided U test to assess differential distributions. *Note: p-values marked with * were too low for Scipy to differentiate from 0.*

Dataset	Metric	AlphaFold version	pLDDT ranges	Difference between means	p-value
S^2_{RCI}	S^2_{RCI}	2	high-mid	0.1417	0*
S^2_{RCI}	S^2_{RCI}	2	high-low	0.4150	0*
S^2_{RCI}	S^2_{RCI}	2	mid-low	0.2732	0*
S^2_{RCI}	ShiftCrypt	2	high-mid	0.0152	0.002
S^2_{RCI}	ShiftCrypt	2	high-low	-0.0201	4.55×10^{-7}
S^2_{RCI}	ShiftCrypt	2	mid-low	-0.0353	9.45×10^{-21}
S^2_{RCI}	Solvent accessibility	2	high-mid	-0.1851	0*
S^2_{RCI}	Solvent accessibility	2	high-low	-0.3521	0*
S^2_{RCI}	Solvent accessibility	2	mid-low	-0.1670	1.93×10^{-297}
S^2	S^2	2	high-mid	0.1352	5.60×10^{-24}
S^2	S^2	2	high-low	0.4238	1.19×10^{-53}
S^2	S^2	2	mid-low	0.2886	2.12×10^{-19}
S^2	S^2	3	high-mid	0.1681	1.74×10^{-43}
S^2	S^2	3	high-low	0.4904	6.36×10^{-38}
S^2	S^2	3	mid-low	0.3223	2.45×10^{-16}
MD	Conf. state var.	2	high-mid	-0.1468	3.46×10^{-143}
MD	Conf. state var.	2	high-low	-0.2197	1.23×10^{-246}
MD	Conf. state var.	2	mid-low	-0.0729	1.48×10^{-22}
MD	Conf. state var.	3	high-mid	-0.1743	1.46×10^{-126}
MD	Conf. state var.	3	high-low	-0.2258	4.6×10^{-155}
MD	Conf. state var.	3	mid-low	-0.0515	2.08×10^{-7}

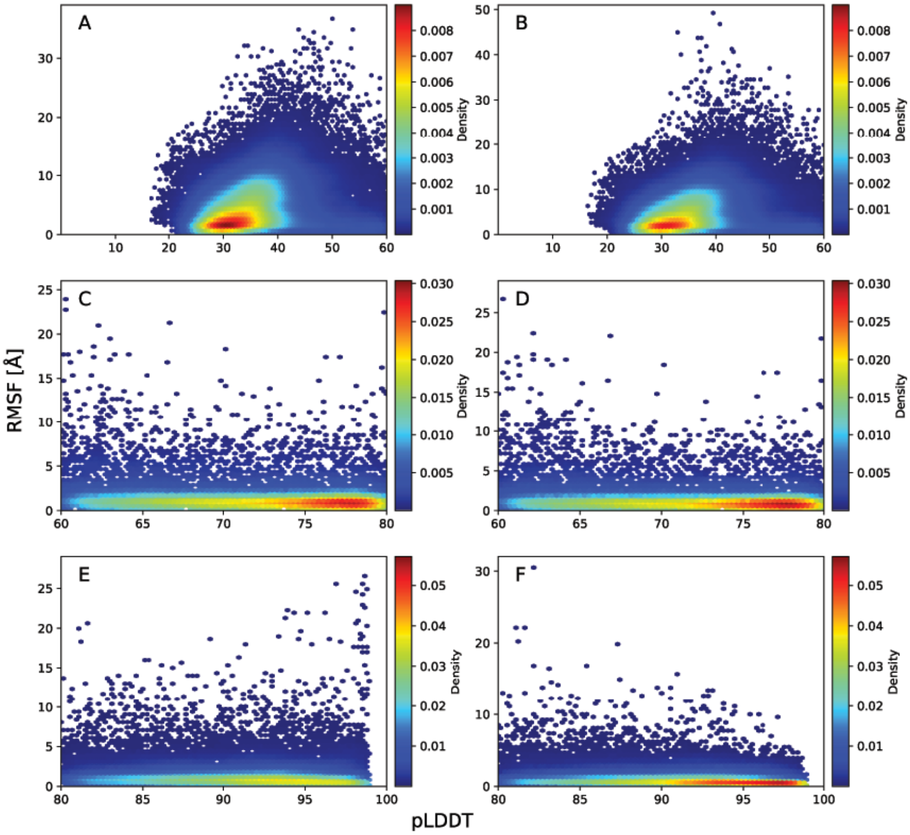
2 Additional information related to NMA

2.1 Generation of NMA data

On the S^2_{RCI} dataset of 762 proteins, WEBnma was carried out on the 762 AlphaFold2 models. Therefore, based on WEBnma output, the final dataset consists of 762 AlphaFold2 models. The predicted RMSF values of each coil residue in all 762 proteins are shown in (Supplementary Fig. 12). The supplementary figure shows some extreme RMSF values in low-pLDDT regions. As explained in the main text, these extreme values can originate artificially from loosely packed stretches in the protein structure. We have therefore adapted the RMSF analysis to reduce the artificial RMSF outliers. As shown in the following (see subsection Truncation criterion), the N- and C-terminal tails from AlphaFold2 models were truncated. The truncation criterion is based on the number of C_α contacts. The final number of truncated proteins in the dataset was 755, and the remaining 7 did not require cutting of termini. Subsequently, normal mode analysis with WEBnma was again performed on these 755 truncated models, and the RMSF was recomputed. The RMSF results are shown for 762 proteins, including both the 755 truncated models and the 7 models that did not require termini cutting (referred to as truncated S^2_{RCI} dataset). Apart from (Supplementary Fig. 12, Supplementary Fig. 13, and Supplementary Table 4) all figures and data in the main document and SI contain the RMSF of truncated S^2_{RCI} dataset. Supplementary Fig. 13 shows the effect of the truncation by comparing the RMSF before truncation and the RMSF after truncation.



Supplementary Fig. 12: **Comparison of pLDDT and RMSF in coils.** pLDDT vs RMSF of 762 proteins for coil residues before truncation. The colour bar represents the Gaussian kernel density estimate of the dataset. The red vertical lines divide the dataset into high pLDDT (≥ 80), mid ($60 \leq$ pLDDT < 80) and low (< 60) pLDDT regions.



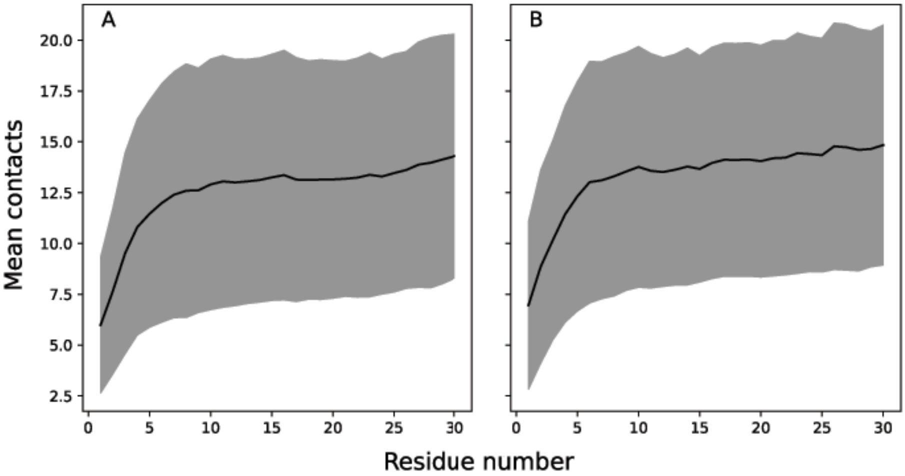
Supplementary Fig. 13: **Comparison of pLDDT and RMSF in coils.** RMSF vs pLDDT of amino acid residues exhibiting coils in non-truncated (A, C, E) and truncated (B, D, F) AlphaFold2 structures in low-pLDDT (A, B), mid-pLDDT (C, D), and c) high-pLDDT (E, F) regions. Only the amino acids that are present in both the non-truncated and truncated AlphaFold2 models are included.

Supplementary Table 4: **RMSF values of are grouped according to pLDDT in low-pLDDT, mid-pLDDT and high-pLDDT for AlphaFold2 models before truncation.** The table reports the minimum, maximum, mean, and standard deviation for each group.

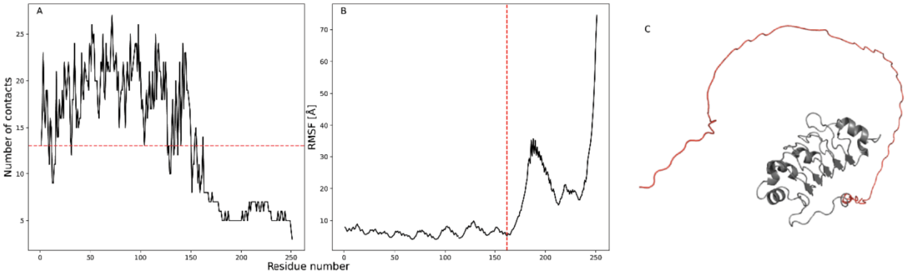
Secondary structure	pLDDT range	min	max	mean	std
Coil	low	0.22	184.99	6.28	5.83
Strand	low	0.28	8.05	1.81	1.66
α-helix	low	0.24	25.98	2.11	2.15
Turn	low	0.22	37.15	2.62	2.74
3_{10}-helix	low	0.29	27.87	2.27	3.29
Bridge	low	0.31	19.52	1.78	2.50
Coil	mid	0.21	112.52	3.00	5.16
Strand	mid	0.23	21.63	1.49	1.60
α-helix	mid	0.18	23.61	2.02	2.37
Turn	mid	0.20	33.05	2.04	2.56
3_{10}-helix	mid	0.21	27.39	2.11	3.07
Bridge	mid	0.26	13.06	1.66	2.04
Coil	high	0.15	33.97	1.58	1.84
Strand	high	0.15	29.67	1.35	1.49
α-helix	high	0.14	24.68	1.57	1.87
Turn	high	0.15	32.52	1.63	1.91
3_{10}-helix	high	0.20	30.04	1.37	1.55
Bridge	high	0.15	29.29	1.48	1.78

2.2 Truncation criterion

For determining N- and/or C-termini truncation, the C_α contacts were assessed within a 10 Å (1 nm) cut-off for each protein in the dataset. In proteins, helices and strands consistently exhibited significant contacts, surpassing approximately 13 contacts per residue across the dataset with lower RMSF ($<$ 20 Å) as shown in (Supplementary Fig. 14). An example is shown in Supplementary Fig. 15. In contrast, coils showed fewer than 13 contacts per residue and showed very high RMSF ($>$ 50 Å). Thus, a 13-contact cut-off was selected to truncate the termini. Following this criterion, all first residues with fewer than 13 contacts were cut both in the N-terminal and C-terminal. Consequently, if the first residue of an N- or C-terminal has $\geq$ 13 contacts, this terminal was not truncated. Only the termini were truncated, so an accidental low contact region in the core of the protein would not get cut.



