Prediction of A. thaliana's MCTP4 Structure using Deep Learning-Based tools and Exploration of Transmembrane domain Dynamics using Coarse-Grained Molecular Dynamics Simulations

Sujith Sritharan,[†] Raphaelle Versini,[†] Jules Petit,[‡] Emmanuelle Bayer,[‡] and Antoine Taly[*,†]

†*Laboratoire de Biochimie Théorique, CNRS (UPR9080), Université Paris Cité, F-75005 Paris, France*

‡*Laboratoire de Biogenèse Membranaire, Unité Mixte de Recherche 5200, Université de Bordeaux, Centre National de la Recherche Scientifique, Villenave d'Ornon, France.*

E-mail: taly@ibpc.fr

Abstract

Multiple C2 Domains and Transmembrane region Proteins (MCTPs) in plants have been identified as important functional and structural components of plasmodesmata cytoplasmic bridges, which are vital for cell-cell communication. MCTPs are endoplasmic reticulum (ER)-associated proteins which contain three to four C2 domains and two transmembrane regions. In this study, we created structural models of *Arabidopsis* MCTP4 ER-anchor transmembrane region (TMR) domain using several prediction methods based on deep learning. This region, critical for driving ER association, presents a complex domain organization and remains largely unknown. Our study

demonstrates that using a single deep-learning method to predict the structure of membrane proteins can be challenging. Our deep learning models presented three different conformations for the MCTP4 structure, provided by different deep learning methods, indicating the potential complexity of the protein's conformational landscape. For the first time, we used simulations to explore the behaviour of the TMR of MCTPs within the lipid bilayer. We found that the TMR of MCTP4 is not rigid, but can adopt various conformations including some not identified by deep learning tools. These findings underscore the complexity of predicting protein structures. We learned that combining different methods, such as deep learning and simulations, enhances our understanding of complex proteins.

INTRODUCTION

Plasmodesmata (PD) are intercellular channels found in plants that allow for the communication and transport of molecules between adjacent cells.[1]PD have a unique membrane organization characterized by tight membrane contact sites, consisting of two concentric membranes - the plasma membrane (PM) and the endoplasmic reticulum (ER).[1] The regulation of intercellular trafficking through these channels is essential for plant growth, development, and defense against biotic and abiotic stresses.[2]

The ER-PM tethering machinery of membrane contact sites in plasmodesmata has been hypothesised to play a crucial role in PD formation, reshaping, and proper function.[1,2] Members of the MCTP family are plasmodesmata-localised and act as ER-PM tethers.[3] These tethering proteins selectively concentrate at plasmodesmata, bind membranes together, and were hypothesised to control the exchange of information between cells.[2]

MCTP proteins consist of two transmembrane regions and three or four tandem C2 domains. The C2 domains act as PM docking sites through interaction with anionic sites, while the transmembrane domain insert into the ER membrane. This structural organization is essential for their function in PD. While C2 domains are relatively conserved and the 3D

structure is more easily accessible, the TMR presents a complex domain organization that has not yet been elucidated.[3]

Recently, a variety of novel protein structure prediction tools, namely Alphafold and RosettaFold, as well as new strategies based on large language models, including ESMFold and OmegaFold, have emerged.[4–7] These tools are powerful and trained to recognize evolutionary preserved structural motifs. However, they are also very recent, and we are just beginning to understand their limitations. One such limitation is that membrane proteins are underrepresented in the training datasets, which could lead to uncertain models. On the other hand, new methods based on language models, such as the one developed by Lin et al. (2022), are not limited by experimental data.[7] These methods employ the protein sequence, potentially generating more accurate predictions for membrane proteins.

Despite these challenges, in this study, we employed several available deep learning algorithms in the literature to test for convergence, aiming to overcome potential limitations and obtain reliable predictions for the MCTP4 ER-anchor TMR domain. First, we generated different models using various deep learning-based prediction tools, namely Alphafold(AF), Alphafold multimer(AFM), RosettaFold(RF), Tr-RosettaFold(TR), ESMFold(ESM), and OmegaFold(OF).[4–9] We compared the results of these different prediction tools to ensure the reliability of the generated models. Finally, we employed molecular dynamics (MD) simulations with the Martini 3 coarse-grained and Charmm36 all-atom force fields, along with principal component analysis. We examine the structure and temporal flexibility of the TMR, as well as its behavior within a lipid bilayer. We find that MD explores the conformational space sampled by DL tools and beyond.

MATERIAL AND METHODS

Jupyter Notebook

All of the analysis and data used in this study have been documented and made available for reference and reuse. These scripts are hosted in Jupyter Notebooks, a popular open-source web application that allows for the creation and sharing of documents. The notebooks can be accessed from the associated GitHub repository at https://github.com/Jouffluu/Molecular-modelling-MCTP.

Predicted Models

The three-dimensional structure of the full-length *Arabidopsis thaliana* MCTP4 also called FT-interacting protein 4 (Uniprot: Q9C8TM3) was obtained using deep-learning prediction tools. We utilized Alphafold (version 2.2),[4] Alphafold multimer (version 2.2),[8] Omegafold (version 1.1.0),[6] and ESMfold (version 1.0.3),[10] which were run on a local cluster. For Rosettafold, we used the public webserver (https://robetta.bakerlab.org/submit.php),[5] and for Tr-rosetta, we utilized the webserver (https://yanglab.nankai.edu.cn/trRosetta).[9]

Computational Detail

All simulations, both Coarse-Grained (CG) and All-Atom (AA), were performed with GROMACS 2021.5 simulation package, with system setup facilitated by the CHARMM-GUI webserver.[11] For each simulation type, we initiated the process with an energy minimization phase, followed by a series of equilibration steps, before proceeding to the production simulations.

System preparation

We utilized the Martini maker from the CHARMM-GUI web server for system construction and mapping of atomistic structures to CG Martini 3 models.[12] Our simulations targeted residues 550 to 776, encompassing 50 residues preceding the transmembrane domain, the transmembrane (TM) helix domains (600 to 750), and extending to the end of the protein. To investigate the stability of the TM helix, we performed simulations in PIPC/PIPE (80:20) bilayer membranes. Each model was oriented using the PPM server by incorporating the topology of the N-terminus of the first chain from the PDB file. Subsequently, each system was solvated in water, neutralized, and supplemented with 0.15 M NaCl. Additionally, we created all-atom systems using the membrane builder from CHARMM-GUI,[13] following the same process with the CHARMM36 force field, and constructed the membranes with equivalent lipids.

CG-MD simulations

In all simulations, GROMACS 2021.5 simulation package was used with the Martini 3 force field.[14,15] The protocol consisted of an initial energy minimization phase, followed by a multi-stage equilibration process, and concluded with a production simulation. Energy minimization was carried out for two iterations of 5,000 steps each using the steepest descent method. The simulations were subsequently equilibrated through five stages, employing time steps of 2, 5, 10, 15, and 20 fs. A target temperature of 300 K was maintained with the v-rescale thermostat, with a coupling constant of 1 ps. An semiisotropic pressure of 1 bar was maintained using the Parrinello-Rahman barostat,[16] with a compressibility of 4.5 × 10-5 bar-1 and a relaxation time constant of 12 ps. Long-range interactions were treated with a cutoff radius of 1.1 nm for both van der Waals and Coulombic interactions, using a switching function from 1.0 nm for van der Waals. The production simulations were performed using an NPT ensemble with a time step of 20 fs for a total simulation time of 3 microseconds. Multiple replicates were conducted for the system, and the final phase of the simulation was

executed with no restraints.

