Annex 3

ROMANIAN ACADEMY
School of Advanced Studies of The Romanian Academy
Institute of Biochemistry

PhD THESIS SUMMARY

BIOINFORMATICS AND BIOCOMPUTING TECHNIQUES FOR LIGAND INTERACTIONS AND BIOACTIVE COMPOUND ANALYSIS AND DEVELOPMENT

SCIENTIFIC COORDINATOR
Dr. Andrei-Jose Petrescu

PH.D. CANDIDATE:
Şulea Teodor Asvadur

2025

Contents

PART I – GENERAL INTRODUCTION

CHAPTER 1 – AIM OF THE WORK

Technological developments in biological data generation, as well as larger data storage solutions and faster data management have been a boon of the 21st century (Carugo and Djinović-Carugo, 2023). Researchers have now the daunting task of analyzing this massive amount of data for which advanced methods must be developed (and made as user-friendly as possible) in order to turn data into useful knowledge. Starting with the mid 2010s the new cryo-EM technologies revolutionized the field and allowed structural insights into increasingly complex biomolecular systems; and as interpreting cryo-EM data critically depends on accurate molecular models, the second revolution came about at the beginning of the 2020s in the form of AI-driven automatic modelling platforms – an unprecedented shake up in the field.

Given this context, the purpose of the presented work was:

a) to examine the limits and to improve the capabilities of the new generation of AI technologies in protein structural prediction;

b) to contribute, beyond structure prediction, to the development of new tools for exploring the conformational space of molecular systems – which are relevant in describing biomolecular interactions and processes in normal biological conditions and

c) to develop computational workflows for solving intricate structural problems related to several specific protein-protein and protein-ligand complexes of medical relevance.

The work covering these goals is structured in five chapters grouped in three parts as follows:

The first part consisting of two chapters focuses on in-silico molecular modelling and results obtained on some relevant protein families using the new generation of AI-driven methods in both their automatic and customized flavors. This first part starts with a brief theoretical chapter, overviewing homology modelling techniques, both the "bespoke" and the more recent "neural network"-based ones.

The second chapter of this part is dedicated to a thorough investigation of the capabilities and limits of the new generation AI-driven methods that became commonplace during the last years of my PhD studies and just received the Nobel Prize in 2024. While these methods are known to perform well in simple cases, we tested them in the more challenging case of multistate multidomain proteins taking as an example the coiled-coil family of Nucleotide-binding Oligomerization Domain-like (NOD-like) receptors on which our group (Department of Bioinformatics and Structural Biochemistry – DBSB) has gained a vast expertise over the past 15 years. Our results indicate that by fine-tuning what the neural network input was fed – the AlphaFold platform was able to distinctly model the "active" and "inactive" forms of the protein. This result indicates that our restrictive input method might be useful in multistate multidomain proteins for guiding the automatic modelling workflow toward specific conformations for proteins where multiple solved conformations exist.

While structural predictive modelling is indispensable as a preliminary virtual description of biomolecular systems, this falls short in explaining the interactions and processes taking place in these systems which in natural biological conditions are driven by the laws of thermodynamics. As molecules are in perpetual thermal morphing, developing novel, faster techniques describing their possible states is of great importance. Given this, the second part of my thesis describes our contributions to the development of Robosample a new molecular sampling software platform, based on a previously published method falling under the "Enhanced Sampling" flavor which, given the increasing size of structurally solved systems (Berman, Vallat and Lawson, 2020), have seen more and more use in the last decade (Shen, Zhou and Shi, 2023). The idea of molecular sampling based on robotic algorithms emerged from the fruitful collaboration between Dr. Laurențiu Spiridon from our department (DBSB-IBAR) and Prof. David Minh from the Illinois Institute of Technology and Associate Director at the Center for Interdisciplinary Scientific Computation, collaboration which led to the development of the original method.

The most significant benefit of sampling the conformational space of molecular systems at room temperature is the possibility to precisely estimate the binding free energy of molecular complexes. This is why the final part of the thesis is dedicated to solving three intricate problems relevant in molecular medicine by combining modelling and simulation for calculating the binding free energies of protein-peptide and protein-ligand complexes in medical relevant systems. Along with the modelling and simulation steps, this work relies on a diverse set of free energy computation techniques, ranging from endpoint methods (MM-GBSA) to methods based on alchemical transformations (HRE/MBAR) sampling.

The first chapter of this part is dedicated to the modelling of Thrombopoietin Receptor (TpoR – of unknown structure at that time of publication) and its interaction with mutant calreticulin, an interaction which was shown to be responsible for ~40% of myeloproliferative neoplasms. This work is the result of a longstanding collaboration between DBSB-IBAR and the Cell Signaling group at the Université Catholique de Louvain headed by Prof. Ștefan N. Constantinescu which provided the HDX experimental constraints and validated our structural predictions, predictions which were further recently confirmed by cryo-EM results on the TpoR structure (Tsutsumi *et al.*, 2023a; Sarson-Lawrence *et al.*, 2024).

This is followed by a chapter featuring the use of such free energy computations in immunobiology in a work that has resulted from the collaboration of our DBSB department with the Department of Molecular Cell Biology of IBAR related to the recognition by the T-Cell Receptor (TCR) of an HLA system loaded with tyrosinase YMD epitope (369-YMDGTMSQV-377) which was shown to be relevant in melanoma. In this study we first assessed the formation of HLA-YMD complex and zoomed out afterwards to looking at the ternary complex which forms when the HLA-YMD system interacts with the hyper-variable region of the TCR, alchemical methods were used to sample its conformational space and Multistate Bennett Acceptance Ratio was used to compute the Binding Free Energy.

The final chapter of this part features the use of free energy estimations in pharmacology and is dedicated to a work performed within the framework of the collaboration of our DBSB department with the group of Prof. Bogdan Amuzescu from the Faculty of Biology of the University of Bucharest on the interaction of NaV 1.5 ion channel with cenobamate, a small molecule used as an antiepileptic drug. In this work modelling of the ion channel was performed by Dr. Amuzescu's team, as well as the initial molecular docking of the cenobamate ligand to the ion channel, while we performed the generation of the mutant structures, system construction (i.e. building the lipid bilayer, embedding the different channels into it, parametrizing the system) and the molecular dynamics simulations. Using the generated trajectory, binding free energies were computed between the native/mutant forms of the receptor and the cenobamate molecule using equally-spaced frames from the MD simulation.

PART II – PROTEIN STRUCTURE PREDICTION AND MODELLING

CHAPTER 2 – ASSESSING THE LIMITS OF NOVEL AI BASED STRUCTURAL PREDICTION METHODS

1. Introduction

Structural biology has seen a transformative boost at the dawn of the 2020s with the advent of automatic Deep Learning Modelling techniques such as AlphaFold2 & 3 (AF2/3) (Jumper *et al.*, 2021; Abramson *et al.*, 2024), RoseTTAFold All-Atom (Krishna *et al.*, 2024) (RFAA) or OmegaFold (Wu *et al.*, 2022) (OF) that superseded traditional manual homology modeling workflows. The AF & RFAA methods rely basically on the same workflow as traditional homology modeling steps which include multiple sequence alignment (MSA), searching for templates and mitigating between them in building models of a target sequence based on a massive Neural Network training aimed at optimizing the local structure and amino acid contacts. On the other hand, OmegaFold takes a different approach predicting the structure of a protein directly from its single primary sequence by using a protein language model based on a geometry-inspired transformer. One of the most important advantages of all these methods is the speed in proposing structural models of a given sequence. In matter of hours for AF & RFAA or even minutes for OF anyone can get predictive models of a protein sequence. Many reviews highlight also the high accuracy of such automatic models especially in the case of single domain targets which are highly homologous to existing experimentally solved structures. On the other hand, such automatic procedures return blunt, black-box models and deny any human intervention driven by extra experimental constraints or information, or by the flair and gained experience of a researcher which, especially in remote homology cases, might be relevant.

By contrast, traditional manual homology modelling is far more tedious and may result in less optimized structures but has the advantage of being far more flexible allowing researchers on one hand to take into account experimental constraints or any other experimental observations and on the other hand to refine models by experimental hypothesis testing in a trial-and-error manner, and therefore driven by a heuristic approach. Thus, heuristic modeling is prone to better tackle remote homology targets or more complex systems such as state dependent protein structures, multidomain proteins or hetero-molecular systems.

In this context, given that over ~60% of known proteomes comprise complex protein sequences exhibiting more than one structural domains it is highly relevant to determine how well the novel AI driven structural prediction platforms preform on such systems and better understand their limitations in generating overall plausible structural models and find ways to increase their prediction capabilities especially when such proteins were shown to adopt multiple conformations along their functional cycle. In other words, it is highly instructive to understand how automatic predictors work in cases in which in structural databases the same protein was solved in multiple structural states and modeling faces a multivalued sequence to structure (1D→3D) mapping problem.

A good case study to assess the performance of the novel AI platforms on more complex protein systems, on which heuristic modelling was shown to be effective, is that of the multi-state, multi-domain protein family of CNL NOD-Like (Coiled-Coil Nucleotide-Binding Oligomerization Domain-Like) receptors. Basically, these proteins consist of three canonical domains with casual N-terminal or C-terminal extensions:

(1) a CC (Coiled Coil) "connector"

(2) an NBD (Nucleotide Binding Domain) "switch"

(3) an LRR (Leucine-Rich Repeat) "sensor"

In turn, upon activation (induced by a pathogen molecular effector) the NBD "switch" suffers an internal conformational transition in which the Arc2 subdomain rotates 180^{o} around the NBS-Arc1 region by releasing the ADP cofactor (specific to the inactive state) and exchanging it with an ATP that stabilizes the active conformation.

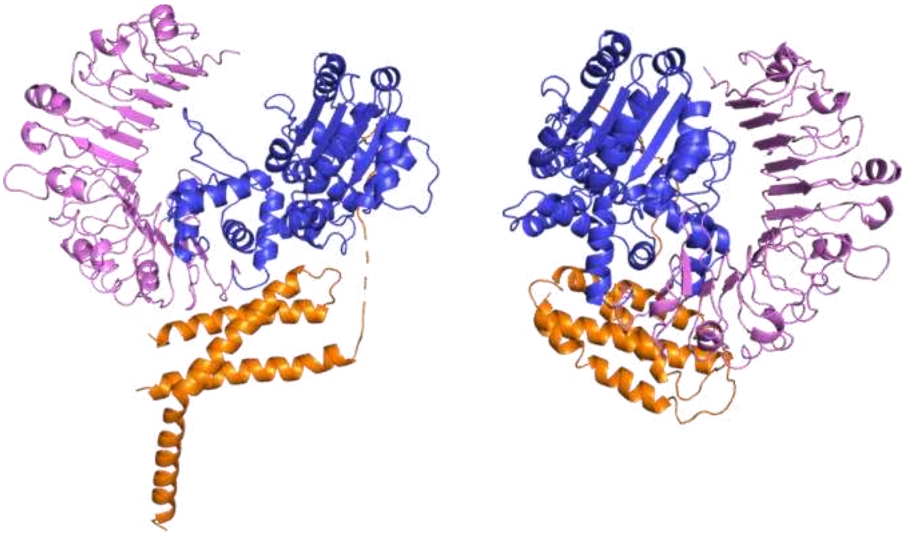

Figure 2-1. Active (6J5T) and inactive (6J5W) conformations of the ZAR1 protein, aligned using the NBD domain.