Supplementary Fig. 14: **Mean number of contacts**. Mean number of Cα contacts within a 10 Å cut-off for first 30 residues depicting N-terminal (A) and last 30 residues depicting C-terminal (B) averaged over all 762 proteins. The black line represents the mean contacts, with the standard deviation shown as grey shaded area.



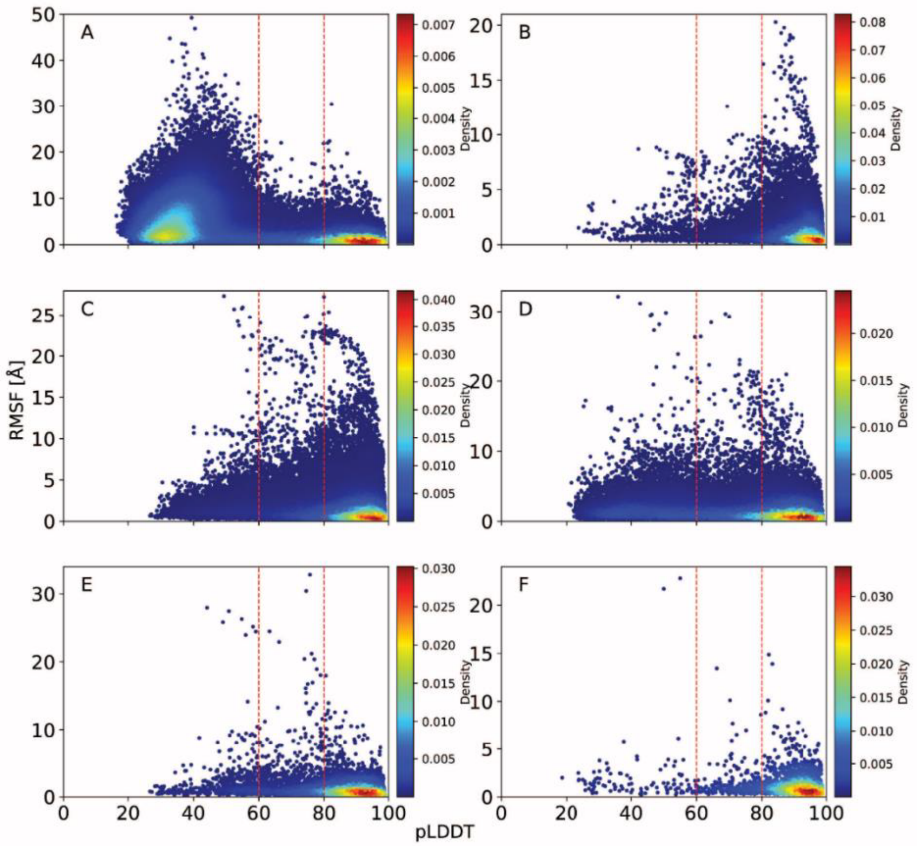
Supplementary Fig. 15: **Number of Cα contacts profile.** (A) and RMSF profile (B) of Q92688. The red dashed line in A represents the contact cut-off (13 contacts) and red dashed line in B represents the RMSF at contact cut-off. The 3D structure of Q92688 is shown in C with a red highlighted region for truncation.

2.3 Additional analysis of RMSF and correlation with pLDDT or S^2_{RCI}

The RMSF values for the dataset with truncated dataset are further analysed (now 762 proteins) according to their secondary structure element as predicted by STRIDE.

The six considered secondary structure elements are coil, strand, α-helix, turn, 3_{10}-helix, and bridge. The tables report the minimum, maximum, mean, and standard deviation for the RMSF values in each secondary structure group. Supplementary Table 5 gives three columns according to the pLDDT value as given by AlphaFold2: low-pLDDT, mid-pLDDT, and high-pLDDT. Supplementary Table 6 gives three columns according to the S^2_{RCI} value as included in the truncated S^2_{RCI} dataset: flexible, ambiguous, and rigid.

Next, the Pearson correlation coefficient between the RMSF values and the pLDDT values were computed in each group (low-pLDDT, mid-pLDDT, and high-pLDDT) in (Supplementary Table 7). Moreover, the Pearson correlation coefficient between RMSF and pLDDT was computed, without considering the subgroups of pLDDT (Supplementary Table 7). Similarly, the Pearson correlation coefficient between RMSF and S^2_{RCI} was computed for each group (flexible, ambiguous, and rigid), and without considering subgroups of S^2_{RCI} (Supplementary Table 8). For both RMSF and pLDDT, RMSF and S^2_{RCI}, the Pearson correlation was calculated for each secondary structure group, and without the classification of secondary structure. Next, the Pearson correlation coefficient between the RMSF values and S^2_{RCI} can also be computed for each individual AlphaFold2 and NMR model in the truncated S^2_{RCI} dataset. This is reported as a histogram in Supplementary Fig. 18 (blue) using the RMSF values of the 746 AlphaFold2 models and Supplementary Fig. 18 (yellow) using the RMSF values of 14,069 NMR models (as explained in the results section 3.5.2).



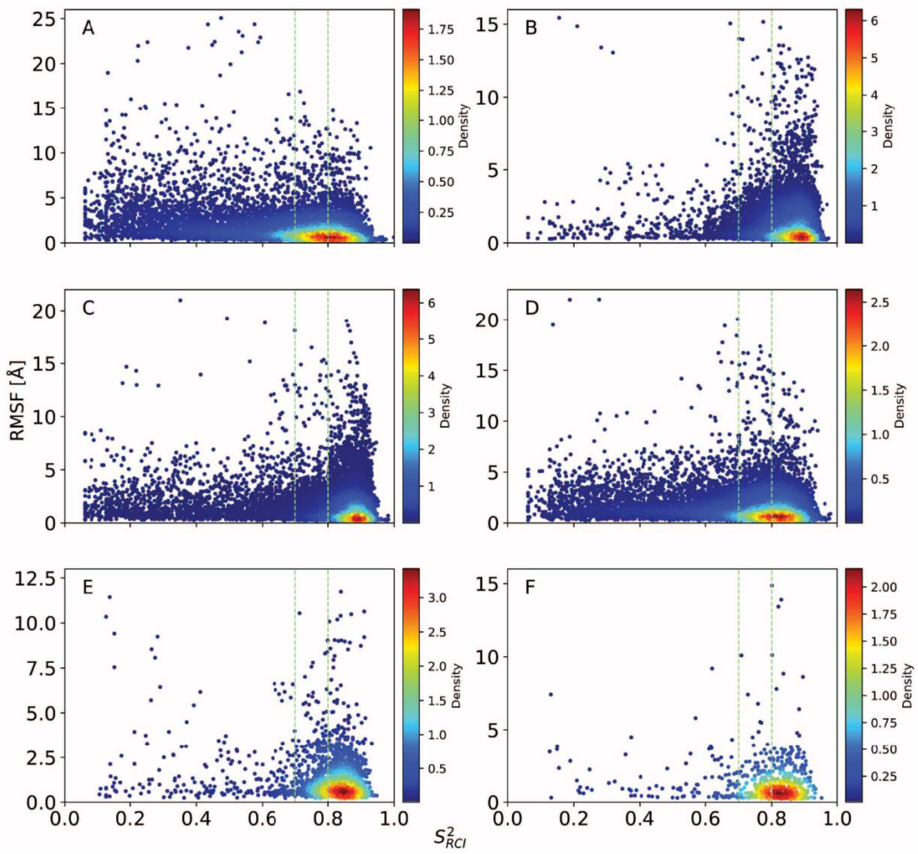
Supplementary Fig. 16: **Comparison of pLDDT and RMSF.** RMSF values versus pLDDT value of each amino acid, visualised with a Gaussian kernel estimator for S^2_{RCI} data set. One subplot for each secondary structure element: A) coil (N = 105,172), B) strand (N = 54,786), C) α-helix (N = 109,639), D) turn (N = 58,328), E) 310-helix (N = 7,931), and F) bridge (N = 2,445), where N represents number of amino acid residues. The red vertical lines divide the dataset into high pLDDT (≥ 80), mid ($60 \leq$ pLDDT < 80) and low (< 60) pLDDT regions.

Supplementary Table 5: **RMSF values are grouped according to pLDDT in low-pLDDT, mid-pLDDT and high-pLDDT.** The table reports the minimum, maximum, mean, and standard deviation for each group.

Secondary structure	pLDDT range	min	max	mean	std
Coil	low	0.22	49.24	5.65	4.43
Strand	low	0.24	8.82	1.89	1.86
α-Helix	low	0.25	27.35	1.87	1.83
Turn	low	0.20	32.16	2.16	2.03
3_{10}-helix	low	0.25	27.95	2.08	3.19
Bridge	low	0.22	22.78	1.89	3.00
Coil	mid	0.21	26.71	1.89	2.18
Strand	mid	0.21	12.60	1.35	1.46
α-Helix	mid	0.18	27.26	1.84	2.26
Turn	mid	0.15	29.68	1.74	2.16
3_{10}-helix	mid	0.21	32.80	1.89	2.78
Bridge	mid	0.26	13.43	1.34	1.53
Coil	high	0.14	30.48	1.27	1.29
Strand	high	0.14	20.28	1.11	1.13
α-Helix	high	0.13	25.39	1.38	1.65
Turn	high	0.13	20.95	1.33	1.36
3_{10}-helix	high	0.16	17.96	1.14	1.16
Bridge	high	0.15	14.87	1.20	1.16

Supplementary Table 6: **RMSF values are grouped according to $S^2_{\mathbf{RCI}}$ values in flexible ($<$ 0.7), ambiguous (0.7-0.8), and rigid($>$ 0.8).** The table reports the maximum, maximum, mean, and standard deviation for each group.