AA-MD simulations

In our AA-MD (All-Atom Molecular Dynamics) simulations, we constructed three distinct transmembrane systems (AF, ESM, OF) using the CHARMM36m force field via CHARMM-GUI. To mimic the environment of plant endoplasmic reticulum, we opted for a lipid composition of 20% PLPE and 80% PLPC. A Verlet cutoff scheme was employed for interaction handling. Van der Waals interactions were addressed with a cutoff and force-switch at 1.0 nm, while electrostatic interactions were managed via the Particle Mesh Ewald method, both set to a cutoff of 1.2 nm. Pressure coupling was achieved using the Parrinello-Rahman method with a semiisotropic type.

Contact analysis

To quantify contacts between transmembrane (TM) helices among the different models predicted by the algorithms, we used the MDAnalysis library to calculate distances between the centers of mass of the residues.[17,18] During the simulations, contacts between the beads representing the helices were defined using a cutoff distance of 12 Å, and contact occupancies were calculated for each bead pair. Subsequently, a contact map was generated using the R language for statistical computing (R Core Team, 2022) for each model.[19]

Principal Component Analysis and Clustering

To reveal the most important motions in the TM helices, we employed principal components analysis (PCA) using the tools provided in the GROMACS software package.[14] For all models, we first fitted the trajectories where the transmembrane (TM) helix part was stable within the membrane for 3 microseconds to ensure that the phosphates of the membrane remained in the same position. We then concatenated these adjusted trajectories.

The covariance matrix was calculated on the backbone atoms of the residues located in the membrane (HP1 and HP2). The first two principal components were then plotted using the matplotlib library for statistical visualization.[20] The results of the PCA were clustered using the K-Medoids method. Utilizing the KMedoids class from the scikit-learn-extra library,[21] we specified five clusters for categorization. We determined this optimal number of clusters using the elbow method. After fitting the model to the PCA data, the coordinates of the cluster centroids were determined.

Data analysis and visualisation

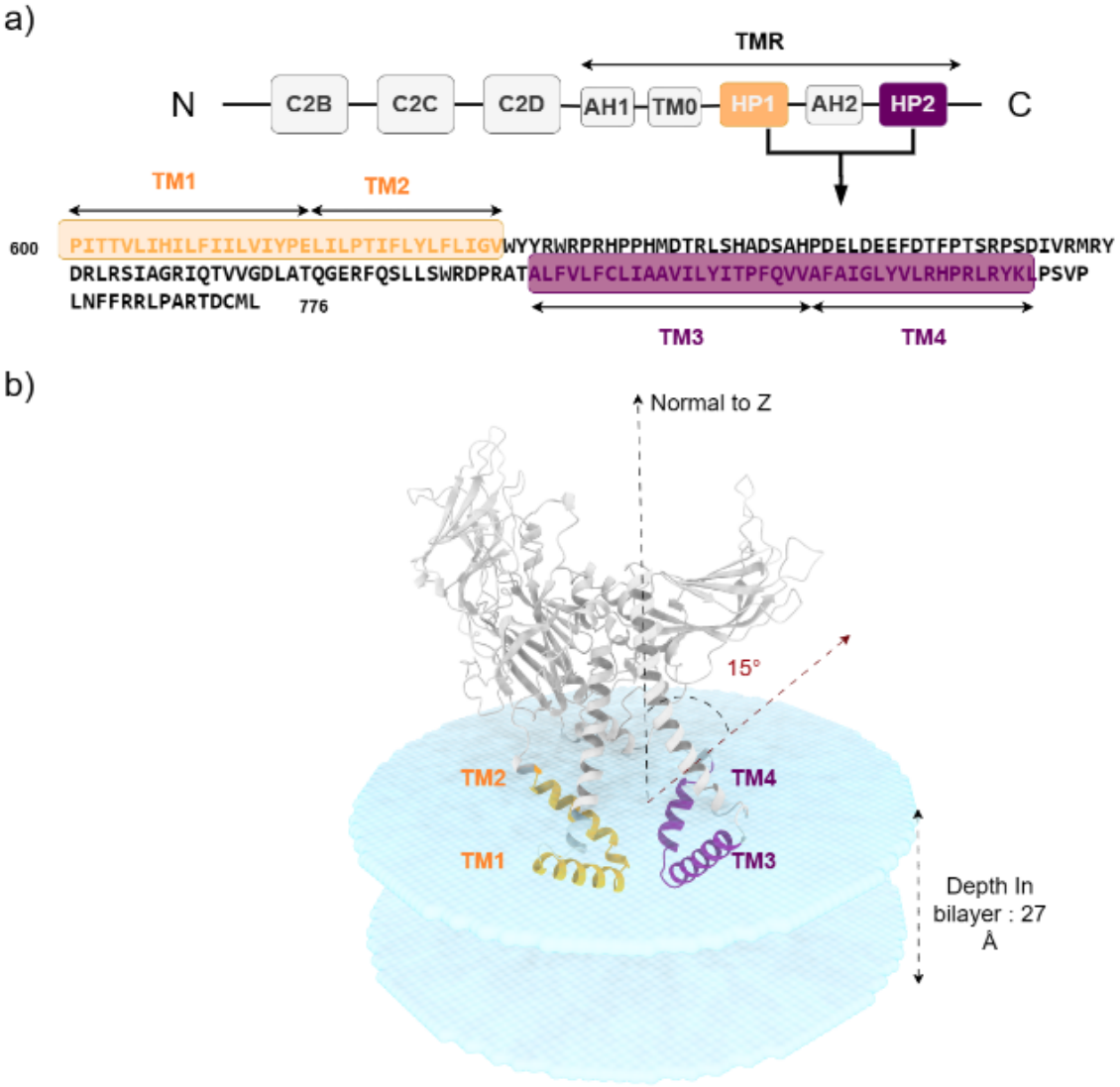
Some basic analysis tasks, such as the calculation of the root mean square deviation (RMSD) or the distances between centers of the geometry of the two helices, were performed using the tools provided in the GROMACS 2021.5 software package.[14] For visualization purposes, we constructed a density plot to provide a view of the distribution of data points within the first two principal components (PC1 and PC2). More complex analyses, such as the tilt angle calculation, were implemented in the Python programming language with extensive use of the MDAnalysis (https://github.com/rversin/MD-TMtools). VMD[22] were used to visualise the trajectories and ChimeraX was used to analyse alphafold output.[23]

RESULTS

Models of MCTP4 ER-TMR domain

We generated six models of MCTP4 ER TMR domain using six different prediction methods, in addition to a partial model of MCTP4 TMR domain previously constructed by Modeller and used as a reference.[24] This later work utilized bioinformatic tools, hydrophobic clusters analysis and molecular dynamics, to delineate transmembrane helices within the MCTP4 protein.[24] We decided to use the resulting definition of membrane domains. Therefore, we extracted the 550-776 region from each model. We previously identified five putative

subdomains within the approximately 200-residue sequence of MCTP4 transmembrane domain (TMR). These subdomains include an N-term amphipathic helix (APH1), a putative transmembrane domain (TMD0), a hairpin transmembrane domain (HP1) composed of two transmembrane helices (TM1 and TM2), a second, longer amphipathic helix (APH2), and another hairpin transmembrane domain (HP2) which is also composed of two transmembrane helices (TM3 and TM4).[24] These subdomains are illustrated in Figure 1.



(a)

Figure 1: a) MCTP4 TMR having 5 elements: 2 transmembrane hairpins (HP1 and HP2) and 3 amphipathic helices (AH1, TM0 and AH2). b) AF MCTP4 model oriented in the lipid bilayer by OPM server.[25]

Subsequently, an alignment and Root Mean Square Deviation (RMSD) calculation were performed on the extracted regions and on the whole model. The results of this analysis are displayed in Table 1 and Figure 3(g-i).

Table 1: RMSD values for the transmembrane domain (550-776) (upper half) and for the alpha carbon of entire models (lower half). For each pair of models, the RMSD value is indicated for each measurement.