2. Extended A. Thaliana family

In order to test the performance of AI-driven modelling, we selected a set of representative *A. Thaliana* CNL proteins. We retrieved a set of 1257 sequences from the NLRScape (Martin *et al.*, 2023) web server, which we filtered by domain organization, discarding any that did not contain all 9 CNL motifs. The resulting set was clustered using MMSeqs2 at 70% coverage and 70% identity, resulting in 36 groups, of which 4 were eliminated.

The 32 representative sequences were then modelled using locally installed AlphaFold2 (AF2) and OmegaFold (OF); and over the web using AlphaFold3 (AF3) and RoseTTAFold All-Atom (RFAA) via the NeuroSnap servers. Additionally, models of these sequences were retrieved for comparison from the AlphaFold Database (AFDB) (Varadi *et al.*, 2024). Given that novel AF3 and RFAA web implementations allow protein modeling along with their ligands we used them to model the 32 CNL both bare and in interaction with ADP and ATP. All structures were minimized using OpenMM (Eastman *et al.*, 2024). Model quality was asses using MolProbity (Williams *et al.*, 2018). The RFAA models were very low quality and simulated annealing was unsuccessfully used to attempt to improve them. These were stable and were confined in a conformational pool of no more than 2Å.

3. ZAR1 – Comparison of models to solved structure

Since ZAR1's structure has been solved, it offers us the ability to directly evaluate the performance of the above software. The CC domain has been solved in both "active" and "inactive" conformations. The two structures present different CC domains, as can be seen in Figure 2-2. As such, default AF2, AF3, RoseTTAFold All-Atom and OmegaFold generated a structure that combines features from both of them into an implausible model.

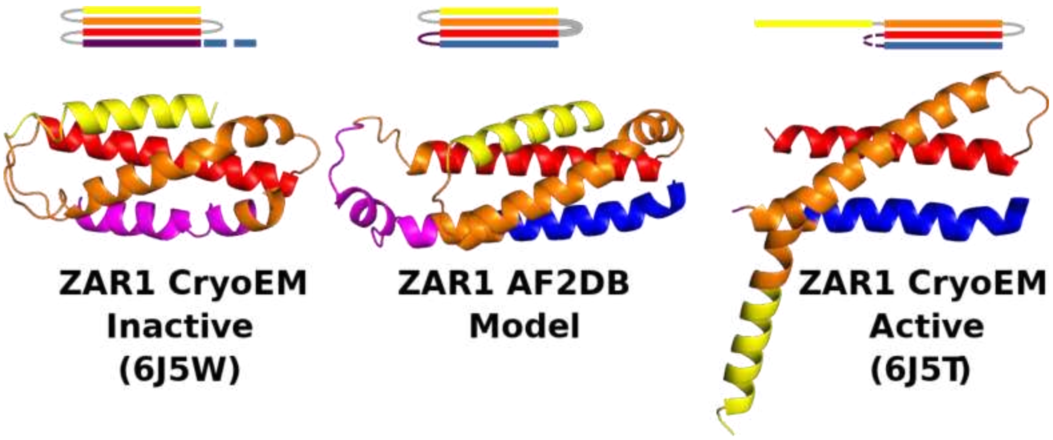

Figure 2-2. Comparison between active, inactive Cryo-Electron Microscopy solved structures and model retrieved from the AlphaFold2 Database

Model Name	CC	NBS-Arc1	ARC2	LRR
AF2 – Database	14.46	1.86	1.63	1.45
AF2 – Active Control	1.39	0.49	0.42	0.53
AF2 – Inactive Control	18.51	1.86	1.51	0.84
AF2 – Active MSA	1.55	0.59	0.61	0.96
AF2 – Inactive MSA	14.40	1.79	1.53	0.97
AF3 – ADP	14.41	1.80	1.64	1.42
AF3 – ATP	14.39	1.89	1.58	1.33
AF3 – No Ligand	14.45	1.92	1.59	1.51
RFAA – ADP	14.42	2.04	1.64	1.86
RFAA – ATP	14.50	2.05	1.78	2.02
RFAA – No Ligand	14.45	1.92	1.65	1.96
OmegaFold	14.48	1.87	1.33	3.11

Table 2-1. Per domain RMSD(Å) for the highest ranked model of ZAR1, using the "active" state as reference.

Model Name	CC	NBS-Arc1	Arc2	LRR
AF2 – Database	12.42	1.33	0.88	1.34
AF2 – Active Control	19.19	1.89	1.45	0.78
AF2 – Inactive Control	0.80	0.69	0.35	0.47
AF2 – Active MSA	18.48	1.88	1.51	0.90
AF2 – Inactive MSA	12.54	0.86	0.44	0.75
AF3 – ADP	11.99	1.31	0.98	1.22
AF3 – ATP	12.11	1.35	0.87	1.16
AF3 – No Ligand	12.08	1.26	0.93	1.37
RFAA – ADP	12.74	1.41	1.02	1.77
RFAA – ATP	12.65	1.53	1.04	1.92
RFAA – No Ligand	12.36	1.42	1.11	1.89
OmegaFold	12.99	1.24	0.87	3.04

Table 2-2 Per domain RMSD(Å) for the highest ranked model of ZAR1, using the "inactive" state as reference

Model Name	RMSD vs active	RMSD vs inactive
AF2 – Database	22.158	6.004
AF2 – Active Control	0.832	23.036
AF2 – Inactive Control	22.612	0.675
AF2 – Active MSA	1.271	22.873
AF2 – Inactive MSA	21.997	5.837
AF3 – ADP	22.119	5.785
AF3 – ATP	22.076	5.879
AF3 – No Ligand	22.191	5.889
RFAA – ADP	22.406	6.756
RFAA – ATP	22.646	6.792
RFAA – No Ligand	22.424	6.027
OmegaFold	22.374	6.442

Table 2-3. Global RMSD(Å) for the highest ranked models for each generation method, relative to the active/inactive crystal structure

The models' LRR domains show an accurate beta-sheet network, but a higher propensity to form compact helical structures than the solved crystal structures.

When comparing Global RMSD values, it clearly appears that the AI models have a tendency to model CNLs in the "inactive" state which is more compact. This might drive the network towards the form that minimizes the exposed surface and is more energetically favorable. Additionally, the models present a propensity to over compact the structure, judging from the tighter binding between the domains, as measured using the Prodigy software (Vangone and A. M. Bonvin, 2015; Xue *et al.*, 2016)

4. Analysis of the extended CNL set models

The other CNL sequences were modelled using the same procedure described for ZAR1 above. Since the system has multiple possible states, it was important to study the conformation each software generates for each sequence.

We mapped the position of the CC, Arc2 and LRR domains' center of mass and the plane formed by the "VG" motif (the start of the NBS domain), the NBS' and the Arc1's centers of mass (origin plane). Using this 3D representation, we can successfully distinguish the "active" from the "inactive" conformations. This distribution is presented in Figure 2-3, for a subset of models.

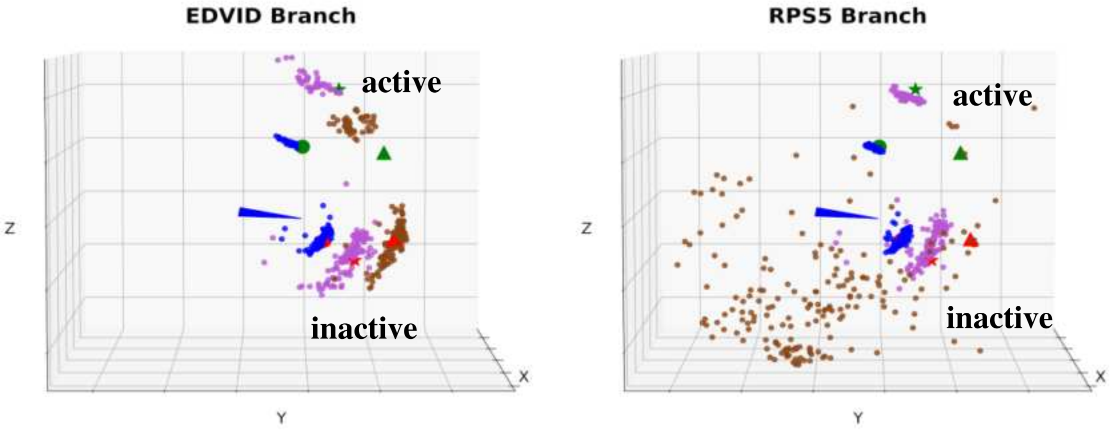

Figure 2-3. Distribution of the CC (brown dots), Arc2 (blue dots) and LRR (brown dots) domains relative to the same domains of the experimental structures (green triangle, circle and star are active CC, ARC2 and LRR respectively; red triangle, circle and star are inactive CC, ARC2 and LRR respectively)

As presented, all models on the "EDVID" branch adopt either an "inactive" or an "active" conformation, with a strong bias towards the compact "inactive" model. Also, modelling in the presence of ATP does not generate the "active" conformation, indicating that ligand specificity is not taken into account by either AF3 or RFAA. In contrast, the "RPS5" branch models display a very different and much more scattered distribution than the "EDVID" branch. A detailed analysis of this branch reveals that the CC does not interface with the NBD or LRR domains, pointing to a different signaling mechanism. For CNL proteins from other families (*Solanum tuberosum*, *Triticum aestivum* and *Hordeum vulgare),* the bias for the ”inactive” form holds.

5. Correcting output prediction by input filtering in multistate situations

In order to drive the network towards a plausible conformation and away from the “chimeric” structure of the CC presented above, we enriched/restricted the input data that AlphaFold2 draws features from. By firstly adding more recent solved structures of Nod-like receptor proteins from the PDB and classifying them as either “active” or “inactive”, we were able to drive the network towards one of these conformations. Secondly, given that both the CC and the LRR domains have well-defined architectures – CATH (Waman *et al.*, 2024) designations 1.20 Up-Down Bundle and 3.80.10 Leucine-Rich Repeat respectively – the AF2 PDB snapshot was filtered to contain only proteins containing at least one of these domains, with only this reduced set being used. Further, given that the MSA generated should reflect features specific to NLR proteins, only NLR sequences were used as input.