Secondary structure	S^2_{RCI}	min	max	mean	std
Coil	flexible	0.21	25.06	2.43	2.60
Strand	flexible	0.23	15.44	1.38	1.66
α-Helix	flexible	0.23	20.99	2.04	2.06
Turn	flexible	0.19	22.01	1.89	1.79
3_{10}-helix	flexible	0.23	11.44	1.70	1.94
Bridge	flexible	0.29	9.20	1.52	1.49
Coil	ambiguous	0.20	16.88	1.40	1.58
Strand	ambiguous	0.20	15.16	1.28	1.33
α-Helix	ambiguous	0.21	16.51	1.56	1.72
Turn	ambiguous	0.20	17.37	1.52	1.70
3_{10}-helix	ambiguous	0.21	10.54	1.45	1.45
Bridge	ambiguous	0.22	10.09	1.27	1.34
Coil	rigid	0.19	14.74	1.30	1.47
Strand	rigid	0.17	14.77	1.13	1.25
α-Helix	rigid	0.15	19.04	1.23	1.35
Turn	rigid	0.20	14.70	1.35	1.40
3_{10}-helix	rigid	0.21	11.74	1.21	1.32
Bridge	rigid	0.21	14.87	1.37	1.73



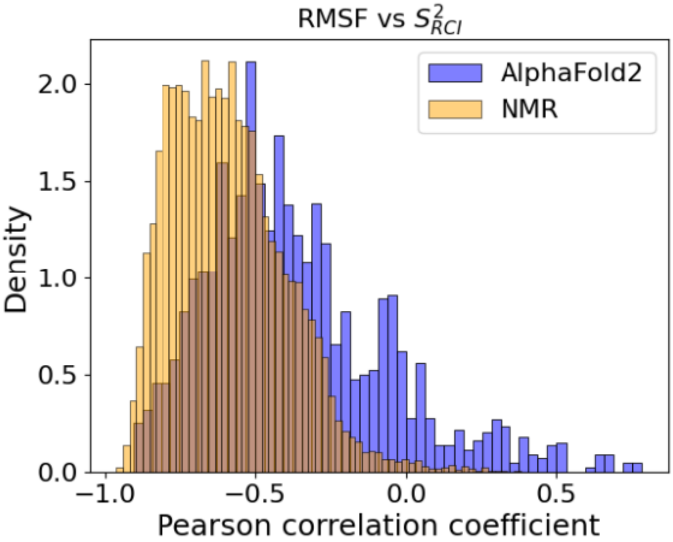
Supplementary Fig. 17: **Comparison of $S^2_{\mathbf{RCI}}$ and RMSF.** RMSF values versus S^2_{RCI} value of each amino acid, visualised with a Gaussian kernel estimator for truncated S^2_{RCI} data set. One subplot for each secondary structure element: A) coil (N = 11,634), B) strand (N = 18,640), C) α-helix (N = 25,861), D) turn (N = 14,759), E) 310-helix (N = 2,250), and F) bridge (N = 670), where N represents number of amino acid residues. The green vertical lines divide the dataset into flexible (< 0.70), ambiguous ($0.70 - 0.80$), and rigid (≥ 0.80) regions.

Supplementary Table 7: **Pearson correlation coefficients and p-values are provided for RMSF and pLDDT**. These correlations are analysed as well as for the full range of pLDDT, with and without the classification of secondary structure elements (all SS).

Subset	Secondary Structure	Pearson correlation coefficient	p-value
low-pLDDT	Coil	0.16	0.00
mid-pLDDT	Coil	-0.17	8.98 x 10^{-55}
high-pLDDT	Coil	-0.16	1.90 x 10^{-157}
all pLDDT	Coil	-0.43	0.00
low-pLDDT	Strand	0.21	4.62 x 10^{-6}
mid-pLDDT	Strand	-0.01	4.51 x 10^{-1}
high-pLDDT	Strand	-0.12	8.70 x 10^{-165}
all pLDDT	Strand	-0.12	1.07 x 10^{-185}
low-pLDDT	α-helix	0.16	9.41 x 10^{-36}
mid-pLDDT	α-helix	-0.04	9.19 x 10^{-7}
high-pLDDT	α-helix	-0.10	1.28 x 10^{-183}
all pLDDT	α-helix	-0.11	4.49 x 10^{-319}
low-pLDDT	Turn	0.07	9.71 x 10^{-15}
mid-pLDDT	Turn	-0.04	1.93 x 10^{-6}
high-pLDDT	Turn	-0.16	7.83 x 10^{-208}
all pLDDT	Turn	-0.20	0.00
low-pLDDT	3_{10}-helix	0.13	1.28 x 10^{-3}
mid-pLDDT	3_{10}-helix	-0.04	1.91 x 10^{-1}
high-pLDDT	3_{10}-helix	-0.17	3.16 x 10^{-41}
all pLDDT	3_{10}-helix	-0.19	1.08 x 10^{-67}
low-pLDDT	Bridge	0.07	4.99 x 10^{-1}
mid-pLDDT	Bridge	0.00	9.63 x 10^{-1}
high-pLDDT	Bridge	-0.18	5.08 x 10^{-17}
all pLDDT	Bridge	-0.14	3.51 x 10^{-12}
low-pLDDT	All	-0.04	1.93 x 10^{-26}
mid-pLDDT	All	-0.07	2.99 x 10^{-47}
high-pLDDT	All	-0.13	0.00
all pLDDT	All	-0.24	0.00

Supplementary Table 8: Pearson correlation coefficients and p-values are provided for RMSF and S^2_{RCI}. These correlations are analysed as well as for the full range of pLDDT, with and without the classification of secondary structure elements (all SS).

Subset	Secondary Structure	Pearson correlation coefficient	p-value
Flexible S^2_{RCI}	Coil	-0.26	3.22 x 10^{-66}
Ambiguous S^2_{RCI}	Coil	-0.07	7.65 x 10^{-5}
Rigid S^2_{RCI}	Coil	-0.05	1.27 x 10^{-3}
All S^2_{RCI}	Coil	-0.32	1.06 x 10^{-278}
Flexible S^2_{RCI}	Strand	-0.11	4.84 x 10^{-3}
Ambiguous S^2_{RCI}	Strand	-0.04	2.57 x 10^{-2}
Rigid S^2_{RCI}	Strand	-0.06	6.85 x 10^{-15}
All S^2_{RCI}	Strand	-0.08	5.2 x 10^{-26}
Flexible S^2_{RCI}	α-helix	-0.01	5.62 x 10^{-1}
Ambiguous S^2_{RCI}	α-helix	-0.08	2.07 x 10^{-4}
Rigid S^2_{RCI}	α-helix	-0.05	2.66 x 10^{-14}
All S^2_{RCI}	α-helix	-0.15	4.72 x 10^{-124}
Flexible S^2_{RCI}	Turn	-0.11	9.50 x 10^{-13}
Ambiguous S^2_{RCI}	Turn	-0.03	5.21 x 10^{-2}
Rigid S^2_{RCI}	Turn	-0.03	8.36 x 10^{-3}
All S^2_{RCI}	Turn	-0.15	1.15 x 10^{-76}
Flexible S^2_{RCI}	3_{10}-helix	-0.16	1.23 x 10^{-2}
Ambiguous S^2_{RCI}	3_{10}-helix	-0.02	7.13 x 10^{-1}
Rigid S^2_{RCI}	3_{10}-helix	-0.08	3.34 x 10^{-3}
All S^2_{RCI}	3_{10}-helix	-0.14	4.67 x 10^{-11}
Flexible S^2_{RCI}	Bridge	-0.14	1.34 x 10^{-1}
Ambiguous S^2_{RCI}	Bridge	-0.18	1.44 x 10^{-2}
Rigid S^2_{RCI}	Bridge	-0.07	1.91 x 10^{-1}
All S^2_{RCI}	Bridge	-0.07	6.56 x 10^{-2}
Flexible S^2_{RCI}	All	-0.17	6.97 x 10^{-75}
Ambiguous S^2_{RCI}	All	-0.05	4.03 x 10^{-10}
Rigid S^2_{RCI}	All	-0.06	3.59 x 10^{-44}
All S^2_{RCI}	All	-0.22	0.00



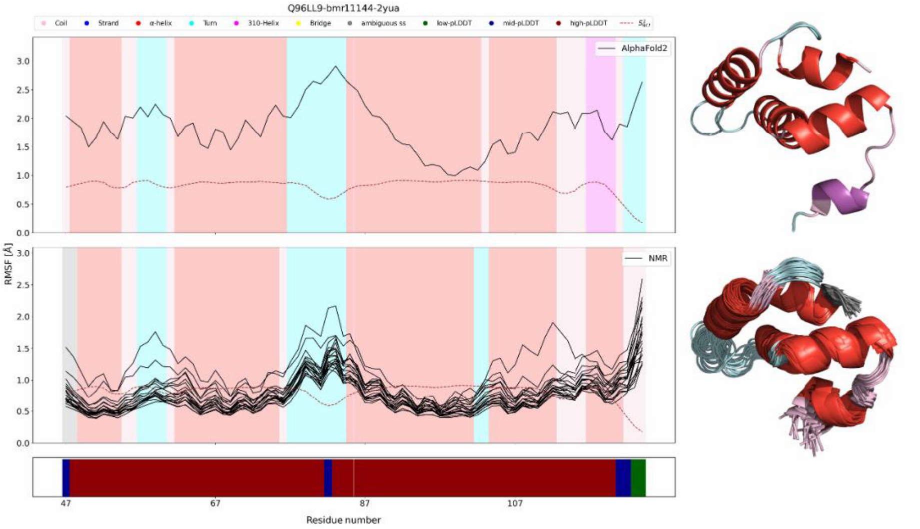
Supplementary Fig. 18: **Pearson correlation coefficients between RMSF and $S^2_{\mathbf{RCI}}$.** Distribution of Pearson correlation coefficients between RMSF values and S^2_{RCI} values of amino acids for AlphaFold2 models (blue) and NMR models (yellow).