	AF	AFM	Rose	Tr-rose	Esmfold	Omegafold
AF	0	13.23	14.29	13.55	14.28	15.63
AFM	8.03	0	6.18	3.07	5.57	15.67
Rose	18.04	16.58	0	6.56	8.13	16.17
Tr-rose	8.75	4.43	17.16	0	4.95	15.37
Esmfold	16.11	15.13	19.28	13.35	0	16.13
Omegafold	26.40	27.00	27.00	27.37	31.33	0

The PLDDT score was used to assess the confidence of the 3D structure prediction algorithms. A low PLDDT score indicates low confidence in the 3D structure, while a high score indicates higher confidence. Confidence scores varied among the models, particularly in the membrane region, as illustrated in Supplementary Figure 8. Among the prediction methods, ESM appeared to be the most confident in its predictions for both TM domains, while AlphaFold (AF2) seemed to be less confident in the HP2 domain compared to the other methods (Figure 2). The other methods showed relatively high confidence in their predictions.

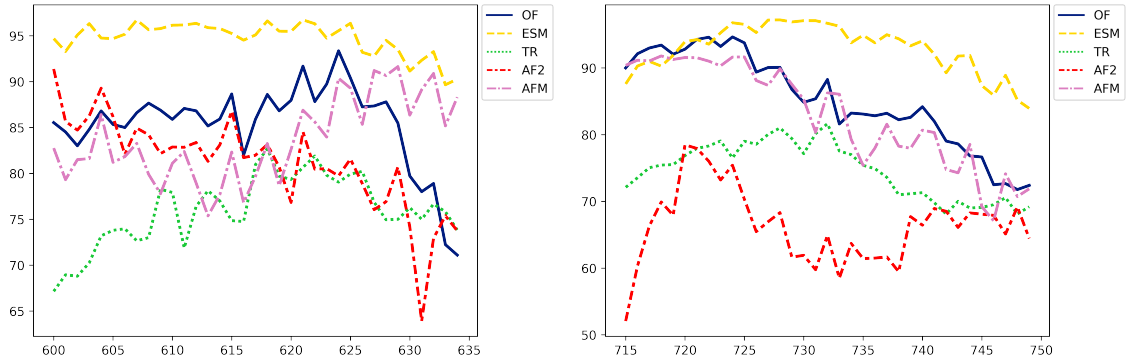
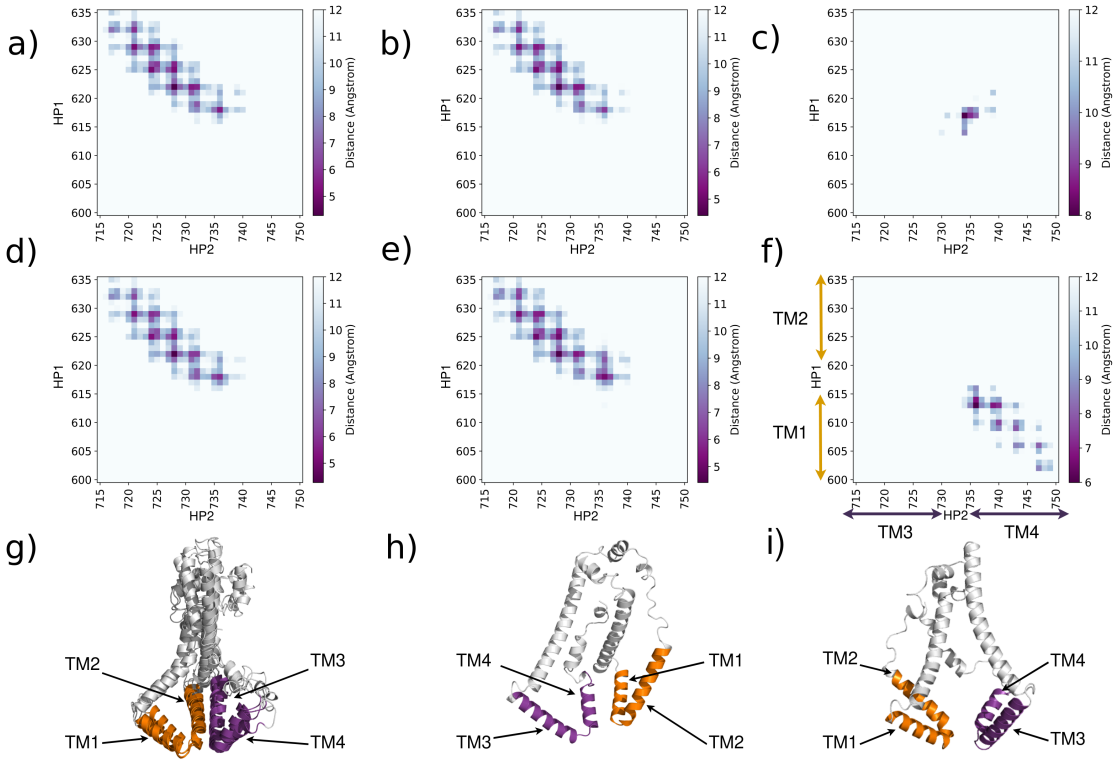


Figure 2: Evaluation of model predictions using the pLDDT score as a function of residues in the HP1 (left) and HP2 (right) regions. The curves of different colors represent models predicted by various prediction methods: AlphaFold (AF, red), OmegaFold (OF, blue), TR (green), ESM (yellow), and AlphaFold Multimer (AFM, magenta)

When comparing the contact maps, which are two-dimensional representations of three-

dimensional structures,[26,27] of the models predicted by each method, we observed that ESM, TR, RF, and AFM predictions showed a similar pattern (Figure 3 a-b,d-e). In these conformations, the TM2 and TM3 helices (615-635, 715:735) were found to be in close proximity. On the other hand, the models made by AF2 and OF displayed different conformations in the membrane domain. In AF2's model, the TM2-TM3 helices were separated by more than 8 angstroms with almost no contact (Figure 3e/i). Furthermore, it's root mean square deviation (RMSD) was high, over 13 angstrom, compared to other models (Table 1). This shows that AF2's model was quite different in terms of structure. When looking at the conformation obtained by OF, the close connection was between helices TM1 and TM4 (residues 600:615, 735:750), as seen in Figure 3f/h. OF's model also had a high RMSD, around 15, compared to the other models (Table 1). These large RMSD values and different helix arrangements highlight the range of protein conformations predicted by different methods.



(a)

Figure 3: Contact maps between the residues of transmembrane HP1 and HP2 of different protein structures: a) ESM, b) AFM, c) AF, d) TR, e) RF and f) OF. The distances were calculated using the alpha carbon atoms. Heat map coloring indicates the distance between residues, ranging from dark purple for short distances to white for longer distances(12 A). g) superposition of ESM, AFM, TR and RF transmembrane domain models. h) and j) are OF and AF transmembrane domains.

Dimers Because other MCTP have been proposed to dimerize,[28] we wanted to determine whether the ER-domain of MCTP4 has the potential to form dimers. The analysis of the dimer produced by Alphafold multimer revealed that intra-subunits contacts were similar to those found in the ESM, RF, and TR models, between TM2 and TM3. In terms of inter-subunits contacts, the TM1 helix of one monomer and the TM4 helix of the other one were found to interact (Figure 4 a/c). Furthermore, the algorithm exhibits high confidence in the intra-domain but also in the inter-subunits-interactions (Figure 4 b/d).