Regarding the ZAR1 sequence, models generated with these input filtering methods in place are much closer to the crystalized structures, as measured by RMSD (Table 2-1 and Table 2-2): the CC domain is modelled after the Active conformation in both of the “active” sets, while the “inactive” form is properly modelled when no MSA is fed into the network (Inactive Control) – as reflected in the global RMSD values (Table 2-3). However, all other domains have lower RMSD to their respective experimentally solved structures than any of the other generated models.

The extended CNL set also benefited from this input filtering, as distinct “active” and “inactive” models were built for all sequences.

6. Conclusions

At the domain level, the folded structures of the NBD-Arc1, Arc2, and LRR modules, which belong to the CATH 3.40.50.300, 1.10.533.10, and 3.80.10 topologies respectively, have been predicted with an accuracy of less than 2 Å by all evaluated platforms. However, in the CC region, the RMSD from the experimental structure is greater than 12Å. Analysis of the solutions provided by AI platforms for this region suggests that they combine structural information from both the ADP-Inactive and ATP-Active structures of ZAR1.

At the global level, it has been shown that AI driven methods favor the more compact, inactive state. Only by selectively filtering the input with specific structural data, could models for the active state of all the CNL proteins be generated. Notably, the inclusion (or exclusion) of the ligand did not bias the network in favor of any conformation, which indicates that the AF3 and RFAA pipelines only account for ligand position after modelling the surrounding receptor, thereby ignoring any “induced fit” effect. Incorporating ligand data into these pipelines represents an important step in modelling proteins with multiple meta-stable states.

The work described here outlines a protocol for driving fast AI-based modelling towards desired conformations, without the computational cost of retraining a neural network. Speaking of retraining, it is important to note that AlphaFold (or rather DeepMind) does not provide training scripts or training data, making generation of specifically-trained networks impossible. On this note, there have been efforts in opening up these kinds of software, such as the OpenFold (Ahdritz *et al.*, 2024) software, which allows for retraining and produces results on par with AlphaFold2. Additionally, it provides a computationally efficient method of generating proteins in multimer conformation (such as the “active” CNL conformation), without the computational cost of running AlphaFold2-Multimer, which for modelling a pentameric structure would be prohibitively high.

PART III – BEYOND STRUCTURE PREDICTION - CONFORMATIONAL SAMPLING BY MOLECULAR SIMULATION

CHAPTER 3 – IMPROVED SAMPLING VIA USE OF GIBBS SAMPLING ON INTERNAL DOFS

1. Materials and methods

The software described in this part is the Robosample software , developed by Spiridon et al (Spiridon *et al.*, 2020). It applies high-performance constrained dynamics algorithms to the simulation of biologically relevant systems. It applies a mix of Blocked Gibbs sampling with HMC – Constrained Dynamics Hamiltonian Monte Carlo (CDHMC) - to efficiently sample conformational space.

It uses internal Bond/Angle/Torsion (BAT) coordinates to sample the conformational spaces of complex biological systems much more efficiently than traditional MD simulation. The original Constrained Dynamics Hamiltonian Monte Carlo (CDHMC) paper (Spiridon and Minh, 2017) outlines how the use of Gibbs sampling together with BAT coordinates allows for the correct recovery of the Boltzmann Distribution.

In Robosample, atoms or groups of atoms are grouped together to form ”rigid bodies”, meaning that none of their internal coordinates are allowed to vary. Different rigid bodies are bound together via different joints (such as Pin, Spherical or Cartesian). In this way, a robot graph (rigid bodies and different types of joints) is overlayed on top of the chemical graph (atoms and bonds). By using different combinations of rigid body definitions and joint types, different subsets of the conformational space are sampled. By sampling correlated degrees of freedom together, a more efficient exploration of the conformational space can be achieved.

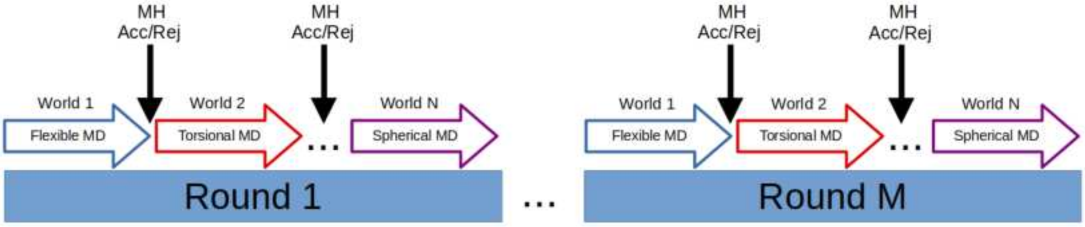

Figure 3-1. Schematic representation of a Robosample simulation, consisting of M rounds, each with N worlds

To evaluate the effectiveness of Robosample, we conducted simulations on two molecular systems: alanine dipeptide and a more complex model featuring the first glycan of the hepatitis C virus E2 protein (E2N1). In this model, the glycan is attached to an E2-derived [-12 + 6] peptide at the ASN17 site. This glycan is believed to play a role in shielding the virus from the host immune system by either delaying or reducing recognition by anti-E2 antibodies (Prentoe *et al.*, 2019).

The alanine dipeptide system was simulated in multiple blocking schemes:

- All rotatable bonds made flexible, with either Pin (TD), Cylinder (CYL) or Ball joints – “All Flexible”
- Only the N-Cα and Cα-C bonds made flexible, with either Pin (TD), Cylinder (CYL) or Ball joints – “Ramachandran Dynamics”

In order to ensure ergodicity, the above six worlds were simulated together with a cartesian world.

2. Results and discussions – Alanine Dipeptide

Despite being small compared to other biomolecules, alanine dipeptide (N-acetylalanine-N-methylamide) has a highly complex and frustrated potential energy surface (PES), making it a challenging system for molecular sampling techniques. Mapping its configuration space onto the φ and ψ dihedral angles helps retain the key maxima and minima of the PES while reducing complexity. As a result, the potential of mean force in the φ/ψ domain is commonly used to evaluate the accuracy and efficiency of sampling methods (Montgomery Pettitt and Karplus, 1985). All simulations were able to recover the free energy surface of the ALA2 system.

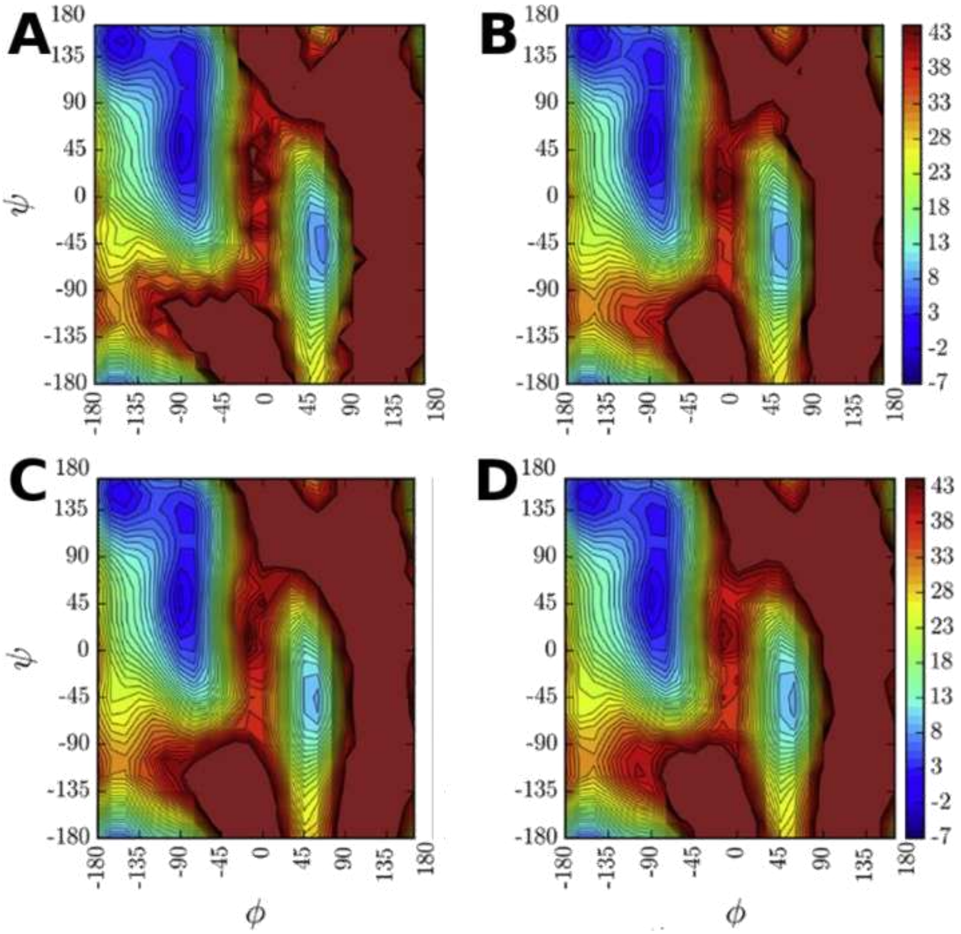

Figure 3-2 Free energy surfaces of ALA2, based on four types of HMC simulations: A - Fully Flexible, B - Mixed-RamaTD, C - Mixed-RamaBall, D - Mixed-RamaCYL

The fact that the free energy surfaces are consistent among themselves shows that CDHMC simulations can recover free energy surfaces using either Pin, Cylindrical or Spherical joints, independent of the size of the bodies.

The Mean First Passage Times between free energy basins provide insight into the relative efficiency of the simulations done. As can be seen in Table 3-1, the fully flexible simulation has the slowest overall transitions. The "All-Bonds" simulations perform better than the fully flexible, with the TD regimen performing best. Ramachandran dynamics outperforms the all-bonds dynamics.