2.4 Examples of $S^2_{\mathbf{RCI}}$ and RMSF correlation for Alphafold2 and NMR models

The correlation between the per-residue RMSF and per-residue S^2_{RCI} values of a given protein is generally expected to be negative, because rigid regions would correspond to high RMSF and low S^2_{RCI}. There were a few NMR models that showed (unexpected) positive correlation between S^2_{RCI} and RMSF.

2.4.1 Example of a protein where the AlphaFold2 model has a slightly weaker negative $S^2_{\mathbf{RCI}}$ versus RMSF correlation than the NMR model

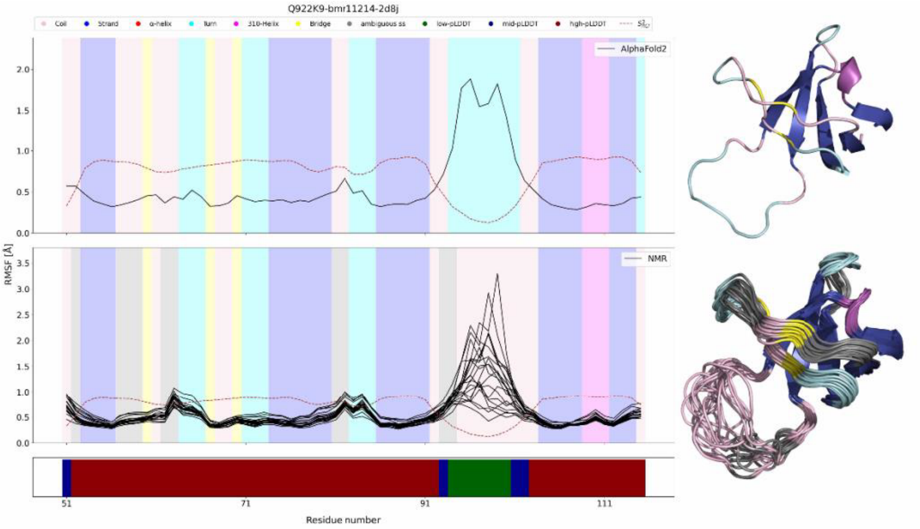
The Pearson correlation coefficient between RMSF and S^2_{RCI} values is -0.52 for the AlphaFold2 model of protein Q96LL9-2YUA (BMRB id 11144). For the 20 NMR structures, 19 have a Pearson correlation coefficient between RMSF and S^2_{RCI} that lie in the range -0.65 to -0.81 and the remaining model shows -0.40. Therefore, the AlphaFold2 model has weaker correlation than most of the NMR models. In addition, in the figures below, the NMR model shows residues with ambiguous secondary structure (grey). The ambiguous residues were computed by the dumb consensus of secondary structure with a 70% threshold within the NMR ensemble.



Supplementary Fig. 19: **Example of a protein (Q96LL9-2YUA).** The 3D structures with secondary structure mapping colours are shown on the right: AlphaFold2 model (right top) and NMR ensemble (right bottom, 20 NMR models shown at once). Comparing S^2_{RCI} (red, dashed) and RMSF (black line) of the AlphaFold2 model and the NMR ensemble (for this protein, there were 20 NMR models within the ensemble). The secondary structure is indicated with shaded regions. The pLDDT of the sequence is shown below the plots. (Color legends at the top of the figure.)

2.4.2 Example of a protein where the AlphaFold2 model has a slightly stronger negative $S^2_{\mathbf{RCI}}$ versus RMSF correlation than (most of) the NMR models

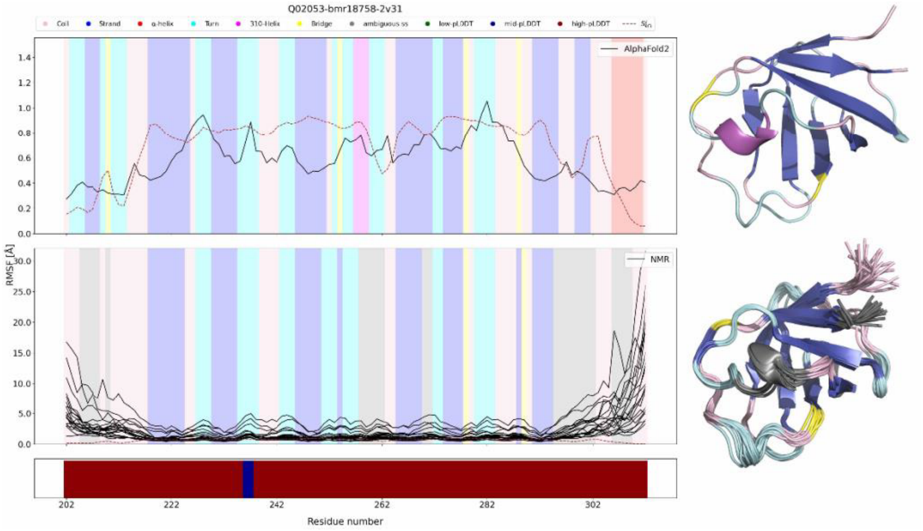
The Pearson correlation coefficient between RMSF and S^2_{RCI} values is −0.91 for the AlphaFold2 model of protein Q922K9-2D8J (BMRB id: 11214). For the 20 NMR models within the ensemble, the Pearson correlation coefficient between RMSF and S^2_{RCI} lie in the range −0.63 to −0.91, where 19 models show correlation coefficient below −0.91. Therefore, the AlphaFold2 model has stronger negative correlation than most of the NMR models.



Supplementary Fig. 20: **Example of a protein (Q922K9-2D8J).** The 3D structures with secondary structure mapping colours are shown on the right: truncated AlphaFold2 model (right top) and NMR ensemble (right bottom, 20 NMR models within the ensemble shown at once). Comparing S^2_{RCI} (red, dashed) and RMSF (black line) of the AlphaFold2 model and the NMR models (for this protein, there were 20). The secondary structure is indicated with shaded regions. The pLDDT of the sequence is shown below the plots.

2.4.3 Example of a protein where the AlphaFold2 model has a (unexpected) positive $S^2_{\mathbf{RCI}}$ versus RMSF correlation, while the NMR models show negative correlation

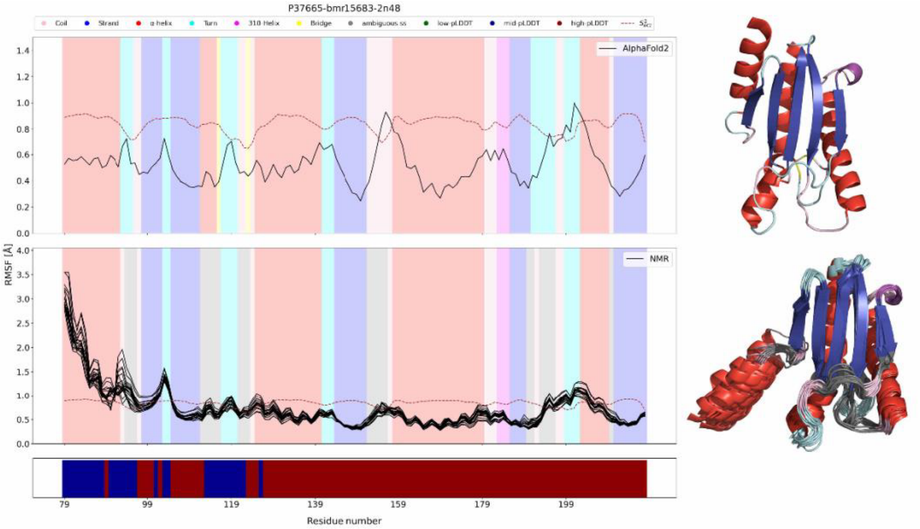
The Pearson correlation coefficient between RMSF and S^2_{RCI} values is 0.63 for the AlphaFold2 model and ranges from -0.62 to -0.85 for the 20 NMR models within the ensemble of protein Q02053-2V31 (BMRB id 18758).



Supplementary Fig. 21: **Example of a protein (Q02053-2V31).** The 3D structures with secondary structure mapping colours are shown on the right: truncated AlphaFold2 model (right top) and NMR ensemble (right bottom, 20 NMR models within the ensemble shown at once). Comparing S^2_{RCI} (red, dashed) and RMSF (black line) of the AlphaFold2 model and the NMR models (for this protein, there were 20). The secondary structure is indicated with shaded regions. The pLDDT of the sequence is shown below the plot.

2.4.4 Example of a protein where the AlphaFold2 model has a negative $S^2_{\mathbf{RCI}}$ versus RMSF correlation, while the NMR structure shows (unexpected) positive correlation.

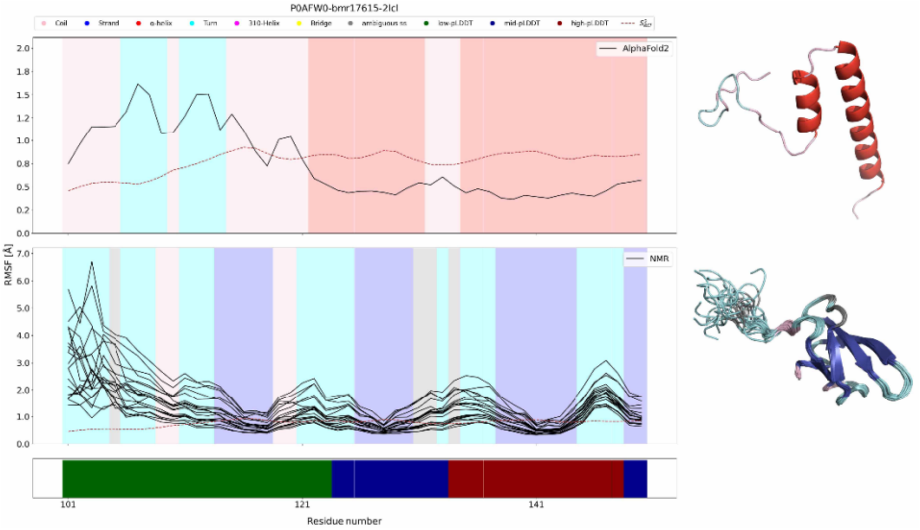
The Pearson correlation coefficient between RMSF and S^2_{RCI} values is -0.40 for the AlphaFold2 model and ranges from 0.00 to 0.09 for the 20 NMR structures of protein P37665-2N48. (BMRB id 15683).