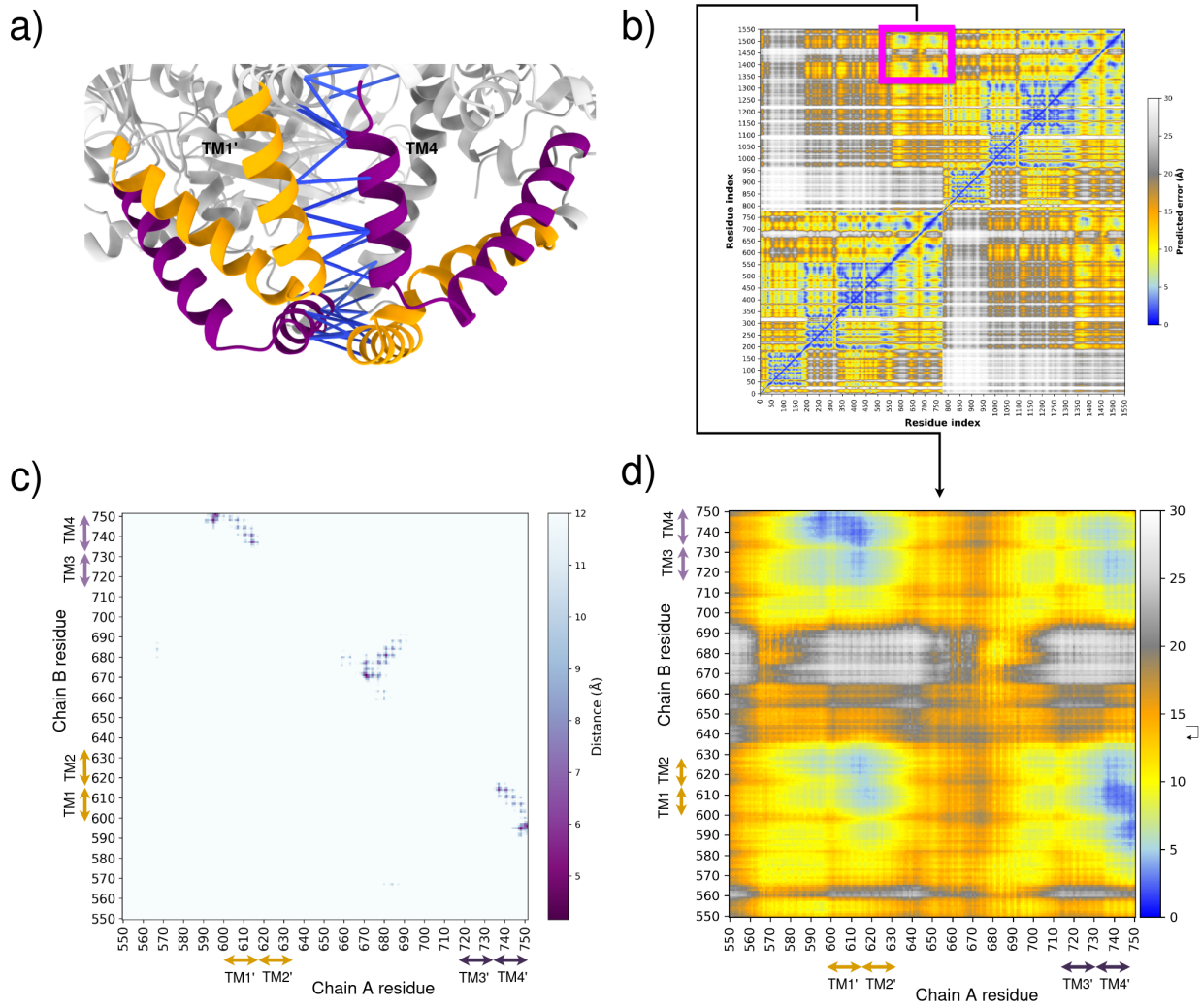


Figure 4: a) AlphaFold multimer prediction of TMR MCTP4 protein. The HP1 domain is represented in orange and the HP2 domain in purple. The pLDDT score between residues of interchain domains is represented by lines, with the blue color indicating a high score. b) Predicted Aligned Error (PAE) for the MCTP4 dimer, extracted by AlphaPickle.[29] PAE provides an estimate of the error in the predicted alignment of residues. c) Inter-chain contact map of TMR. d) Close-up view of the PAE for the TMR of MCTP4

In every model, the MCTP4 TMR incorporates two hairpins predicted to be in the bilayer membrane. While four of the models converge on similar arrangements and distances between the hairpins, two models provide notably different configurations. To investigate the stability of each model and whether they would interconvert, we studied them using molecular dynamics simulations.

Molecular Dynamics Simulations

We performed coarse-grained molecular dynamics simulations with the Martini 3 force field, to investigate the behaviour of the TMR domains of each model, produced by the deep learning methods, included in a lipid bilayer that mimics the composition of the ER, as detailed in the materials and methods section. Ten replicates were conducted for each model, each lasting 3 microseconds. During the initial simulations, we observed that certain TM domains (HP1 or HP2 or both) exited the membrane, which could potentially introduce variability in our analyses (see Supplementary Material Figure 10-15). To ensure consistency, we decided to exclude these simulations from our analysis. We then repeated the simulations until we obtained four simulations for each starting model whose TM domains remained inside the membrane throughout the entire simulation. This approach allowed us to obtain a consistent dataset for further analysis.

Interestingly, out of the four simulations conducted for each of the AF2 and ESM models, one simulation from each model exhibited significant variations in the distances between the TM2 (HP1) and TM3 (HP2) helices during the simulation, with both proximity and separation observed. Despite these movements, these particular simulations demonstrated stability within the membrane, indicating that the helices were not confined to a single conformation (see Supplementary Materials Figure 13). This also show that the system's movement was not ballistic, which suggest that the simulations might have converged to equilibrium.

To gain a deeper understanding of the dynamic behaviour of these helices within the membrane, we conducted further characterization using principal component analysis, to get insights into the conformational dynamics of TM domains in lipid bilayer membranes.

Principal Component Analysis and Clustering

We employed principal component analysis (PCA) as a statistical method to elucidate the most significant motions of the TM helices within the lipid bilayer. In PCA, the Cartesian

coordinates (X, Y, Z) of each atom were used as descriptors to capture the accessible degrees of freedom of the protein. Specifically, we selected the residues of the TM helices located in the membrane to perform the PCA. The resulting conformations from the simulations were projected onto the first two principal components, which accounted for 49% of the variability observed in our simulations (see Figure 5a). Each data point in the figure represents a conformation from the trajectory of an MCTP TMR model obtained through different methods, with the respective starting points of each model also displayed. Despite distinct starting points for each model, the conformations generated by the simulations appear to convergence towards two distinct basins (see Figure 5c). The main basin regroups the four similar models (ESM, RF, TR and AFM) and most of the structures including some coming from simulations started from AF2 and OF. The second basin does not include any of the starting structures but is formed by structures extracted from the simulations started from various models (AF2, ESM, RF and TR).

To identify the centers of these basins, we performed clustering using the k-meloids algorithm on the two principal components extracted from PCA. This allowed us to identify five distinct clusters within the conformational space (see Figure 5b). From each cluster, we extracted representative structures that best represented the characteristics of that particular cluster.

For the first cluster (Cluster A), we obtained a representative structure that closely resembled the initial "consensus" starting point of the simulation. This cluster is part of the main basin, suggesting that the simulation explored a conformational space similar to the initial structure throughout most of the simulation time.

In the second cluster (Cluster B), we obtained a representative structure that exhibited significant structural changes compared to the initial starting points. In this structure, helices TM2 are in contact with both TM3 and TM4 helices, which is novel and not captured by deep learning methods (see Figure 5d). This suggests that the simulation explored a different conformational space and underwent substantial structural rearrangements.

In the third cluster (Cluster C), which is also part of the main basin, the representative structure shows contacts corresponding to the TM1-TM4 and TM2-TM3 helices. This indicates a unique conformation not present in the initial models, further highlighting the simulation's ability to explore alternative conformations.