From\To	C5	PPII	C7$_{eq}$	α$_L$
Fully Flexible				
C5	3.2 ± 0.2	9.1 ± 0.5	14.3 ± 1.7	5380.0 ± 285.0
PPII	9.6 ± 0.9	5.7 ± 0.4	10.9 ± 1.8	5380.0 ± 285.0
C7$_{eq}$	16.5 ± 1.1	12.6 ± 0.6	2.8 ± 0.2	5370.0 ± 285.0
α$_L$	502.0 ± 195.0	499.0 ± 194.0	491.0 ± 194.0	14.1 ± 4.4
Mixed-RamaTD				
C5	1.8 ± 0.03	4.0 ± 0.1	2.8 ± 0.1	631.0 ± 92.2
PPII	2.5 ± 0.1	3.2 ± 0.1	2.6 ± 0.1	631.0 ± 92.2
C7$_{eq}$	3.2 ± 0.2	4.7 ± 0.2	1.5 ± 0.02	630.0 ± 92.2
α$_L$	50.9 ± 1.1	52.4 ± 1.2	49.4 ± 1.1	8.3 ± 1.3
Mixed-RamaCYL				
C5	2.3 ± 0.03	5.2 ± 0.1	3.5 ± 0.02	615.0 ± 33.6
PPII	3.2 ± 0.1	4.1 ± 0.1	3.3 ± 0.005	615.0 ± 33.4
C7$_{eq}$	4.1 ± 0.1	5.9 ± 0.1	1.9 ± 0.01	614.0 ± 33.6
α$_L$	46.4 ± 4.5	48.4 ± 4.6	44.7 ± 4.7	12.0 ± 1.4
Mixed-RamaBall				
C5	2.4 ± 0.01	5.2 ± 0.01	3.1 ± 0.1	518.0 ± 18.6
PPII	3.1 ± 0.01	4.1 ± 0.04	3.0 ± 0.1	518.0 ± 18.5
C7$_{eq}$	3.7 ± 0.01	5.6 ± 0.02	1.9 ± 0.02	518.0 ± 18.4
α$_L$	37.0 ± 1.4	39.1 ± 1.5	35.8 ± 1.6	12.8 ± 1.0

Table 3-1 Average Mean First Passage Times (MFPT) of Fully Flexible and Ramachandran simulations, expressed in molecular dynamics steps. Note: in the original paper, these were divided by 200 to be comparable to the paper originally describing the CDHMC method.

3. Results and discussions – E2N1 Glycoprotein

Figure 3-3 shows that in the same amount of time, mean value and fluctuations of the RMSD are much larger for the CDHMC simulations than for the fully flexible HMC simulations. An overlap of a subset of frames shows a clearer difference between the two methods (Figure 3-4).

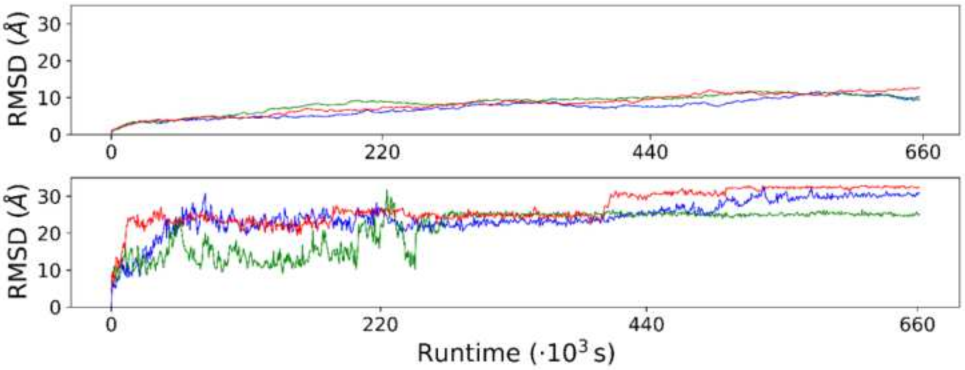

Figure 3-3 RMSD Timeseries for MD (Top) and HMC (Bottom) simulations

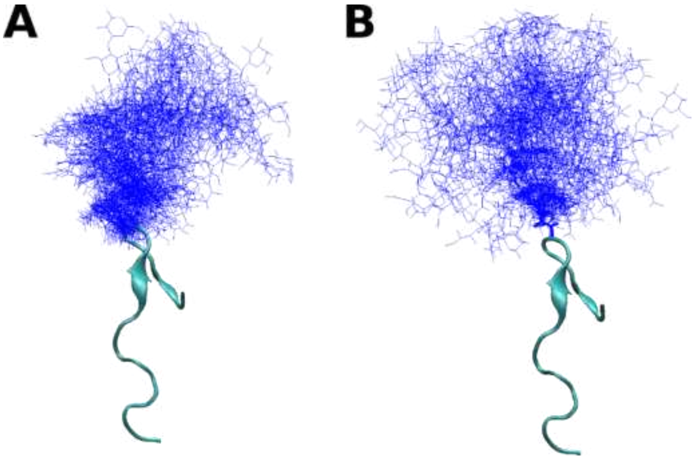

Figure 3-4 Overlapped frames from MD (A) and HMC (B) trajectories. For clarity, only the polysaccharide coordinates have been included from all frames, with the polypeptide kept from the first frame

4. Conclusions

The results presented in this work suggest that Robosample provides a robust framework for performing constrained molecular simulations while ensuring ergodicity through Gibbs sampling. Beyond the traditional torsional and angle/torsion mobilities commonly used in constrained simulations (Vaidehi and Jain, 2015), Robosample expands the range of motion by incorporating additional robotic mobility types, leveraging the mechanical joints available in Simbody. This enables the exploration of arbitrary degree-of-freedom (DOF) couplings, including weak bond/torsion interactions found in cylindrical joints.

It has been shown that through parameter optimization and different joint types, the software is able to reproduce the ALA2 free energy surface in a more efficient way than is possible via classical MD simulation. Additionally, Ramachandran dynamics using spherical joints, which have not been used in other internal DoF MD software, have the shortest MFPT for the rarest transition.

Using an actual system (the E2N1 glycoprotein), we have shown that exploration of the conformational space of highly flexible structures can be more efficiently sampled if Robosample is used. The fact that using alternating constraints speeds up space exploration means that there is, in theory, an optimal combination of Gibbs blocks which may explore any space faster than simple MD. Currently, we are exploring the use of Replica Exchange in tandem with the HMC algorithm, in order to overcome larger potential energy barriers.

PART IV – USING MODELING AND SIMULATION FOR INVESTIGATING COMPLEX MOLECULAR INTERACTIONS

CHAPTER 4 – MODELLING THROMBOPOIETIN RECEPTOR COMPLEXATION BY MUTANT CALRETICULIN

1. Introduction

In the current chapter, we employed an integrative strategy to dissect the molecular mechanism by which del52/ins5 calreticulin mutants (mCALR) specifically and persistently binds to TpoR and to shed light on how this interaction drives TpoR dimerization and activation. Understanding how frameshift mutations in a key chaperone result in novel binding capabilities is crucial both from a conceptual standpoint and for therapeutic development. In order to reveal the molecular basis of the formation of a permanent TpoR-mCALR complex we proceed to accurately modelling the two players in an effort to locate putative hotspots responsible for their interaction as presented below.

2. TpoR Homology Modelling

In the first stage, the TpoR sequence was analyzed in order to generate multiple secondary structure profiles, solvent accessibility profiles and intrinsic disorder profiles. These profiles were used to refine the alignment between the target sequence and the templates used.

The next stage consisted of searching homologous structures to be used in homology modelling. Thus, the TpoR sequence was subject to searches by molecular threading, using PHYRE2 (Kelley *et al.*, 2016). The first result (RCSB Code: 1ERN (Livnah *et al.*, 1999)) is a crystal of the Erythropoietin Receptor (EpoR). 20 EpoR structures were gathered from the RCSB database, and the 1CN4 (Syed *et al.*, 1998) structure was chosen to be used as template for the membrane-distal Cytokine Receptor Module (CRM). It presents around 25% identity to the TpoR sequence. Even though the target sequence can be modelled via homology modelling, using EpoR, it displays a large, 65 amino acid-long insertion in its 2nd domain, for which no homologous structures were found. Interestingly, automatic models generated with the state-of-the-art platform AlphaFold2 were unrealistic. The solution proposed for this 65aa insertion was a random coil wobbling apart from the main folded body of TpoR, being completely accessible to the solvent despite its sequence secondary structure and accessibility propensities.

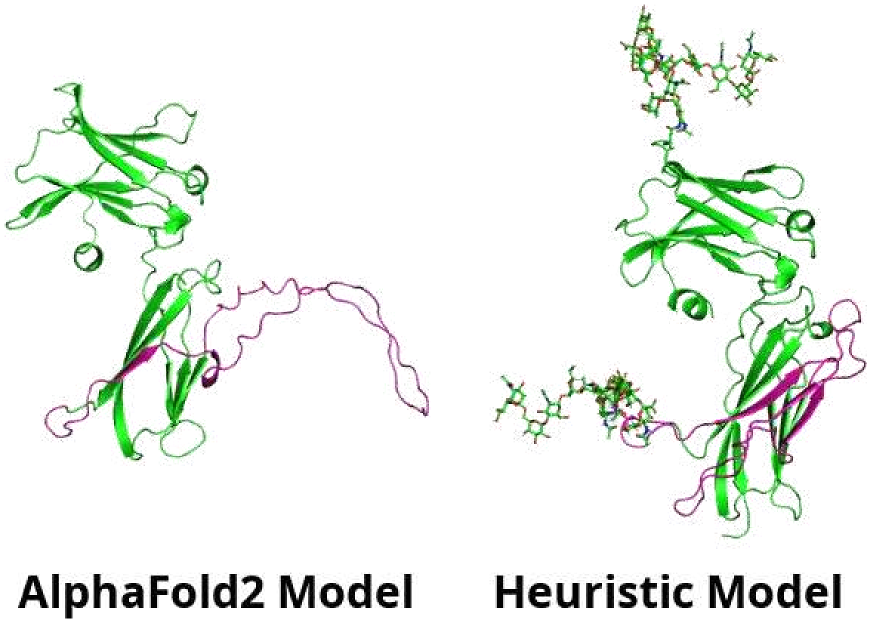

Figure 4-1 Comparison between AlphaFold2 model and our heuristic model. The purple part represents the 65aa insertion. The glycans are represented using sticks.

Hence, in order to more realistically model this insertion, searches for more appropriate molecular architectures were employed. The reference structure, EpoR, displays a β-sandwich-type architecture, in which one of the faces has 3 extended structures and the other has 4 (β-sandwich 3/4). This architecture is part of the CATH 2.60.40.10 superfamily, which also includes β-sandwich 4/5 architectures. The assumption that TpoR adopts such an architecture started from the observation that 2 extended structures were predicted in the large insertion of the TpoR sequence. Thus, a 4/5 β-sandwich architecture was proposed for TpoR. The immunoglobulin group of the 2.60.40.10 superfamily is known for adopting this architecture, with the same topology as the one proposed. The CATH representative for the immunoglobulin group is the 2E8 antibody for the LDL receptor, RCSB Code: 12E8 (Trakhanov *et al.*, 1999).

The membrane-proximal CRM of the TpoR extra-cellular domain (ECD) was modelled using the AlphaFold2 software, which was run on our local computational cluster. The generated model was attached to the CRM 1 model, using MODELLER v.9.12 (Webb and Sali, 2018). 4 cysteine sulfide bonds were added based on distance between neighboring cysteine residues.