Supplementary Fig. 22: **Example of a protein (P37665-2N48).** The 3D structures with secondary structure mapping colours are shown on the right: truncated AlphaFold2 model (right top) and NMR ensemble (right bottom, 20 NMR models within the ensemble shown at once). Comparing S^2_{RCI} (red, dashed) and RMSF (black line) of the AlphaFold2 model and the NMR models (for this protein, there were 20). The secondary structure is indicated with shaded regions. The pLDDT of the sequence is shown below the plots.

2.4.5 Example of a protein where the 88% of overlapping amino acid sequence between AlphaFold2 NMR shows conflicting secondary structure.

The Pearson correlation coefficient between RMSF and S^2_{RCI} values is -0.55 for the AlphaFold2 model and ranges from -0.74 to -0.86 for the 20 NMR structures of protein P0AFW0-2LCL (BMRB id 17615).



Supplementary Fig. 23: **Example of a protein (P0AFW0-2LCL).** The 3D structures with secondary structure mapping colours are shown on the right: AlphaFold2 model overlapping with NMR sequence (right top) and NMR ensemble (right bottom, 20 NMR models within the ensemble shown at once). Comparing S^2_{RCI} (red, dashed) and RMSF (black line) of the truncated AlphaFold2 model and the NMR models (for this protein, there were 20). The secondary structure is indicated with shaded regions. The pLDDT of the sequence is shown below the plot. The sequence in the RMSF plot shows sequence from 101-150 amino acids, while the structure shows 101-161 amino acids.

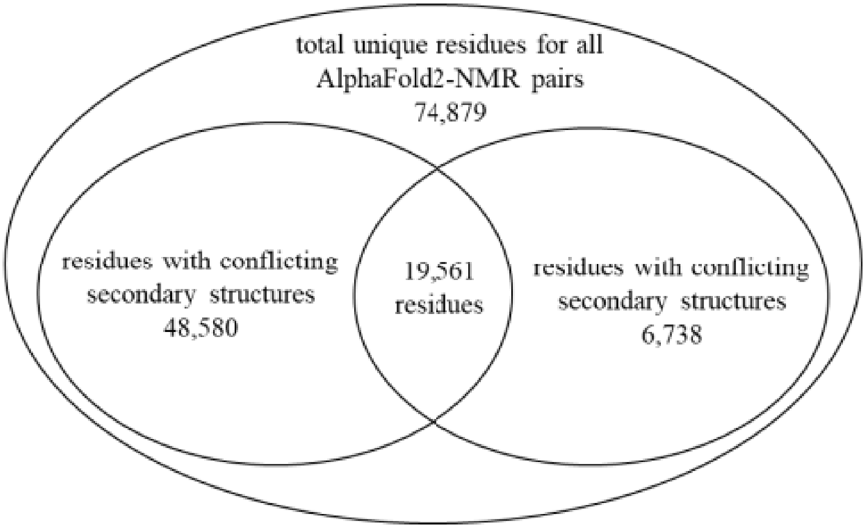
3 Conflicting secondary structure elements between AlphaFold2 and NMR models

Using STRIDE, a secondary structure (SS) element is assigned to each residue of the AlphaFold2 model of a protein, and to each residue of the models in the NMR ensemble of the protein. When a residue has an equal assignment in all models (one AlphaFold2 model and one (or more) NMR models), we say that the residue has identical SS. When a residue has a different assigned SS in the AlphaFold2 model compared to its assigned SS in all the protein's NMR models, we say that the residue has a conflicting SS. Besides these residues with conflicting SS and identical SS, there is a third group of residues: a residue might have an AlphaFold2 assigned SS that is identical to the SS in some of the NMR models but conflicting in some of the other NMR models of the protein.

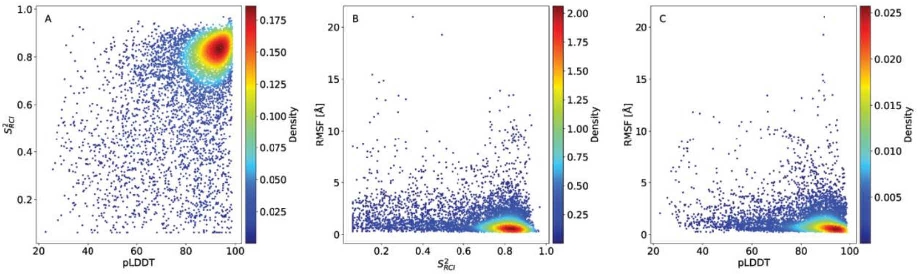
There are 746 unique proteins with 746 AlphaFold2 models and 746 NMR ensembles (totaling 14,069 NMR models), corresponding to 14,069 AlphaFold2-NMR pairs (see main text). Out of the 74,879 unique residues of these proteins that are present in the AlphaFold2 sequence and the NMR models (overlapping), several residues (19,561) from one or more NMR models of the same ensemble exhibit indeed both conflicting and identical secondary structures. This variability arises because different NMR models within the same ensemble can show different secondary structures for the same residues. These residues are shown as the overlap between conflicting and identical secondary structures in Supplementary Fig. 24. The distribution of S^2_{RCI}, RMSF, and pLDDT for residues with conflicting secondary structures (6,738 residues) is shown in Supplementary Fig. 25. The Pearson correlation coefficient between S^2_{RCI} and RMSF for residues with conflicting secondary structures (SS) is -0.17 (p-value $= 1.26 \times 10^{-44}$, $N = 6,258$ where N is the number of amino acids with S^2_{RCI} available values). For S^2_{RCI} and pLDDT, the Pearson correlation is 0.44 (p-value $= 0.94 \times 10^{-308}$, $N = 6,258$), and it is -0.14 (p-value $= 6.96 \times 10^{-35}$, $N = 6,738$) for RMSF and pLDDT.

We also examined the conflicting SS residues for each structure in 14,069 AlphaFold2-NMR pairs, identifying a total of 14,006 AlphaFold2-NMR pairs with conflicting SS residues. For these 14,006 pairs, we computed the difference in the Pearson correlation coefficients ($\rho_{k,m}^{\mathrm{NMR}}$ and $\rho_{k,m}^{\mathrm{AF2}}$, for detailed explanation see results section 3.5.2 in main) of AlphaFold2-NMR pairs (Eq. 1). The values $\Delta\rho_{k,m} > 0$ indicates that $\rho_{k,m}^{\mathrm{NMR}}$ is stronger than $\rho_{k,m}^{\mathrm{AF2}}$, while $\Delta\rho_{k,m} < 0$ indicates that $\rho_{k,m}^{\mathrm{AF2}}$ is stronger than $\rho_{k,m}^{\mathrm{NMR}}$. Out of 14,006 AlphaFold2-NMR pairs, 9,994 showed stronger $\rho_{k,m}^{\mathrm{NMR}}$, and the remaining 4,012 pairs showed stronger $\rho_{k,m}^{\mathrm{AF2}}$. The distribution of conflicting SS residues for both cases is shown in Supplementary Fig. 26. For AlphaFold2 models where correlation between S^2_{RCI} vs RMSF is stronger than NMR models, the percentage of conflicting SS residues range from 0.86% to 80.48%, with an average of $19.46 \pm 10.18\%$ conflicting SS residues across the overlapping sequences of AlphaFold2 and NMR models. In comparison, for NMR models, the range is from 1.01% to 88.00% with an average of $19.07 \pm 9.91\%$ conflicting SS residues.

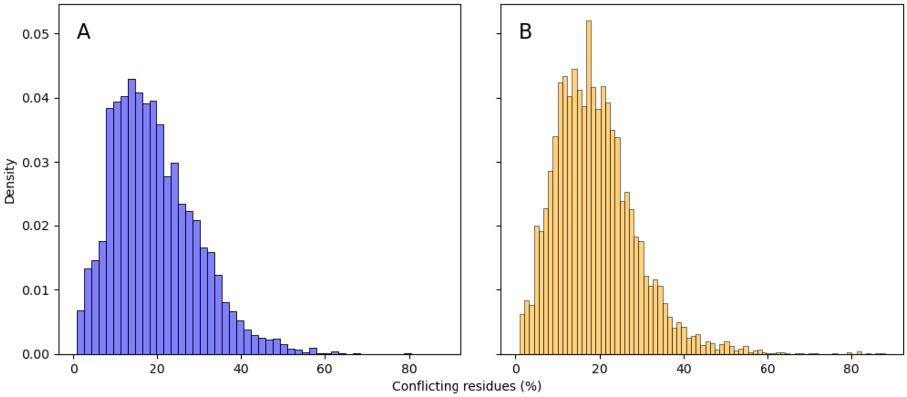
$$\Delta\rho_{k,m} = \rho_{k,m}^{\mathrm{AF2}} - \rho_{k,m}^{\mathrm{NMR}} \tag{1}$$



Supplementary Fig. 24: **Total conflicting secondary structure residues.** The Venn diagram representing total number of unique residues for 14,069 AlphaFold2-NMR pairs with conflicting secondary structure residues and identical secondary structure residues.



Supplementary Fig. 25: **Conflicting secondary structure residues.** A) S^2_{RCI} vs pLDDT, B) S^2_{RCI} vs RMSF, and C) pLDDT vs RMSF of 6,738 residues with conflicting secondary structures between AlphaFold2-NMR pairs are shown. The A, B, and C are visualized with a Gaussian kernel estimator between their corresponding x-axis and y-axis variables.