In the fourth cluster, (Cluster D), the representative structure illustrates contacts between TM1-TM3 and TM1-TM4 helices. This suggests a distinct conformation, emphasizing the diversity of conformations accessible to the system.

Finally, in the fifth cluster (Cluster E), the representative structure corresponds to a conformation where the TM domains are separated (5d). This suggests that this simulation explored a wider range of conformational space, further highlighting the dynamic nature of the system under study.

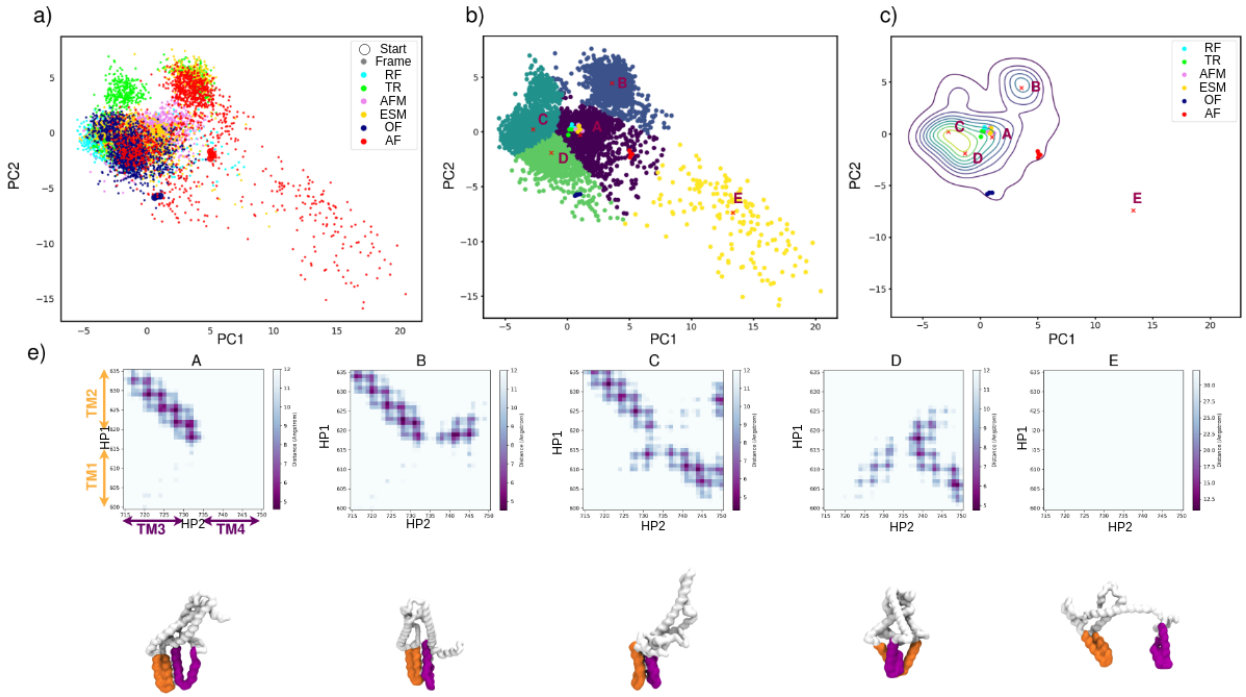


Figure 5: Principal Component Analysis (PCA) plots, clustering and density representation. a) The plot shows a projection based on a PCA's first two principal components (PCs). Large points represent the starting points of each simulation for each model, and small points are conformations generated by the simulation in each model. Each model is represented by its own colour: AlphaFold (AF, red), OmegaFold (OF, blue), TR-rosetta (TR, green), ESMFold (ESM, yellow), and AlphaFold Multimer (AFM, magenta). b) The centroids of each basin have been determined using the K-means clustering method. c) A density plot showing the projection of the two PCs, where each point corresponds to the starting point of each model. d) The representative structure of each cluster is shown with a contact map corresponding to contacts between transmembrane domains.

The PCA and clustering analysis have provided deeper insights into the structure and dynamics of the MCTP4 TMR. Our findings indicate that the MCTP4 TMR is not rigid, but displays substantial potential for structural rearrangement within the bilayer. There's a prominent convergence towards two main conformational basins, signifying two key structural states that the TMR can adopt. Furthermore, we identified certain novel helical contacts, implying that the MCTP4 TMR can explore a broad conformational space. This underscores the dynamic and versatile nature of the MCTP4 TMR in its biological context.

Atomistic Simulations for Enhanced Analysis

To complement and reinforce our analyses based on coarse-grained simulations, we also carried out atomistic simulations of the TMR domains. These simulations were performed on the three forms of TMR domains (ESM, AF and OF) obtained through deep learning methods, aiming to validate the stability and reliability of the predicted conformations. For each form, three replicas were conducted, each lasting 250 nanoseconds. Root Mean Square Deviation (RMSD) analyses confirmed the stability of these replicas suggesting a high reliability of the observed conformations. (see Supplementary 7).

Furthermore, Principal Component Analysis (PCA) revealed that each replica was firmly anchored in its own conformational space, without significant overlap between different spaces. This observation underlines that the replicas did not explore other conformational regions, indicating a clear segregation and distinction between the conformations of each TMR domain form.(see Supplementary 6).

These atomistic results complement our coarse-grained simulations by demonstrating that the TMR domain forms, while being stable and distinct within their respective conformational spaces, do not visit other potential conformations. This highlights the advantage of coarse-grained simulations in exploring a wide array of conformations, thus offering a broader perspective on the conformational dynamics of proteins in membrane environments, complemented by the precision of atomistic simulations to validate and deepen our understanding of specific molecular dynamics.

Combined, these approaches provide a comprehensive and detailed view of the dynamics and stability of TMR domains within the membrane, thereby enriching our understanding of the underlying mechanisms governing the behaviour of MCTP4 protein in a biomembrane context.

Discussion

Our study revealed that the MCTP4 TMR domain, located within the ER, consists of two hairpin, each containing two helices. These hairpin are deeply embedded within the lipid bilayer of the ER. The arrangement of these hairpins and their proximity to each other varied, reflecting the complexity and dynamism of the MCTP4 TMR structure within an ER-mimicking bilayer.

Exploration of the conformational landscape by deep learning models.

In this study, we generated six distinct models of MCTP4 TMR using various prediction methods, which allowed us to explore the conformational landscape of the protein. Notably, PLDDT scores revealed varied confidence levels in the predictions of MCTP4's TMR 3D structure, with ESM showing the highest confidence in its predictions for both TM domains. It is interesting to note that four of the prediction methods (ESM, TR, RF, and AFM) converged on a similar MCTP's transmembrane domain, where helices TM2 and TM3 are in close proximity. This suggests that this particular conformation might be a reliable representation of the protein's actual structure. However, AF2 and OF produced distinct models with different helical arrangements. These divergences could indicate the flexibility of MCTP4 TMR and the presence of alternative conformations. This also raises the importance of using multiple prediction methods in parallel to explore the conformational landscape of a protein. By relying on a consensus among several methods, we can potentially obtain a more accurate and reliable picture of the protein's structure.[30] This approach is particularly relevant given the complexity of protein structure prediction and the possibility that different tools have complementary strengths and weaknesses. By combining the results, individual errors can be mitigated, and more robust conclusions can be drawn regarding the likely structure of a protein.