The transmembrane domain (TMD) was modelled after the E chain of the crystal structure with the PDB ID: 6S1K (Cassidy *et al.*, 2020). This crystal structure was found by using BLAST on the sequence of TpoR's TMD. This was done in order to generate a realistic helical structure. The TMD was attached to the ECD using MODELLER. The TMD's tilt angle, relative to the hypothetical membrane it's supposed to traverse was chosen in order to be consistent with experimental data provided (i.e. the distance between adjacent L508 residues should be at most 6Å) and in order to form a cross-shaped dimer similar to the one found by (Defour *et al.*, 2013).

The CALR del52/ins5 monomers were generated using AlphaFold2, and the CALR dimer was generated using RosettaDock (Lyskov and Gray, 2008).

3. Assessing TpoR-CRM1 interaction with mutant Calreticulin.

Sequence analysis of the TpoR D1D2 shows an excess of negative charges. Conversely, the C-terminus of CALR del52 is strongly positive. As such we posited that the interaction detected between CALR and TpoR could be electrostatic in nature.

Through an extensive workflow that involved multiple molecular docking runs, molecular dynamics and simulated annealing simulations, 3 binding poses were determined for the TpoR/mCALR C-terminus dimer. Through the use of Prodigy (Vangone and A. M. J. J. Bonvin, 2015; Xue *et al.*, 2016) and MM-GBSA (Kollman *et al.*, 2000), these were ranked and the highest ranking one was used to generate the complete tetramer model, which was further embedded into a POPC membrane and fully solvated.

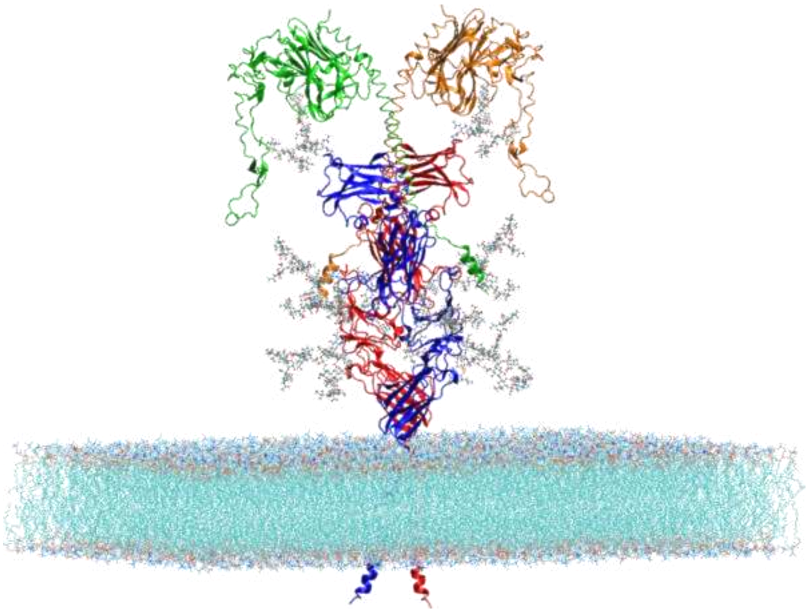

Figure 4-2 Fully solvated TpoR – mCALR system, composed of around 1.5 million atoms.

The CALR ins5 mutant was generated following the same procedure.

The system's stability was tested via a 100ns MD simulation. The number of contacts between the TpoR glycan and the CALR remained consistent.

4. Conclusions

The work presented reveals mechanistic insights into the recognition and subsequent activation of the Thrombopoietin Receptor by pathogenic mutants of calreticulin, triggering myeloproliferative neoplasm. The structural model we created shows that the binding happens on two distinct sites: a physiological CALR/N-glycan binding, which is not specific to the TpoR/CALR pair, and an electrostatic-driven interaction, which is much more stable and leads to blood pathologies. The activation mode we propose can be used in further studies in the development of targeted therapies, targeting, for instance, the N-terminus of the CALR molecule.

The method we employed in discovering the conformation of the TpoR/CALR del52 complex highlights an interesting "anaconda" effect of the CALR molecule. The C-terminus, positively charged as it is, performs a slither-like motion around the membrane distal CRM of the TpoR, until it gets "stuck" on one of the two acidic patches. It is tempting to posit that this way of binding is not unique to this protein pair and may be used as a starting assumption when studying other protein/protein interactions with biological significance.

As a final remark we would like to mention that an experimental cryo-EM structure of TpoR released several months after the publication of our tetramer 2x(TpoR-mCALR) model (Tsutsumi *et al.*, 2023b) confirmed the predicted TpoR structure included in our more complex overall heteromolecular glycoproteic model, as can be seen from Figure 4-3. Moreover, the cryo-EM structure lacks the glycan moiety which – based on MS experimental data – is included in our model. Even more important – the protein core of TpoR in our *heuristic* model matches the cryo-EM structure better than the *automatic* TpoR structural solution proposed at that time by AlphaFold2 as discussed previously (Figure 4-1).

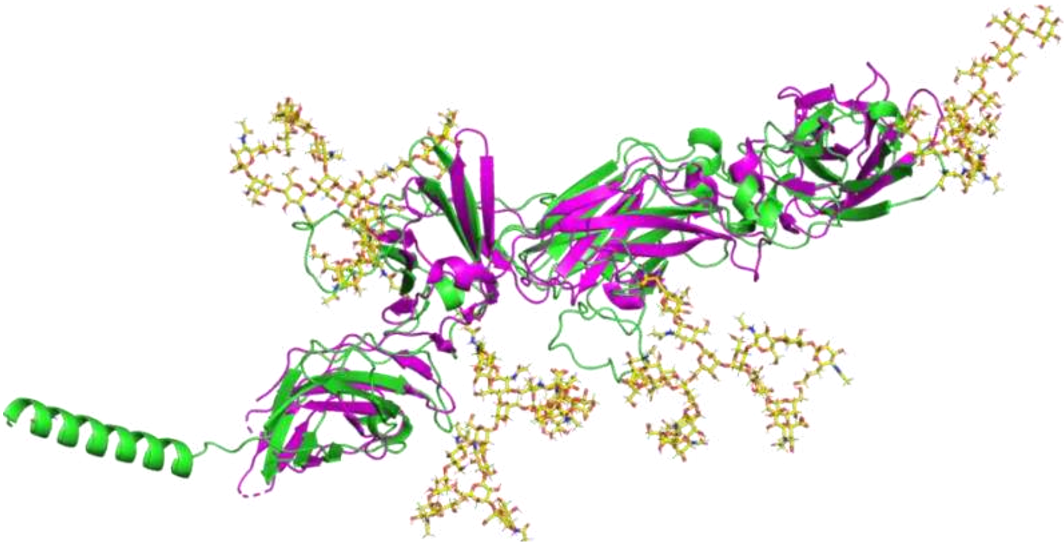

Figure 4-3 Structure alignment between generated model (green and yellow) and solved structure (magenta – RCSB Code: 8G04) of TpoR ECD.

CHAPTER 5 – OXIDATION OF TYROSINASE EPITOPE AND ITS EFFECT ON T CELL REACTIVITY IN MELANOMA

1. Introduction

Intricate bioinformatics and biocomputing workflows can also be used in endeavors related to an in depth understanding of protein – ligand interactions when the ligands are complex flexible peptides.

This chapter presents our endeavor on unraveling the molecular mechanisms leading to the increase in binding free energy of the ternary immunity complex HLA-YMD-TCR upon YMD oxidation - were YMD is the peptide 369-YMDGTMSQV-377 derived from tyrosinase. This presentation is headed by a brief introduction on the three molecular players and the context of this study.

The HLA, MHC I proteins are heterodimers, composed of two protein chains, α and β-microglobulin. The α chain consists of three modules, each playing a distinct role in its interaction with special receptors found on the surface of the T cells. The α3 chain is the only transmembrane chain in the MHC I structure, and it is also responsible for the noncovalent binding to the β-microglobulin chain. In addition, the α3 region interacts with the CD8 receptor found on the surface of T cells. This interaction temporarily blocks the MHC complex while the T cell checks the antigenicity of the peptide bound to the groove that opens between the α1 and α2 domains (Hewitt, 2003).

The T-Cell Receptor (TCR) is a heterodimer composed of two chains, α and β. In very rare cases, it can consist of γ and δ chains instead. The α and β chains each have two distinct regions:

- A constant region, located closer to the T cell membrane (but still extracellular)
- A variable region, which contains three hyper-variable segments that determine peptide complementarity (CDR) and are responsible for binding the antigenic peptide.

2. Modeling of ternary complex, parameter generation for YM_2M_6D, HREX

Broadly speaking, the computational work was done in three stages:

- Modelling the HLA-A02:01/YMD binary system, in both native (WT) and $M2_{SO}M6_{SO}$ forms
- Modelling the HLA-A02:01/YMD/TCR ternary system, in both forms
- Computing the binding free energy of the YMD peptide to the ternary complex.

For the HLA/YMD complex the selection process identified **YL**SPI**A**SPL (Y9L) and M**L**IYS**M**WGK (A14) as the best-fitting templates.

- Y9L (PDB: 5F9J) is naturally bound to HLA-A*02:01, making it a strong match for the experimental haplotype. Additionally, it aligns well with the hydrophobicity and volumetric profiles of YMD and features a tyrosine at position 1, which engages in a stacking interaction with HLA, just like YMD.
- A14 (PDB: 4N8V) originates from the Virion membrane protein A14 and was chosen due to the presence of methionine at position 6, mirroring YMD. However, this structure is associated with HLA-A*11, which differs from the experimental haplotype.

Both templates contain hydrophobic residues at positions 2 and 6, which are involved in oxidation and interact with HLA. However, their spatial arrangements within the HLA binding groove are distinct:

- In Y9L, the side chain at position 6 points directly toward the groove.
- In A14, the side chain at position 6 is oriented parallel to the surface.

These differences provide an opportunity to assess how initial binding configurations influence free energy estimates.

Given that from all the possible oxidized forms of the YMD peptide, it was shown through HPLC experiments that $M2_{SO}M6_{SO}$ was the predominant form, four models of the binary complex were built:

- HLA/YMD Native with M6 pointing towards the HLA groove/towards the solvent
- HLA/YMD oxidized, with the M6 pointing towards the HLA groove/towards the solvent

3. Modelling the HLA/YMD/TCR ternary complex

The construction of ternary HLA-YMD-TCR complex models was carried out in two sequential steps:

- TCR Assembly: Models incorporating patient-specific clone 3 CDR loops (TCRc3) were generated using a joint fragment-based homology modeling approach.
- Docking: The TCRc3 model was docked onto the four HP binary complexes created in the initial phase.