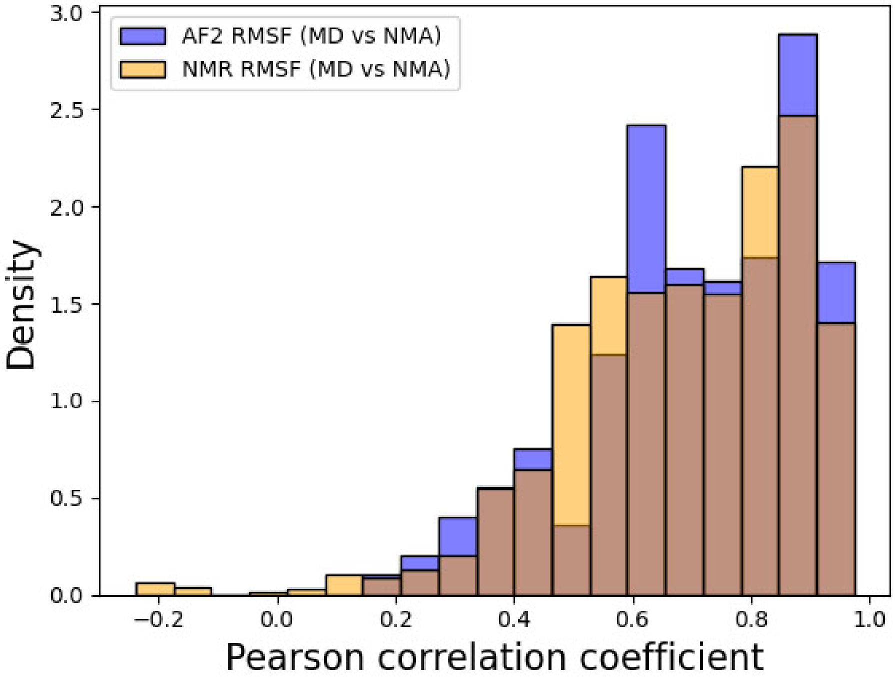
Supplementary Fig. 26: **Distribution of conflicting SS residues in AlphaFold2-NMR pairs.** The distribution of conflicting SS residues is shown as percentage on x-axis for A) AlphaFold2 models where the correlation between S^2_{RCI} vs RMSF is stronger than NMR models, and B) NMR models where the correlation between S^2_{RCI} vs RMSF is stronger than AlphaFold2 models.

4 Correlation assessment between MD and NMA-derived RMSF in the Constava dataset

In the Constava dataset of 100 AlphaFold2 structures, we calculated the Pearson correlation coefficients between the RMSF profiles derived from MD simulations and those obtained from NMA. Out of all structures, all but one protein (PDB ID 2l95) exhibited statistically significant correlations (p-value <0.05). Among these, 91 structures displayed moderate to very strong correlations, with coefficients ranging from 0.40 to 0.95, while the remaining eight structures demonstrated weaker correlations (<0.40) (Supplementary Fig. Supplementary Fig. 27). Importantly, the lower correlation observed in most cases was not directly linked to the pLDDT scores, as most structures were composed predominantly of high-pLDDT regions, with mid- and low-pLDDT regions being infrequent in the core and more commonly found at the termini.

Supplementary Table 9: **Examples of proteins with Pearson correlation coefficients between $\mathbf{S^2_{RCI}}$ vs RMSF.** The Pearson correlation coefficients of specific proteins with their unique UniProt ID are reported for their respective AlphaFold2 and NMR models, including the number of conflicting residues occurring between the overlapping sequence of between the AlphaFold2 and NMR models, and total number of residues in the non-truncated and truncated AlphaFold2 models. For the NMR models, an example of only one structure from the NMR ensemble is provided.

UniProt ID	Correlation coefficient (AF2)	Correlation coefficient (NMR)	# conflicting residues	Total overlapping residues	Conflicting residues (%)	Total # residues (non-truncated)	Total # residues (non-truncated)
C3VPR6	-0.75	-0.73	13.85	87	15.92	1915	1898
O43157	-0.60	-0.46	21.95	112	19.60	2135	2122
O60885	-0.65	-0.78	15.00	83	18.07	1362	1315
P00519	-0.89	-0.83	26.50	97	27.32	1130	1081
P16157	-0.64	-0.66	15.70	104	15.10	1881	1817
P26039	-0.82	-0.78	12.00	132	9.09	2541	2523
P35670	-0.81	-0.86	29.00	161	18.01	1465	1352
P36006	-0.84	-0.87	13.05	69	18.91	1272	1230
P38398	-0.58	-0.57	29.07	104	27.95	1863	1851
P59046	-0.79	-0.68	28.00	93	30.11	1061	1056
Q04656	-0.73	-0.68	57.15	181	31.57	1500	1436
Q53SF7	-0.52	-0.76	19.15	79	24.24	1128	1041
Q63HR2	-0.54	-0.79	23.55	114	20.66	1409	1375
Q92625	-0.58	-0.76	26.00	80	32.50	1134	1069
Q9P212	-0.73	-0.86	25.95	101	25.69	2302	1881



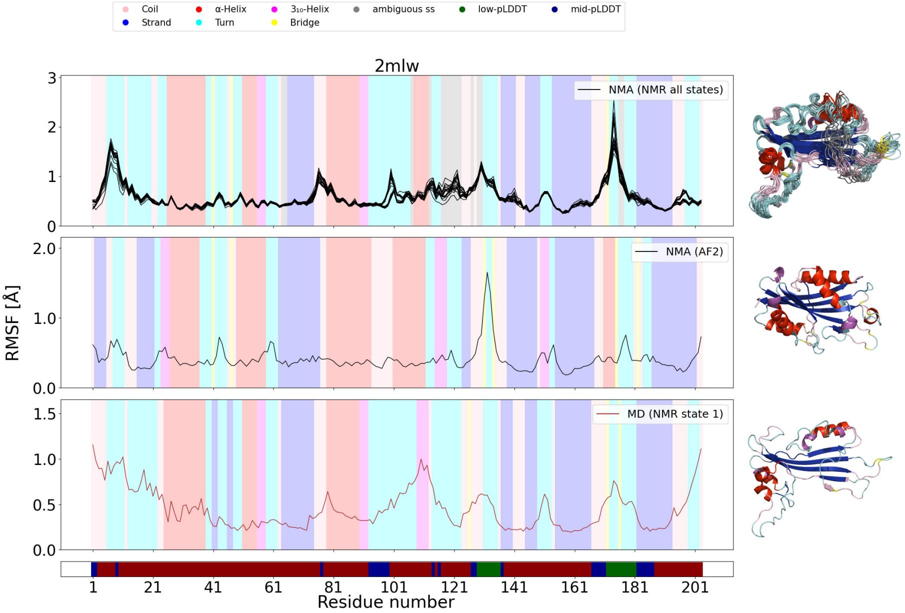
Supplementary Fig. 27: **Pearson correlation coefficients between RMSF (MD) and RMSF (NMA).** Distribution of Pearson correlation coefficients calculated between the RMSF profiles of AlphaFold2 models derived from NMA and those obtained from MD simulations for the Constava dataset.

The differences in RMSF values primarily arise from variations in secondary structure elements between AlphaFold2 models subjected to NMA and their corresponding MD models (first model of the NMR ensemble as .gro files from the Constava dataset) subjected to MD simulations. To illustrate this behavior, we present two specific cases: 2mlw and 2m2u. For these cases, secondary structures were computed using STRIDE for their AlphaFold2 models and MD models. The Pearson correlation coefficient between two RMSF profiles of 2mlw is 0.25 (p-value 1.99×10^{-4}), while 2m2u exhibits a Pearson correlation coefficient of 0.29 (p-value 7.86×10^{-5}) (Supplementary Fig. Supplementary Fig. 28, Supplementary Fig. 29). Although the AlphaFold2 model of 2mlw shows overall high pLDDT scores, the two RMSF profiles from MD and NMA differ significantly, likely due to substantial conflicts in secondary structures across both high- and mid-pLDDT regions in the two models Supplementary Fig. 28. This behavior is also evident in 2m2u, where consistently high pLDDT scores in the AlphaFold2 model are accompanied by minor conflicts in secondary structure compared to its MD model, which may account for significant differences in their RMSF profiles (Supplementary Fig. 29). These findings indicate that conflicts in secondary structures between protein models derived from various computational and experimental methods may lead to differing RMSF values obtained from NMA and MD simulations, particularly when these conflicts involve transitions from rigid to flexible secondary structures and vice versa. However, these discrepancies between RMSFs obtained from two metrics may differ from protein to protein. In such cases, pLDDT may not be a reliable indicator for capturing the gradations of protein dynamics even in high-pLDDT regions. Additionally, truncation of AlphaFold2 models leads to the formation of the new termini, which resulted in slightly different RMSF values at the new termini as observed in NMA compared to the full-length NMR structures analyzed through MD simulations. This discrepancy further contributes to the observed lower Pearson correlation. An example of 2l95 is provided, which shows differences in the RMSF profiles in the C-termini in its AlphaFold2 model and MD model (Pearson 0.20, p-value 0.11) (Supplementary Fig. 30).

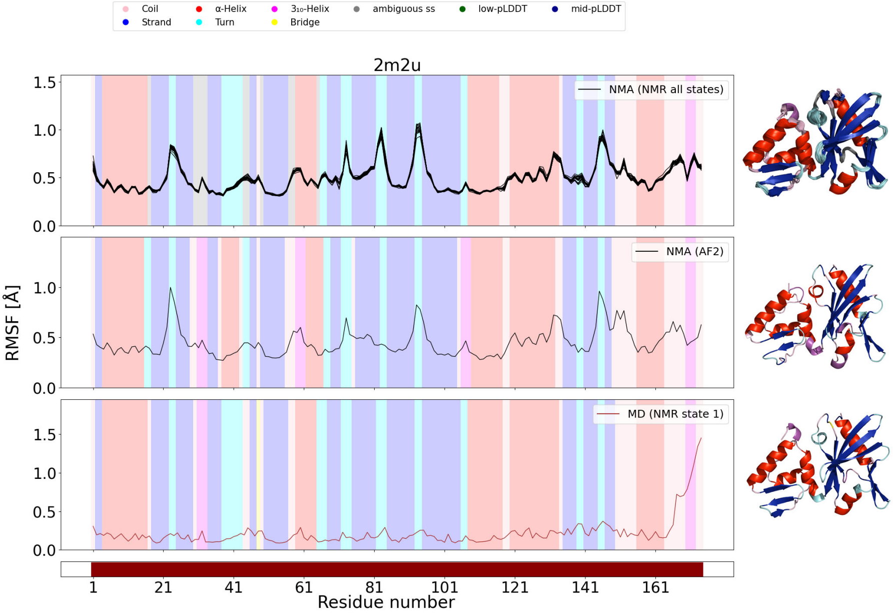
Next, to investigate how fluctuations from MD simulations correlate with those from NMA on NMR models, NMA was carried out on the NMR ensembles from the Constava dataset. Out of the 100 NMR ensembles, one (PDB ID: 1rqs) encountered a WEBnma error (invalid bond distance) for all models within the ensemble. Similar to the correlation analysis presented in the manuscript for S^2_{RCI} and NMA RMSF (section 3.5.2), the Pearson correlation coefficient was computed between the RMSF values from MD simulations and those from NMA for each individual AlphaFold2 model in the Constava dataset and its corresponding NMR models. This is displayed as a histogram in Supplementary Fig. 27 (blue) for the RMSF values of the 99 AlphaFold2 models, and in Supplementary Fig. 27 (yellow) for the RMSF values of the 1,556 NMR models. The RMSFs obtained from two different computational methods, MD simulations and NMA, represent dynamics across multiple timescales and are expected to exhibit a positive correlation. However, an outlier (2l95) with negative correlation coefficient in NMR models was observed due as discussed previously.