Molecular Dynamics Simulations

Simulations Behavior with Martini 3

During our simulations, we investigated the dynamic behaviors of the TM helices of MCTP4 within a lipid bilayer membrane designed to mimic the composition of the endoplasmic reticulum. With Martini 3, we were able to observe TM Helices that show spatial flexibility. Nevertheless, among the initial 10 replicas for each model, we observed a diverse range of helix behaviors within the membrane. Some helices were coming out of the membrane (Supplementary Material Figure 14-19). Consequently, we repeated the simulations multiple times to obtain a consistent and stable set of results within the membrane. In order to ensure the reliability of our analysis, we excluded simulations in which the TM domains exited the membrane. The Martini 3 force field has been developed in part to solve the issue identified with martini2, that tends overestimate protein-protein interactions.[15] However, Martini3 is conversely associated with too hydrophobic alpha-helices. Therefore modifications to the scaling of interactions within the Martini 3 force field have been proposed to solve this issue. Thomasen et al., proposed adjusting protein-protein interaction parameters to 88% of their original value in the Martini 3 force field.[31] The authors found that this scaling leads to results that are closer to experimentally observed protein-membrane interactions, compared to a different scaling factor that modifies protein-water interactions. This did not solve the issue in our case (data not shown).

The plus of Molecular dynamics run on top of Deep-Learning Models

Three starting points were generated by deep learning methods and for each model, 4 times 3 μs simulations were launched. This allowed the exploration of a conformational landscape with two basins. The center of the first basin is very close to the conformation majorly found by deep learning algorithms. However, the other clusters present representative structures with contacts that are not found in deep learning methods, suggesting that with MD sim-

ulation, we were able to explore new conformations which are not found by deep learning methods. It's also interesting to see that the AF2 and OF models start right at the edge of the consensus basin and then move into it. This shows how useful it is to use the physics-based MD simulations together with models made by deep learning. By using both methods, we can benefit from their respective strengths and attain a more robust result.

Monomer versus Dimer contacts

Analyzing the predictions of deep learning models, we have observed significant differences between the TMR models of AF2, OF, and those that have converged, namely RF, AFM, TR, and ESM.

In the TMR model of AF2, the two helices of each transmembrane hairpin do not come into contact. This feature stands in stark contrast with those observed in other models. In the TMR model from OF, it's noteworthy that helices TM1 and TM4 are in contact. Similarly, the converged models exhibit a distinct contact between helices TM2 and TM3. This analysis thus reveals a variety of structural configurations predicted by different deep learning models, underscoring the complexity of transmembrane interactions of MCTP4.

Through the use of coarse-grained molecular dynamics simulations, we found that the interaction between TM2 and TM3 tended to be more stable. This observation is supported by our density map, which reveals two distinct basins. Models that display these interactions are predominantly found in the larger of the two basins.

In contrast, the models produced by AlphaFold (AF) and OpenFold (OF) are positioned away from these basins. This positioning suggests that these models may represent an intermediate state. Further insights can be gleaned from studying the dimeric form of the protein.

Interestingly AF2 and AFM provided different results. This prompted us to explore further the difference. Noteworthy, AF2 showed significant deviations and obtained a low PLDDT score, particularly in the TM regions. This implies that AlphaFold's monomeric

predictions are not always reliable or accurate for certain protein domains. The variation observed between monomeric and multimeric predictions could indicate that the formation of the dimeric structure involves additional interactions or structural rearrangements not captured in the monomeric prediction. A key capability of AF2 is to allow the prediction of contacts from sequence alignments.[32,33] The relationship between sequences and contacts is however partially ambiguous, which has been shown in the case of conformational changes. This in turn triggered the creation of strategies to explore the conformational landscape with AF2 and RoseTTAFold.[34–39] It is therefore tempting to speculate that evolutionary signals are not necessarily captured by AlphaFold's monomeric models for membrane proteins that form homo-oligomers. In the case of AF2, the model might be subject to conflicting constraints corresponding to intra-subunit and inter-subunit contacts. In agreement with that observation, with AFM the inter-subunit contacts between TM1 from one subunit and TM4 helix from the other subunit are evaluated to be positive (Figure 4).

The model produced by OF shows contacts between TM1 and TM4 helices that rather appear to be inter-subunit contacts with AFM. This is coherent with the notion that there are conflicting constraints, based on coevolution, corresponding to intra-subunit and inter-subunit contacts in homo-oligomers, in particular using a monomeric prediction tool such as OF. Interestingly, the exploration of this model with a physics-based method like molecular dynamics could help resolve these conflicting constraints.

Finally, predicting the structures of multimers is a complex task, as assessing interfaces increases the difficulty compared to monomers. This raises a debate on whether to treat interfaces separately, as suggested by Zhu and colleagues.[40] Furthermore, this leads to the question of whether multimeric proteins should be modeled as monomers or oligomers. Oligomeric proteins could benefit from more precise treatment as oligomers rather than monomers, as part of the evolutionary pressure or signal might be associated with interfaces.

Overall, our study on the MCTP4 protein's transmembrane domain showed that predicting protein structures is complicated and needs a mix of methods. By using different tools,

including deep learning and MD simulations, we got a detailed view of how this part of the protein moves and changes conformation. We also learned that it's crucial to think about how proteins group together and that simple models like AlphaFold might not always get it right. This study shows that using multiple approaches is key to understanding protein structures, especially for the parts that are in cell membranes.

Acknowledgments

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (project n° 772103-BRIDGING .), labex DYNAMO (11-LABX-0011) and ANR projects DIVCON (ANR-21-CE13-0016-01) and MITOFUSION (ANR-19-CE11-0018).

References

(1) Tilsner, J.; Nicolas, W.; Rosado, A.; Bayer, E. M. Staying Tight: Plasmodesmal Membrane Contact Sites and the Control of Cell-to-Cell Connectivity in Plants. *Annual Review of Plant Biology* **2016**, *67*, 337–364.

(2) Petit, J. D.; Li, Z. P.; Nicolas, W. J.; Grison, M. S.; Bayer, E. M. Dare to change, the dynamics behind plasmodesmata-mediated cell-to-cell communication. *Current Opinion in Plant Biology* **2020**, *53*, 80–89.

(3) Brault, M. L. et al. Multiple C2 domains and transmembrane region proteins (MCTPs) tether membranes at plasmodesmata. *EMBO reports* **2019**, *20*, e47182.

(4) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589, Number: 7873 Publisher: Nature Publishing Group.

(5) Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876, Publisher: American Association for the Advancement of Science.

(6) Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; Ma, J.; Peng, J. High-resolution de novo structure prediction from primary sequence. 2022; https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1, Pages: 2022.07.21.500999 Section: New Results.

(7) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Costa, A. d. S.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale prediction of atomic level protein structure with a language model. 2022; https://www.biorxiv.org/content/10.1101/2022.07.20.500902v3, Pages: 2022.07.20.500902 Section: New Results.

(8) Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2022**,

(9) Du, Z.; Su, H.; Wang, W.; Ye, L.; Wei, H.; Peng, Z.; Anishchenko, I.; Baker, D.; Yang, J. The trRosetta server for fast and accurate protein structure prediction. *Nature Protocols* **2021**, *16*, 5634–5651.

(10) Cabezudo, A. C.; Athanasiou, C.; Tsengenes, A.; Wade, R. C. Scaling protein-water interactions in the Martini 3 coarse-grained force field to simulate transmembrane helix dimers in different lipid environments. 2022; https://www.biorxiv.org/content/10.1101/2022.09.09.506752v1, Pages: 2022.09.09.506752 Section: New Results.

(11) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **2008**, *29*, 1859–1865.

(12) Hsu, P.-C.; Bruininks, B. M. H.; Jefferies, D.; Cesar Telles de Souza, P.; Lee, J.; Patel, D. S.; Marrink, S. J.; Qi, Y.; Khalid, S.; Im, W. CHARMM-GUI Martini Maker

for modeling and simulation of complex bacterial membranes with lipopolysaccharides. *Journal of Computational Chemistry* **2017**, *38*, 2354–2363.

(13) Wu, E. L.; Cheng, X.; Jo, S.; Rui, H.; Song, K. C.; Dávila-Contreras, E. M.; Qi, Y.; Lee, J.; Monje-Galvan, V.; Venable, R. M.; Klauda, J. B.; Im, W. CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *Journal of Computational Chemistry* **2014**, *35*, 1997–2004.