To construct the TCRc3 model, the HP/HPT-DB database was screened for structurally similar candidates corresponding to the patient-specific clone 3 CDR regions. Sequence analysis revealed that different CDR loops and constant regions had varying best matches in terms of sequence homology. Consequently, a joint fragment-based homology modeling strategy was implemented, following methodologies outlined previously (Slootweg *et al.*, 2013; Rajaraman *et al.*, 2016).

Four reference structures (PDB IDs: 5HHM, 3PWP, 2PYE, and 2BNQ) were selected:

- 5HHM and 3PWP were used to construct the α- and β-chain conserved scaffolds respectively.
- 2PYE and 2BNQ served as templates for assembling the CDR loops of TCRc3

All HLA-A*02:01 and TCRc3 models were generated using MODELLER v9.21. In brief, variable loops bridging the sequence-conserved regions (SCRs) were iteratively optimized via MODELLER’s loop refinement protocol. Sequence-variable regions were subjected to repeated simulated annealing and energy minimization and final models were validated using the MolProbity server, yielding scores of 0.59 for HLA and 1.00 for TCR, confirming their structural quality.

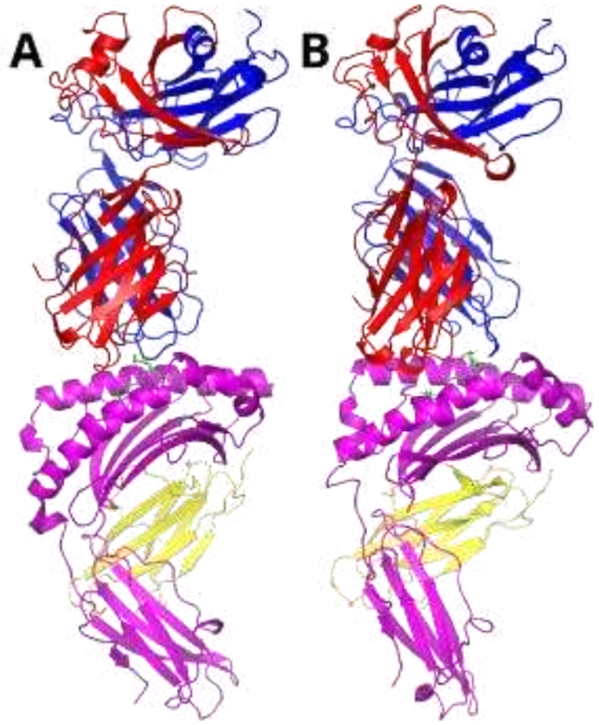

Figure 5-1 Generated HLA/YMD/TCR models, with M6 pointing towards the TCR (↑ configuration); A - Native; B – Oxidized

4. Binding free energy calculation

In this phase, Absolute Binding Free Energy estimates were obtained using Hamiltonian Replica Exchange (HRE), where thermodynamic states were progressively scaled between a fully coupled and a fully decoupled ligand state. This process involves an alchemical transformation, in which the interactions exerted by HLA and TCR on YMD are gradually weakened until they are completely eliminated. The two HLA/YMD/TCR complex models (with wildtype YMD and oxidized YMD) were subjected to a total of 472.5 ns of simulation, across 135 replicas. A replica exchange would be attempted every 10^{-3} ns. The fully coupled trajectories were further used for differential structural analysis, including stability analysis, cluster count analysis and solvent accessibility analysis.

The binding free energy calculations were performed using YANK (Wang *et al.*, 2013), leveraging Hamiltonian Replica Exchange (HRE) for enhanced sampling. The Multistate Bennett Acceptance Ratio (MBAR), implemented via the PyMBAR (Shirts and Chodera, 2008) package, was used to compute binding free energies. The number of replicas and their scaling parameters were determined automatically using YANK's trailblazing module. The peptide was kept inside the binding site by adding harmonic restraints, which would appear before the nonbonded parameters would disappear.

The binding free energy differences ($\Delta\Delta G$) were estimated using a thermodynamic cycle that involves the free binding energies of both forms. The energy difference on the rightmost arrow represents the difference in binding energy given by the oxidation of the methionine residues.

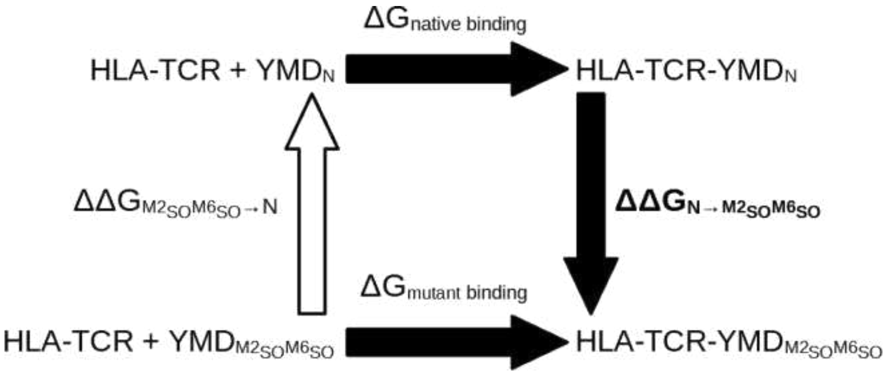
$\Delta G_{native\ binding}$
HLA-TCR + YMD_N
HLA-TCR-YMD_N
$\Delta\Delta G_{M2_{SO}M6_{SO} \rightarrow N}$
$\Delta\Delta G_{N \rightarrow M2_{SO}M6_{SO}}$
$\Delta G_{mutant\ binding}$
HLA-TCR + $YMD_{M2_{SO}M6_{SO}}$
HLA-TCR-$YMD_{M2_{SO}M6_{SO}}$

Figure 5-2 Thermodynamic cycle on which the estimation of oxidation effect was based.

The results of the MBAR calculation are presented in Table 5-1

$\boldsymbol{\Delta G}$ N ↑	$\boldsymbol{\Delta G}$ $M2_{SO}M6_{SO}$ ↑	$\boldsymbol{\Delta G}$ N ↓	$\boldsymbol{\Delta G}$ $M2_{SO}M6_{SO}$ ↓
-43.070 +/- 1.085	-43.277 +/- 0.595	–17.097 ± 0.669	–21.05 ± 0.557
$\boldsymbol{\Delta\Delta G}$↑		$\Delta\Delta G$↓	
~0.0 KCAL/MOL		~ –4.0 kcal/mol	

Table 5-1 Absolute Binding Free Energies for the four systems.

Analysis of Table 5-1 reveals that only the binding free energy difference in the 'down' (↓) configuration ($\Delta\Delta G$↓) is statistically significant. This suggests that the experimentally observed shift in equilibrium toward the oxidized form in the ternary complex is primarily driven by the YMD↓ conformation. Given that oxidation at M6 (M6SO) is expected to directly influence TCR binding, it is somewhat unexpected that the increased affinity for the oxidized form is mainly attributed to the 'down' (↓) configuration, rather than direct TCR interaction. However, binding free energy variations stem from complex thermodynamic equilibria, incorporating both enthalpic and entropic contributions, which result from a vast array of dynamic molecular interactions at room temperature.

Through analysis of the "fully-coupled" replica, it was found that the mutant form of the peptide increased the local stability of the binding pocket. First, the water molecule stability was increased in the mutant form, as measured by water residency time and the average number of water molecules in the site for the two systems.

	N ↓	$M2_{SO}M6_{SO}$ ↓
Average	11.5	15.4
St. Dev	5.8	5.1

Table 5-2 Average number of water molecules around the native/mutant YMD peptide

Secondly, the RMSF of both the YMD peptide and the TCR CDR loops are lower in the mutated form:

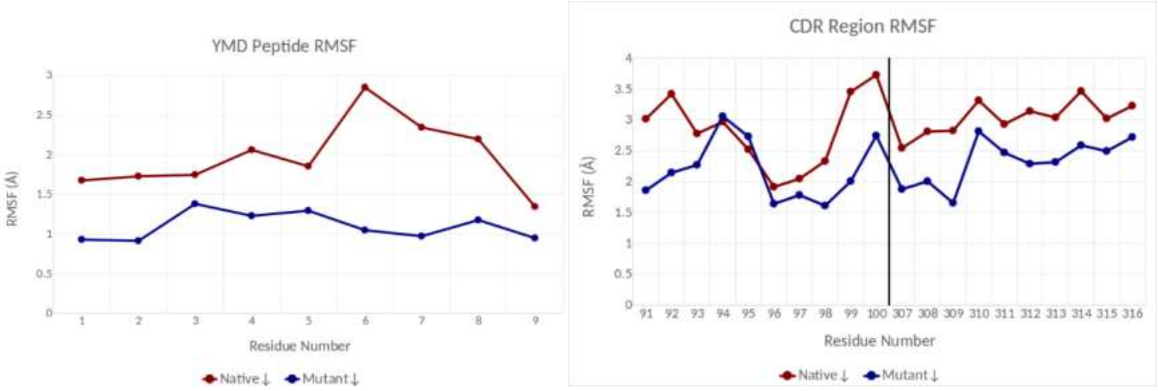

Figure 5-3 RMSF of the YMD Native/Mutant peptide and TCR CDR loops in the HLA/Native and HLA/Mutant complex. The vertical line separates the two CDR regions.

5. Conclusions

In the paper described in the current chapter, we generated a several models of the HLA: A0201/YMD binary complex and their ability to bind to a specific clone of a T-Cell Receptor. Not only did our computational results agree with the experimental findings (described in more detail in the paper), but through our models we were able to add insights into the underlying contribution of the sulfoxide moiety.

It is important to note that a large number of approximations were made during model building. First of all, the fact that the complex consists of multiple molecules that were built through homology modelling, not taken from any crystal structures, adds an unyielding degree of uncertainty. This is exacerbated by the fact that the CDR loops are very highly variable, and are also directly involved in the binding. Moreover, the CD8 protein, which is involved in stabilizing the ternary complex, was not included in the model, given the lack of structural information on its relative orientation.

In conclusion, computational data suggests that oxidation increases binding through an increase in local order of the interface, rather than through a direct interaction with the TCR.

CHAPTER 6 – INTERACTION OF VOLTAGE-DEPENDENT NAV1.5 CHANNELS WITH CENOBAMATE

1. Introduction

The present work presents a combination of molecular modeling and simulation techniques to investigate the effect of eight documented NaV1.5 channel point mutations upon its interaction with cenobamate, a novel anti-epileptic drug, mutations known to produce QT distortions. More specifically, it describes the modelling of the NaV1.5 α domains, as well as eight significant point mutations, the accurate modelling of its environment and its equilibration through increasingly flexible MD simulations. It then describes the method by which we used MD and MM-PBSA to assess whether the mutations described can affect the binding of cenobamate.