To determine whether fluctuations from MD correlates differently with the NMA fluctuations of AlphaFold2 models or of the NMR models, a non-parametric Wilcoxon signed-rank test was performed on the distributions of these correlation coefficients. With a p-value of 0.0001, it indicates a significant difference between the AlphaFold2 and NMR model correlation coefficients. Next, the differences between Pearson correlation coefficients for AlphaFold2 and NMR models were computed (see eq. 5) for the Constava dataset. The differences between the two Pearson correlation coefficients for 1,556 AlphaFold2-NMR pairs ranged from -0.40 to 0.40. Although the correlation coefficients for the AlphaFold2 and NMR models of same protein differ only slightly, with a mean difference 0.02 ± 0.12 over 1,556 pairs. Of the 1,556 pairs, 56.04% (872 pairs from 69 unique PDB IDs) showed a slightly stronger positive correlation in the AlphaFold2 models, while 43.96% (684 pairs and 69 unique PDB IDs) showed a slightly stronger positive correlation in the NMR models. A total of 55 unique PDB IDs exhibit a stronger correlation in both their AlphaFold2 and NMR models (across one or more states). Additionally, 30 PDB IDs (including 2l95) show a slightly stronger correlation exclusively in the AlphaFold2 models compared to the NMR models, while 14 display a stronger correlation exclusively in the NMR models. When comparing RMSF values from NMA of AlphaFold2 models and MD simulations of NMR models, variations in correlation can occur due to differences in how the termini are handled. In AlphaFold2 models, the flexible termini are sometimes truncated, which can reduce RMSF values if these regions adopt more rigid secondary structures. This truncation can occasionally increase the correlation with RMSF values from MD simulations, where the full termini remain intact and retain their flexibility. In contrast, when performing NMA on the full NMR ensemble without truncating the termini, the additional flexibility in these regions may slightly increase or decrease correlation coefficients relative to AlphaFold2 NMA and MD simulations. An example of 2l9n is provided, which shows correlation coefficients ranging from 0.42 to 0.64 across 20 NMR models, and 0.84 in its AlphaFold2 model (Supplementary Fig. Supplementary Fig. 31). In addition, some NMR ensembles exhibit significant variability in their individual RMSF profiles, particularly in highly flexible regions (Supplementary Fig. Supplementary Fig. 31). This variability within the ensemble can further influence the correlation coefficients. Together, these factors along with low-, and mid-pLDDT contribute to the nuanced differences observed in correlation between RMSF profiles across NMA of AlphaFold2, MD simulations, and NMA of the full NMR ensemble.

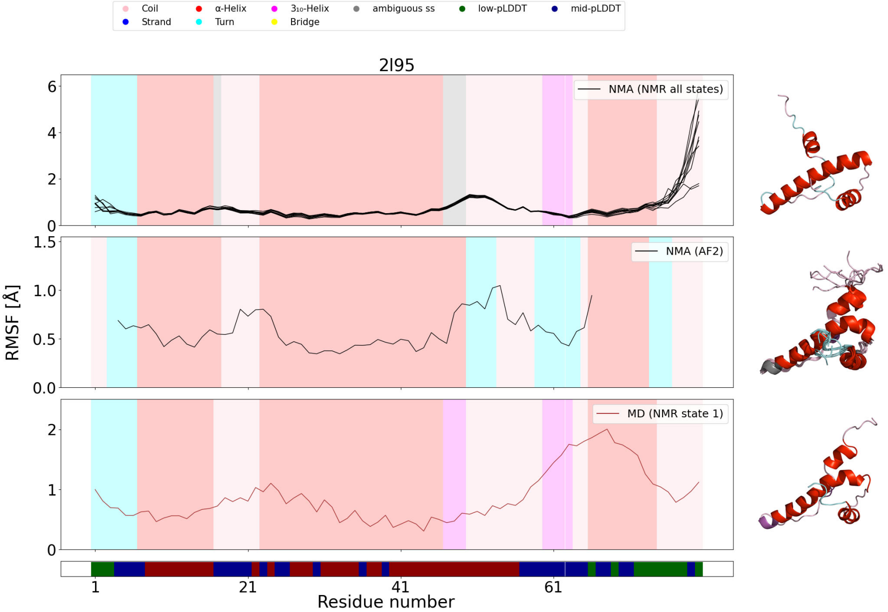
Overall, these results suggest that NMA is capable of capturing trends in protein flexibility that are consistent with those observed in MD simulations. However, when comparing RMSF profiles from these two computational methods, it is important to consider that MD simulations incorporate solvent effects and time-dependent thermal fluctuations, factors that are not captured in NMA. Therefore, while NMA can provide valuable insights into the global motions and flexibility of proteins, a comprehensive comparison between NMA and MD requires careful consideration of the additional dynamic factors that MD simulations account for, quality of input structures, differences in secondary structures, offering a more nuanced and complete understanding of protein dynamics.



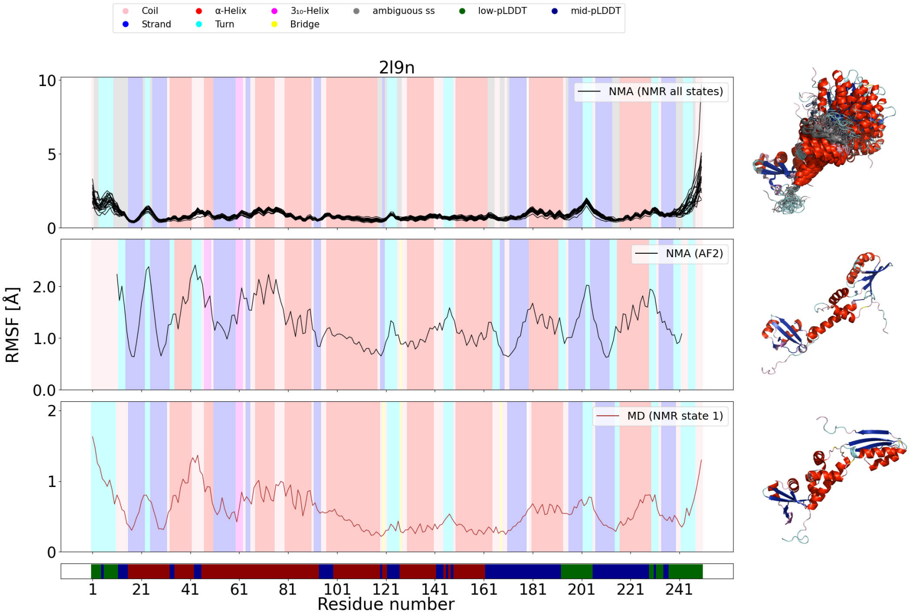
Supplementary Fig. 28: **Example of a protein (2MLW).** The 3D structures with secondary structure mapping colours are shown on the right: NMR ensemble (right top), truncated AlphaFold2 model (right middle), and MD input model (right bottom). Comparing RMSF obtained from NMA (black line) of the each model of the NMR ensemble, AlphaFold2 model and RMSF obtained from MD simulations (brown line) on the input MD model. The secondary structure is indicated with shaded regions. The pLDDT of the sequence is shown below the plot.



Supplementary Fig. 29: **Example of a protein (2M2U).** The 3D structures with secondary structure mapping colours are shown on the right: NMR ensemble (right top), truncated AlphaFold2 model (right middle), and MD input model (right bottom). Comparing RMSF obtained from NMA (black line) of the each model of the NMR ensemble, AlphaFold2 model and RMSF obtained from MD simulations (brown line) on the input MD model. The secondary structure is indicated with shaded regions. The pLDDT of the sequence is shown below the plot.