(14) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: fast, flexible, and free. *Journal of Computational Chemistry* **2005**, *26*, 1701–1718.

(15) Souza, P. C. T. et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature Methods* **2021**, *18*, 382–388, Number: 4 Publisher: Nature Publishing Group.

(16) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

(17) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Dotson, D. L.; Domański, J.; Buchoux, S.; Kenney, I. M.; Beckstein, O. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. Proceedings of the 15th Python in Science Conference. Austin, TX, 2016; pp 98–105.

(18) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32*, 2319–2327.

(19) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2022.

(20) Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007**, *9*, 90–95.

(21) learn-extra Developers, S. Scikit-learn-extra: A set of tools for scikit-learn. https://github.com/scikit-learn-contrib/scikit-learn-extra, 2023; [Online; accessed 04-August-2023].

(22) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(23) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **2021**, *30*, 70–82.

(24) Petit, J. Membrane Tethering in Plant Intercellular Communication : Structure-Function of Multiple C2 domains and Transmembrane Region Proteins (MCTP) at Plasmodesmata ER-PM Membrane Contact Site. Theses, Université de Bordeaux ; Université de Liège, 2022.

(25) Lomize, M. A.; Pogozheva, I. D.; Joo, H.; Mosberg, H. I.; Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Research* **2011**, *40*, D370–D376.

(26) Fakhoury, Z.; Sosso, G. C.; Habershon, S. Generating protein folding trajectories using contact-map-driven directed walks. *Journal of Chemical Information and Modeling* **2023**, *63*, 2181–2195.

(27) Mansoor, S.; Baek, M.; Park, H.; Lee, G. R.; Baker, D. Protein Ensemble Generation through Variational Autoencoder Latent Space Sampling. *bioRxiv* **2023**, 2023–08.

(28) Tran, T. M.; McCubbin, T. J.; Bihmidine, S.; Julius, B. T.; Baker, R. F.; Schauflinger, M.; Weil, C.; Springer, N.; Chomet, P.; Wagner, R., et al. Maize carbohydrate

partitioning defective33 encodes an MCTP protein and functions in sucrose export from leaves. *Molecular plant* **2019**, *12*, 1278–1293.

(29) mattarnoldbio, 2021, DOI: `10.5281/zenodo.5708709`.

(30) Graille, M.; Sacquin-Mora, S.; Taly, A. Best Practices of Using AI-Based Models in Crystallography and Their Impact in Structural Biology. *Journal of Chemical Information and Modeling* **2023**,

(31) Thomasen, F. E.; Skaalum, T.; Kumar, A.; Srinivasan, S.; Vanni, S.; Lindorff-Larsen, K. Recalibration of protein interactions in Martini 3. *bioRxiv* **2023**,

(32) Bhattacharya, N.; Thomas, N.; Rao, R.; Dauparas, J.; Koo, P. K.; Baker, D.; Song, Y. S.; Ovchinnikov, S. Single layers of attention suffice to predict protein contacts. *Biorxiv* **2020**, 2020–12.

(33) Bouatta, N.; Sorger, P.; AlQuraishi, M. Protein structure prediction by AlphaFold2: are attention and symmetries all you need? *Acta Crystallographica Section D: Structural Biology* **2021**, *77*, 982–991.

(34) Mitrovic, D.; McComas, S. E.; Alleva, C.; Bonaccorsi, M.; Drew, D.; Delemotte, L. Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning. *bioRxiv* **2022**, 2022–09.

(35) Wallner, B. AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling. *bioRxiv* **2022**, 2022–12.

(36) Stein, R. A.; Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLOS Computational Biology* **2022**, *18*, e1010483.

(37) Wayment-Steele, H. K.; Ovchinnikov, S.; Colwell, L.; Kern, D. Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *bioRxiv* **2022**, 2022–10.

(38) Schlessinger, A.; Bonomi, M. Exploring the conformational diversity of proteins. *Elife* **2022**, *11*, e78549.

(39) Del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **2022**, *11*, e75751.

(40) Zhu, W.; Shenoy, A.; Kundrotas, P.; Elofsson, A. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *bioRxiv* **2022**,

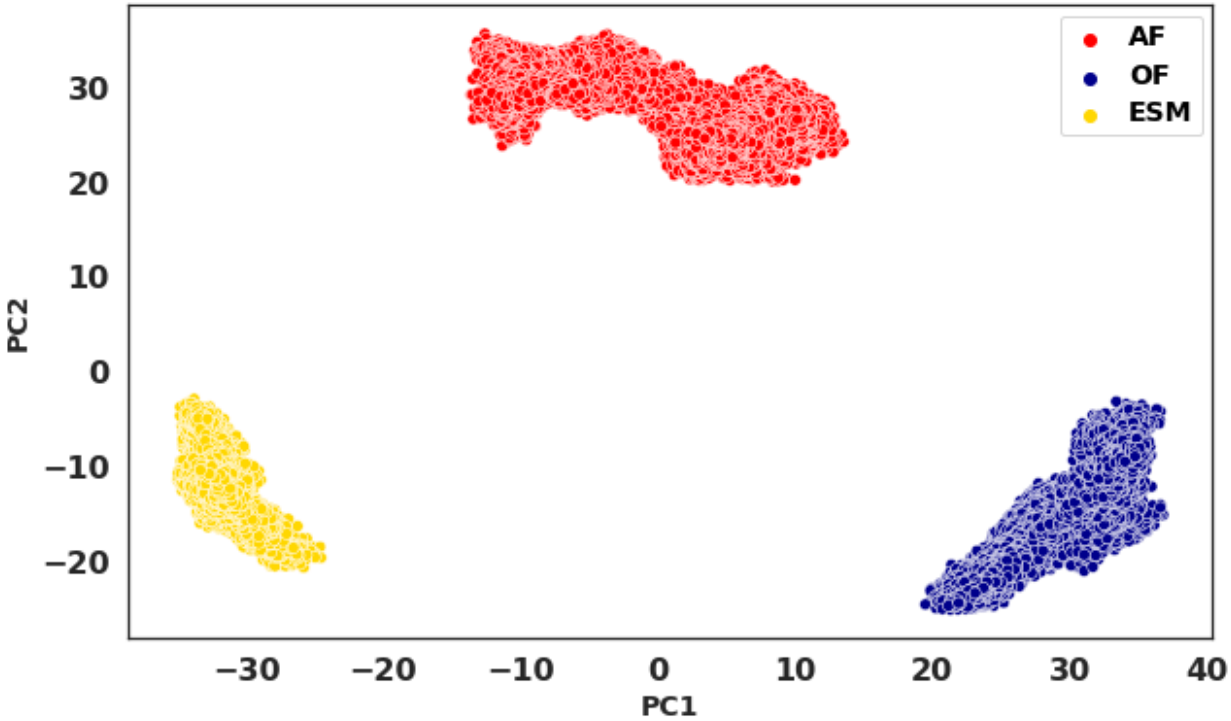
(41) Buchan, D. W. A.; Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research* **2019**, *47*, W402–W407.

(42) Chatzigoulas, A.; Cournia, Z. DREAMM: a web-based server for drugging protein-membrane interfaces as a novel workflow for targeted drug design. *Bioinformatics* **2022**, *38*, 5449–5451.