Structurally, Voltage-dependent Na+ channels (Nav) consist of four homologous pore-forming α domains, each consisting of six transmembrane helices. Between the 3rd and 4th α domains there is a short intracellular loop, which as an inactivation gate, blocking the pore from the inside during sustained membrane depolarization.

2. Model Generation

The model for the wildtype form of Na(v)1.5 was generated using AlphaFold2. The protein was embedded in the DPPC membrane using PACKMOL-Memgen (Schott-Verdugo and Gohlke, 2019). Explicit TIP3P type water with an ionic strength of 150 mM (composed of Na^+ and Cl^-) was used. An approximately 200A x 200A lipid patch was generated, with a 17.5 water buffer above and below the protein, in order to ensure an appropriate cell where the cytosolic domain would not communicate with the extracellular domain across the periodic boundary.

The mutant forms of the proteins were generated after equilibration of the wildtype form. The cenobamate ligand was parametrized using GAFF2.

Docking of the cenobamate molecule to the different mutants was done using Autodock Vina 1.2.5 (Trott and Olson, 2009)

Initial global minimization was performed using OpenMM. All molecular simulations were performed using NAMD3 (Phillips *et al.*, 2020). Each of the 9 forms of the Na(v)-1.5 receptor-Ligand complexes were subjected to 100ns of MD simulation. Considering the size of the system, a timestep of 2fs was used and SHAKE was applied to the hydrogen bonds.

To compute binding free energy, MM-PBSA calculations were performed. Calculation parameters were derived from similar use cases in other published work (Wang *et al.*, 2022). 100 equally-spaced snapshots were used from each simulation.

3. Results – Modelling and Simulation

The α domain of the human NaV1.5 was successfully modeled using AlphaFold2. The model was embedded into a 200Åx200Å patch of DPPC, the resulting system was solvated in an explicit water box, with a 17.5Å buffer above and below the protein. Na^+/Cl^- ions were added to neutralize the system and up to a concentration of 0.15M. The 8 mutants were generated after minimization and equilibration of the wildtype form. After generation, the mutants were minimized to ensure that no clashes would be present after the mutations have been done.

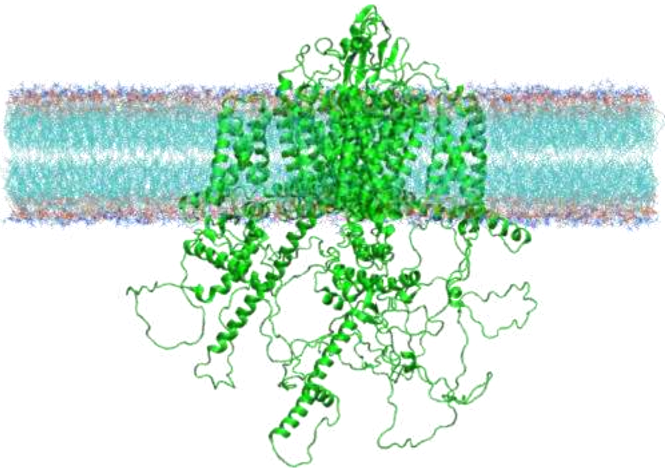

Figure 6-1 NaV1.5 (cartoon representation) embedded in the lipid membrane (line representation). The water and ions have been hidden for clarity

Over the 100ns explicit solvent simulations, the 9 models of NaV1.5 did not suffer any major conformational changes, as measured by the RMSD.

RMSD analysis of the cenobamate molecule, docked to the different forms of the NaV1.5 protein show that only the N1463K mutation, during its equilibration period, allows the ligand to be more flexible, all other mutations make it more rigid than the wildtype form. The tighter binding exhibited with the mutant forms could indicate that there are enthalpic influences that drive the lower estimated K_d values (Table 6-1).

4. Results – Binding Free Energy Calculation

The effect of point mutations on the binding of cenobamate to NaV1.5 has been estimated both via molecular docking and MM-PBSA. As presented in Table 6-1, the docking assays suggest that N927S, N1463K, N1463Y and M1766R bind cenobamate stronger than the wildtype. However, the MM-PBSA computations predict that all mutants bind the cenobamate molecule stronger than the wildtype, with the N932S mutant being the top candidate for binding.

MUTANT	AVERAGE DOCKING SCORE (KCAL/MOL)	$\Delta G_{BINDING}$ MM-PBSA (KCAL/MOL)
WILDTYPE	-5.50	-9.0828+/-3.0573
N927S	-5.62	-11.6598+/-2.3896
N932K	-5.00	-11.6168+/-2.6316
N932S	-5.03	-14.394+/-2.0402
L935V	-4.83	-11.3162+/-2.7667
S1458Y	-5.11	-11.8996+/-2.2859
N1463K	-6.20	-11.8148+/-2.1851
N1463Y	-6.01	-11.0899+/-1.7401
M1766R	-6.03	-10.1092+/-2.0133

Table 6-1 Estimated binding affinity of cenobamate to different forms of human NaV1.5

5. Conclusions

Cenobamate, a novel antiseizure drug, exhibits binding effects on the NaV1.5 channel at clinically relevant concentrations. The presented work shows that the compound binds in the central cavity of the wildtype NaV1.5 and stronger to selected mutant variants. As such, mutations in the cardiac NaV1.5 channel should be considered when prescribing cenobamate, since mutations in said channel could lead to potentially fatal ventricular arrhythmias.

Overall conclusions

The present thesis describes the work undertaken during my PhD training in Dr. Petrescu's Department of Bioinformatics and Structural Biochemistry at IBAR. This was an exciting time that concurred with groundbreaking progress in bioinformatics and biocomputing. This allowed me to get a grip on new state of the art techniques in molecular modeling and simulation and use them to get an in-depth insight into the behavior of several complex biomolecular systems relevant in molecular medicine.

The first part of my work is dedicated to exploring the capabilities and limits of the new generation of DL driven structure prediction platforms in modeling of more complex multi-state, multi-domain protein systems – by taking as a case study a CNL NOD-Like receptor family from *Arabidopsis thaliana* on which our department has significant structural expertise and results over the past decade. Our results indicate that while such automatic modeling platforms show high accuracy in predicting the structure of sequence modules that do fold in well-known, validated topologies they fall short in predicting the configuration of all-alpha coiled-coil regions, remote homologues and regions lacking templates – were the more flexible manual heuristic techniques outperform automatic modeling.

We have developed AlphaFold2 information filtering workflows aimed to drive the modeling process toward specific configurations in multistate instances. This reveals two major research avenues, in tailoring the input data (sequences and structures) on a family-by-family basis and in using multimer structural templates to generate monomer structures with multimer conformation, which is much less time-intensive and resource-consuming.

Both of the described avenues will be explored further after my thesis defense, mostly by writing a fully-fledged pipeline that streamlines the selection of sequences and structures.

Beyond modeling, which returns unique 'frozen' structures, the second part of this thesis presents my contributions to the development of Robosample – a highly efficient simulation platform of molecular conformational sampling based on robotic algorithms developed and coordinated in our Department by Dr. Laurențiu Spiridon.

Robosample – which is the first Romanian molecular simulation platform – is an absolute achievement that is currently in full development and expansion. Hence, I intend to make Robosample my first and foremost work priority after my thesis defense by continuing to optimize the end-user experience and collaborating with others in the group research on new Gibbs block selections. Additionally, I plan to implement Hamiltonian Monte Carlo to complement the existing T-REX implementation and to explore the efficiency of the software for molecular docking simulations.

The third part of my PhD work focusses on using molecular modeling and simulation to assess molecular interactions in complex biomolecular systems relevant in molecular medicine.

For instance, the work on Thrombopoietin receptor and YMD peptide gave both practical insights into molecular mechanisms relevant in oncology by addressing aspects related to myeloproliferative neoplasms and melanoma, respectively. The first one identifies the molecular mechanisms by which a C-terminal frameshift in the Calreticulin gene, expressing a lectin chaperone in ER, induces the constitutive activation of a cytokine receptor, specifically the thrombopoietin receptor, while the second unravels mechanistic details related to the effect of oxidation on YMD tyrosinase epitope recognition, and why it has a greater effect on T-cell reactivity. The work on these systems firstly involved the generation of highly confident computational models of tetrameric/ternary complexes given that no adequate experimental structures have had been solved for either system at the time when the research was conducted. Moreover, these highly accurate models allowed binding free energy calculations to be performed for both systems through various molecular dynamics simulation-based techniques.

Finally, the third study highlights molecular modeling and simulation application in pharmacogenetics by describing our work on the voltage-dependent sodium channel and the effect of different mutations on the binding of cenobamate, a novel anti-seizure drug. Our binding free energy computations showed that cenobamate displays increased affinity to all mutants compared to wild type NaV1.5 channels and binds especially well to the N932S mutation. Thus, it would be best to be mindful of the presence of this mutation when prescribing cenobamate to epileptic patients.

All in all, the results presented herein reveal on one hand the current trends in biocomputing method development and on the other hand the power of computational techniques in molecular life sciences research, especially when these are used in workflows that incorporate experimental results.

List of Personal Contributions

1. Spiridon L, **Şulea TA**, Minh DDL, Petrescu AJ. *"Robosample: A Rigid-Body Molecular Simulation Program Based on Robot Mechanics"*, *Biochimica Biophysica Acta General Subjects* 1864(8): 129616, (2020) DOI: 10.1016/j.bbagen.2020.129616;

   AI: 1.40(**Q2**) **IF**: 3.68
2. Chiritoiu G., Munteanu CVA., **Şulea TA**., Spiridon L., Petrescu AJ., Jandus C., Romero P., Petrescu SM. "Methionine oxidation selectively enhances T cell reactivity against a melanoma antigen", iScience(107205), (2023) DOI: 10.1016/j.isci.2023.107205

   AI: 1.63(**Q1**) **IF**: 6.10
3. Papadopoulos N, Nédélec A, Derenne A, **Şulea TA**, Pecquet C, Chachoua I, Vertenoeil G, Tilmant T, Petrescu AJ, Mazzucchelli G, Iorga BI, Vertommen D, Constantinescu SN "Oncogenic CALR mutant C-terminus mediates dual binding to the thrombopoietin receptor triggering complex dimerization and activation", Nature communications 14(1): 1881, (2023), DOI: 10.1038/s41467-023-37277-3

   AI: 5.61(**Q1**) **IF**: 17.69
4. **Şulea TA**, Draga S, Mernea M, Corlan AD, Radu BM, Petrescu AJ, Amuzescu B . "Differential Inhibition by Cenobamate of Canonical Human Nav1.5 Ion Channels and Several Point Mutants", International journal of molecular sciences 26(1), (2025), DOI: 10.3390/ijms26010358

AI: 1.05(**Q1**) **IF**: 4.90

5. **Şulea TA**, Martin EC, Bugeac CA, Bectaș FS, Iacob AL, Spiridon L, Petrescu AJ . "Lessons from Deep Learning Structural Prediction of Multistate Multidomain Proteins-The Case Study of Coiled-Coil NOD-like Receptors", International journal of molecular sciences 26(2), (2025) DOI: 10.3390/ijms26020500

AI: 1.05(**Q1**) **IF**: 4.90

References

Abramson, J. *et al.* (2024) 'Accurate structure prediction of biomolecular interactions with AlphaFold 3', *Nature*, 630(8016), pp. 493–500. Available at: https://doi.org/10.1038/s41586-024-07487-w.