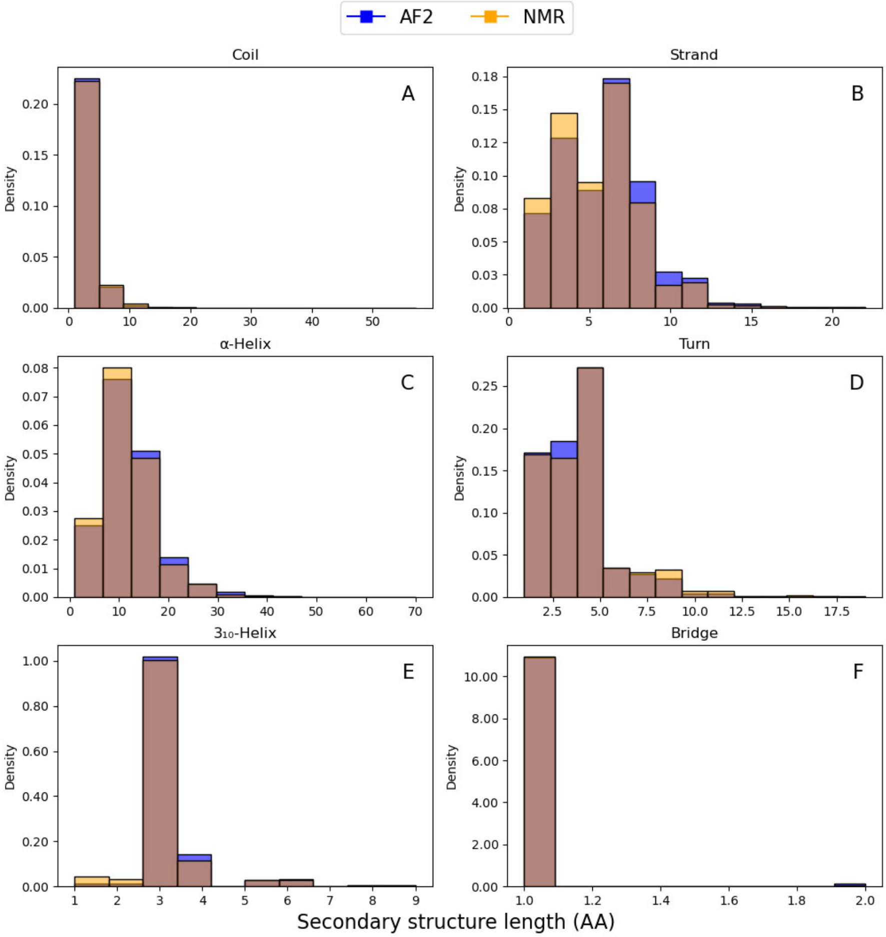
Supplementary Fig. 30: **Example of a protein (2L95).** The 3D structures with secondary structure mapping colours are shown on the right: NMR ensemble (right top), truncated AlphaFold2 model (right middle), and MD input model (right bottom). Comparing RMSF obtained from NMA (black line) of the each model of the NMR ensemble, AlphaFold2 model and RMSF obtained from MD simulations (brown line) on the input MD model. The secondary structure is indicated with shaded regions.



Supplementary Fig. 31: **Example of a protein (2L9N).** The 3D structures with secondary structure mapping colours are shown on the right: NMR ensemble (right top), truncated AlphaFold2 model (right middle), and MD input model (right bottom). Comparing RMSF obtained from NMA (black line) of the each model of the NMR ensemble, AlphaFold2 model and RMSF obtained from MD simulations (brown line) on the input MD model. The secondary structure is indicated with shaded regions. The pLDDT of the sequence is shown below the plot.

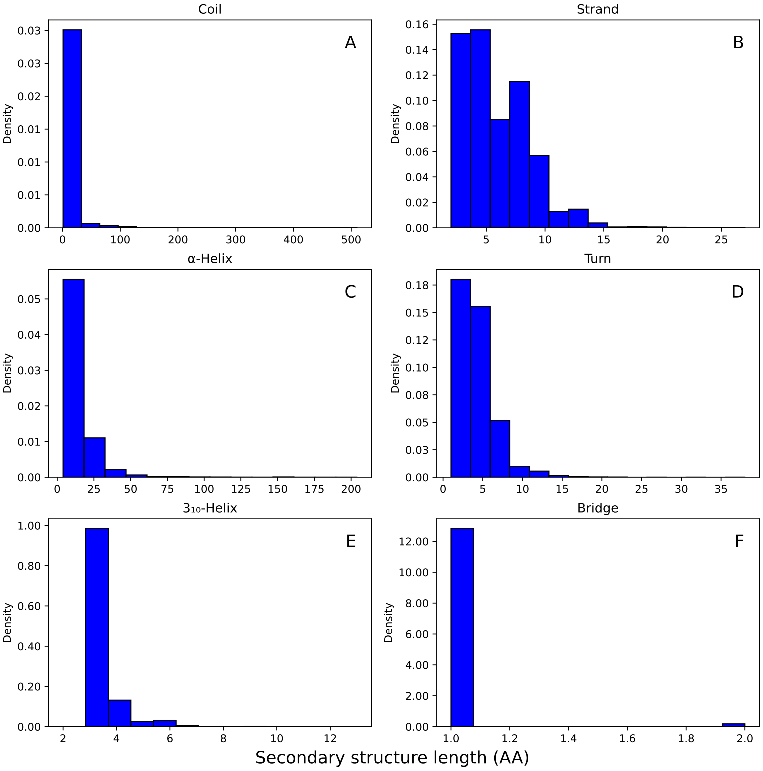
5 Secondary structure content variation

To examine variations in secondary structure content between AlphaFold2 and NMR models within the S^2_{RCI} dataset, the amino acid sequence lengths for six secondary structure types: coil, strand, α-helix, turn, 3_{10}-helix, and bridge were computed. This analysis focused on the overlapping sequence regions between each AlphaFold2 model and its corresponding NMR model. For AlphaFold2 models, secondary structure assignments derived from STRIDE were compared with the STRIDE consensus secondary structures of the NMR models. To ensure consistency in the analysis, secondary structure elements consisting of only a single residue were generally excluded, with the exception of bridges. For each AlphaFold2-NMR pair, if a single-residue secondary structure element, such as a turn, occurred in one model (either NMR or AlphaFold2) but was absent in the corresponding region of the other model, that residue was excluded from the analysis. This step was applied to ensure optimal alignment of secondary structure elements between models and to reduce inconsistencies in the comparative analysis. The distributions of the sequence length of each secondary structure are between AlphaFold2 and their corresponding NMR models are provided in Supplementary Fig. 32. The results show that both AlphaFold2 and NMR models demonstrate similar secondary structure content distributions with minor variations. AlphaFold2 shows a slight preference for longer strands, α-helices, 3_{10}-helices and bridges, likely due to its predictive model favoring more rigid regions compared to the NMR models. Conversely, NMR models tend to slightly favor longer turns and coils. This slight preference for longer turns and coils in the NMR data suggests that these regions may exhibit greater flexibility, reflecting the dynamic behavior of proteins in solution. AlphaFold2, by contrast, may model these regions as slightly shorter and more rigid, aligning with a more static structural view. This difference might reflect NMR's ability to capture a broader range of conformational variability than AlphaFold2.



Supplementary Fig. 32: **Sequence length of secondary structure elements in AlphaFold2 vs NMR models.** The distribution of amino acid (AA) sequence lengths for six secondary structure elements is shown for truncated AlphaFold2 models (blue) and NMR models (yellow) across 746 proteins for S^2_{RCI} data set. The lengths are calculated over overlapping sequences between the AlphaFold2 and NMR models for each protein. The subplots are shown as A) coil (N_{AF2} = 4,914, N_{NMR} = 5,032), B) strand (N_{AF2} = 3,183, N_{NMR} = 2,959), C) α-helix (N_{AF2} = 2,124, N_{NMR} = 2,000), D) turn (N_{AF2} = 4,022, N_{NMR} = 4,184), E) 3_{10}-helix (N_{AF2} =708, N_{NMR} = 418), and F) bridge (N_{AF2} = 674, N_{NMR} = 546), where N represents number of secondary structure elements of varying lengths for AlphaFold2 (AF2) and NMR models.

Next, the histograms for the full-length AlphaFold2 models in the S^2_{RCI} data set are presented in Supplementary Fig. 33. These histograms show that, for all six secondary structure types except coils and α-helices, AlphaFold2 models generally favor shorter lengths. For the 762 models analyzed, coils vary in length from 1 to 510 residues, while α-helices range from 1 to 204 residues, both demonstrating the highest variability in their sequence lengths Supplementary Table 10). As shown in Supplementary Fig. 33, these coils and helices, which span long sequence lengths, are typically located at the termini as they are truncated in Supplementary Fig. 32 and might also occur in low- and mid-pLDDT regions.



Supplementary Fig. 33: **Sequence length of secondary structure elements in AlphaFold2 models.** The distribution of amino acid (AA) sequence lengths for six secondary structure elements is shown for AlphaFold2 models across 762 proteins for S^2_{RCI} data set. The subplots are shown as A) coil (N = 19,412), B) strand (N = 9,585), C) α-helix (7,604), D) turn (N = 15,124), E) 3_{10}-helix (N = 2,456), and F) bridge (N=2,424), where N represents number of secondary structure elements of varying lengths for full-length AlphaFold2 (AF2) models.

Supplementary Table 10: **Sequence length of secondary structure elements**. The table reports the total number, minimum, maximum, mean, and standard deviation for each secondary structure group over 762 full-length AlphaFold2 models of S^2_{RCI} data set.

Secondary structure	**No. of total SS**	**min**	**max**	**mean**	**std**
Coil	19412.00	1.00	510.00	6.91	22.61
Strand	9585.00	2.00	27.00	5.72	2.88
α-helix	7604.00	4.00	204.00	14.84	11.45
Turn	15124.00	1.00	38.00	4.10	2.19
3_{10}-helix	2456.00	2.00	13.00	3.27	0.75
Bridge	2424.00	1.00	2.00	1.01	0.12

List of Publications

1. Dixit, B., Vranken, W., & Ghysels, A. (2023). Conformational dynamics of α-1 acid glycoprotein (AGP) in cancer: A comparative study of glycosylated and unglycosylated AGP. *Proteins: Structure, Function, and Bioinformatics*, prot.26607. https://doi.org/10.1002/prot.26607

2. Gavalda-Garcia, J., Dixit, B., Díaz, A., Ghysels, A., & Vranken, W. (2024). Gradations in protein dynamics captured by experimental NMR are not well represented by AlphaFold2 models and other computational metrics. *Journal of Molecular Biology*, 168900. https://doi.org/10.1016/j.jmb.2024.168900

3. Roca-Martinez, J., Lazar, T., Gavalda-Garcia, J., Bickel, D., Pancsa, R., Dixit, B., . . . & Vranken, W. F. (2022). Challenges in describing the conformation and dynamics of proteins with ambiguous behavior. *Frontiers in Molecular Biosciences*, 9, 959956. https://doi.org/10.3389/fmolb.2022.959956

4. Kagami, L. P., Orlando, G., Raimondi, D., Ancien, F., Dixit, B., Gavaldá-García, J., . . . & Vranken, W. (2021). b2bTools: online predictions for protein biophysical features and their conservation. *Nucleic Acids Research*, 49(W1), W52-W59. https://doi.org/10.1093/nar/gkab425