Ahdritz, G. *et al.* (2024) 'OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization', *Nature Methods*, 21(8), pp. 1514–1524. Available at: https://doi.org/10.1038/s41592-024-02272-z.

Berman, H.M., Vallat, B. and Lawson, C.L. (2020) 'The data universe of structural biology', *IUCrJ*, 7(4), pp. 630–638. Available at: https://doi.org/10.1107/S205225252000562X.

Carugo, O. and Djinović-Carugo, K. (2023) 'Structural biology: A golden era', *PLOS Biology*, 21(6), p. e3002187. Available at: https://doi.org/10.1371/journal.pbio.3002187.

Cassidy, C.K. *et al.* (2020) 'Structure and dynamics of the E. coli chemotaxis core signaling complex by cryo-electron tomography and molecular simulations', *Communications Biology*, 3(1), p. 24. Available at: https://doi.org/10.1038/s42003-019-0748-0.

Defour, J.P. *et al.* (2013) 'Tryptophan at the transmembrane-cytosolic junction modulates thrombopoietin receptor dimerization and activation', *Proceedings of the National Academy of Sciences of the United States of America*, 110(7), pp. 2540–2545. Available at: https://doi.org/10.1073/pnas.1211560110.

Eastman, P. *et al.* (2024) 'OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials', *Journal of Physical Chemistry B*, 128(1), pp. 109–116. Available at: https://doi.org/10.1021/acs.jpcb.3c06662.

Hewitt, E.W. (2003) 'The MHC class I antigen presentation pathway: strategies for viral immune evasion', *Immunology*, 110(2), pp. 163–169. Available at: https://doi.org/10.1046/j.1365-2567.2003.01738.x.

Jumper, J. *et al.* (2021) 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596(7873), pp. 583–589. Available at: https://doi.org/10.1038/s41586-021-03819-2.

Kelley, L.A. *et al.* (2016) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols*, 10(6), pp. 845–858. Available at: https://doi.org/10.1038/nprot.2015-053.

Kollman, P.A. *et al.* (2000) 'Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models', *Accounts of Chemical Research*, 33(12), pp. 889–897. Available at: https://doi.org/10.1021/ar000033j.

Krishna, R. *et al.* (2024) 'Generalized biomolecular modeling and design with RoseTTAFold All-Atom', *Science*, 384(6693), p. eadl2528. Available at: https://doi.org/10.1126/science.adl2528.

Livnah, O. *et al.* (1999) 'Crystallographic Evidence for Preformed Dimers of Erythropoietin Receptor Before Ligand Activation', *Science*, 283(5404), pp. 987–990. Available at: https://doi.org/10.1126/science.283.5404.987.

Lyskov, S. and Gray, J.J. (2008) 'The RosettaDock server for local protein-protein docking', *Nucleic Acids Research*, 36(Web Server), pp. W233–W238. Available at: https://doi.org/10.1093/nar/gkn216.

Martin, E.C. *et al.* (2023) 'NLRscape: an atlas of plant NLR proteins', *Nucleic Acids Research*, 51(D1), pp. D1470–D1482. Available at: https://doi.org/10.1093/nar/gkac1014.

Melo, F. and Feytmans, E. (1997) 'Novel knowledge-based mean force potential at atomic level', *J. Mol. Biol*, (267), pp. 207–222. Available at: https://doi.org/10.1186/1471-2105-7-324.

Montgomery Pettitt, B. and Karplus, M. (1985) 'The potential of mean force surface for the alanine dipeptide in aqueous solution: a theoretical approach', *Chemical Physics Letters*, 121(3), pp. 194–201. Available at: https://doi.org/10.1016/0009-2614(85)85509-3.

Phillips, J.C. *et al.* (2020) 'Scalable molecular dynamics on CPU and GPU architectures with NAMD', *Journal of Chemical Physics*, 153(4). Available at: https://doi.org/10.1063/5.0014475.

Prentoe, J. *et al.* (2019) 'Hypervariable region 1 and N-linked glycans of hepatitis C regulate virion neutralization by modulating envelope conformations', *Proceedings of the National Academy of Sciences*, 116(20), pp. 10039–10047. Available at: https://doi.org/10.1073/pnas.1822002116.

Rajaraman, J. *et al.* (2016) 'An LRR/Malectin Receptor-Like Kinase Mediates Resistance to Non-adapted and Adapted Powdery Mildew Fungi in Barley and Wheat', *Frontiers in Plant Science*, 7. Available at: https://doi.org/10.3389/fpls.2016.01836.

Sarson-Lawrence, K.T.G. *et al.* (2024) 'Cryo-EM structure of the extracellular domain of murine Thrombopoietin Receptor in complex with Thrombopoietin', *Nature Communications*, 15(1), p. 1135. Available at: https://doi.org/10.1038/s41467-024-45356-2.

Schott-Verdugo, S. and Gohlke, H. (2019) 'PACKMOL-Memgen: A Simple-To-Use, Generalized Workflow for Membrane-Protein–Lipid-Bilayer System Building', *Journal of Chemical Information and Modeling*, 59(6), pp. 2522–2528. Available at: https://doi.org/10.1021/acs.jcim.9b00269.

Shen, W., Zhou, T. and Shi, X. (2023) 'Enhanced sampling in molecular dynamics simulations and their latest applications—A review', *Nano Research*, 16(12), pp. 13474–13497. Available at: https://doi.org/10.1007/s12274-023-6311-9.

Shirts, M.R. and Chodera, J.D. (2008) 'Statistically optimal analysis of samples from multiple equilibrium states', *Journal of Chemical Physics*, 129(12). Available at: https://doi.org/10.1063/1.2978177.

Slootweg, E.J. *et al.* (2013) 'Structural Determinants at the Interface of the ARC2 and Leucine-Rich Repeat Domains Control the Activation of the Plant Immune Receptors Rx1 and Gpa2 1 [C][W][OA]', 162(July), pp. 1510–1528. Available at: https://doi.org/10.1104/pp.113.218842.

Spiridon, L. *et al.* (2020) 'Robosample: A rigid-body molecular simulation program based on robot mechanics', *Biochimica et Biophysica Acta - General Subjects*, 1864(8). Available at: https://doi.org/10.1016/j.bbagen.2020.129616.

Spiridon, L. and Minh, D.D.L. (2017) 'Hamiltonian Monte Carlo with Constrained Molecular Dynamics as Gibbs Sampling', *Journal of Chemical Theory and Computation*, 13(10), pp. 4649–4659. Available at: https://doi.org/10.1021/acs.jctc.7b00570.

Syed, R.S. *et al.* (1998) 'Efficiency of signalling through cytokine receptors depends critically on receptor orientation', *Nature*, 395(6701), pp. 511–516. Available at: https://doi.org/10.1038/26773.

Trakhanov, S. *et al.* (1999) 'Structure of a monoclonal 2E8 Fab antibody fragment specific for the low-density lipoprotein-receptor binding region of apolipoprotein E refined at 1.9 Å', *Acta Crystallographica Section D Biological Crystallography*, 55(1), pp. 122–128. Available at: https://doi.org/10.1107/S090744499800938X.

Trott, O. and Olson, A.J. (2009) 'AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading', *Journal of Computational Chemistry*, 31(2), p. NA-NA. Available at: https://doi.org/10.1002/jcc.21334.

Tsutsumi, N. *et al.* (2023a) 'Structure of the thrombopoietin-MPL receptor complex is a blueprint for biasing hematopoiesis', *Cell*, 186(19), pp. 4189-4203.e22. Available at: https://doi.org/10.1016/j.cell.2023.07.037.

Tsutsumi, N. *et al.* (2023b) 'Structure of the thrombopoietin-MPL receptor complex is a blueprint for biasing hematopoiesis', *Cell*, 186(19), pp. 4189-4203.e22. Available at: https://doi.org/10.1016/j.cell.2023.07.037.

Vaidehi, N. and Jain, A. (2015) 'Internal Coordinate Molecular Dynamics: A Foundation for Multiscale Dynamics', *The Journal of Physical Chemistry B*, 119(4), pp. 1233–1242. Available at: https://doi.org/10.1021/jp509136y.

Vangone, A. and Bonvin, A.M. (2015) 'Contacts-based prediction of binding affinity in protein–protein complexes', *eLife*, 4. Available at: https://doi.org/10.7554/eLife.07454.

Vangone, A. and Bonvin, A.M.J.J. (2015) 'Contacts-based prediction of binding affinity in protein–protein complexes', *eLife*, 4(JULY2015), pp. 1–15. Available at: https://doi.org/10.7554/eLife.07454.

Varadi, M. *et al.* (2024) 'AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences', *Nucleic Acids Research*, 52(D1), pp. D368–D375. Available at: https://doi.org/10.1093/nar/gkad1011.

Waman, V.P. *et al.* (2024) 'CATH 2024 : CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds', (xxxx). Available at: https://doi.org/10.1016/j.jmb.2024.168551.

Wang, K. *et al.* (2013) 'Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics', *Journal of Computer-Aided Molecular Design*, 27(12), pp. 989–1007. Available at: https://doi.org/10.1007/s10822-013-9689-8.

Wang, S. *et al.* (2022) 'MM/PB(GB)SA benchmarks on soluble proteins and membrane proteins', *Frontiers in Pharmacology*, 13. Available at: https://doi.org/10.3389/fphar.2022.1018351.

Webb, B. and Sali, A. (2018) 'Comparative Protein Structure Modeling Using MODELLER', *Physiology & behavior*, 176(1), pp. 139–148. Available at: https://doi.org/10.1117/12.2549369.Hyperspectral.

Williams, C.J. *et al.* (2018) 'MolProbity: More and better reference data for improved all-atom structure validation', *Protein Science*, 27(1), pp. 293–315. Available at: https://doi.org/10.1002/pro.3330.

Wu, R. *et al.* (2022) 'High-resolution de novo structure prediction from primary sequence', *bioRxiv* [Preprint]. Available at: https://doi.org/10.1101/2022.07.21.500999.

Xue, L.C. *et al.* (2016) 'PRODIGY: A web server for predicting the binding affinity of protein-protein complexes', *Bioinformatics*, 32(23), pp. 3676–3678. Available at: https://doi.org/10.1093/bioinformatics/btw514.