SUPPLEMENTARY

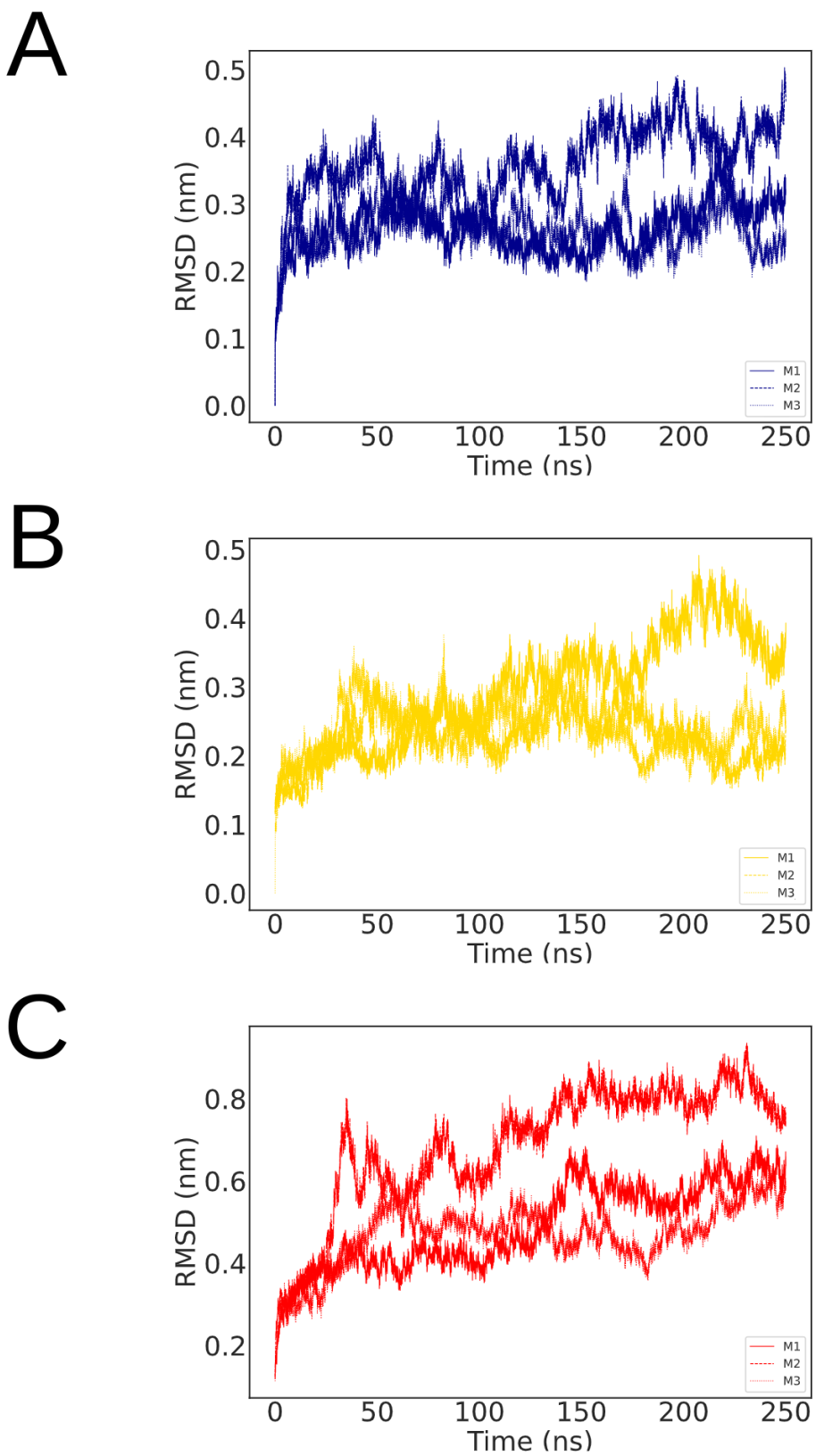
Table 2: Statistical information on each model from OPM server.[25]

Models	**Tilt**	**Depth**
AF	15 ± 0°	27.2 ± 1.2 Å
AFM	40 ± 0°	31.2 ± 0.7 Å
OF	37 ± 4°	31.2 ± 0.8 Å
ESM	34 ± 0°	27.8 ± 2.4 Å
RF	72 ± 1°	17.8 ± 1.2 Å
TR	33 ± 2°	30.8 ± 1.1 Å



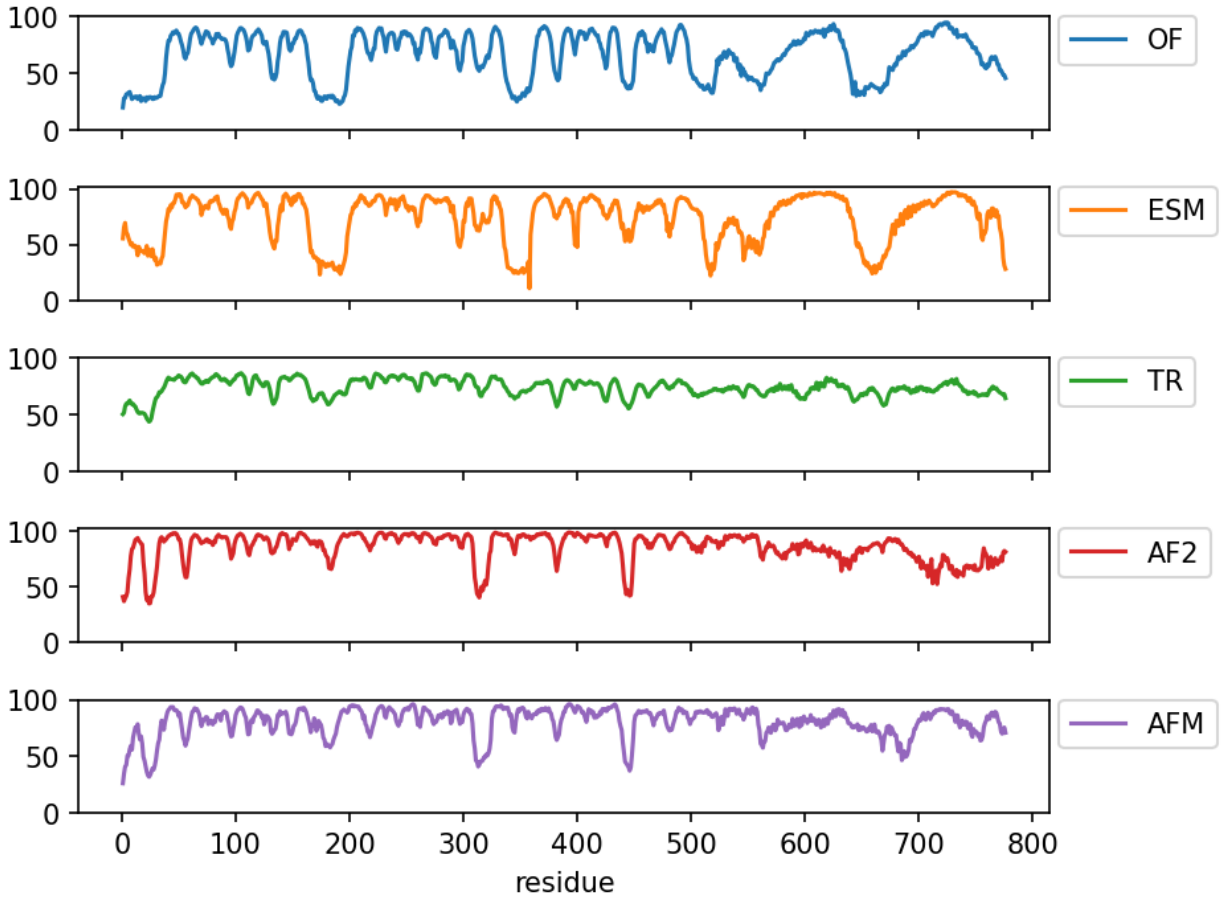
(a)

Figure 6: Projection of the first principal components (PC1 against PC2) from all atom simulation. Each point represents an observation. Colors represent different models: Red for AF2, Navy Blue for OMEGA, and Gold for ESM



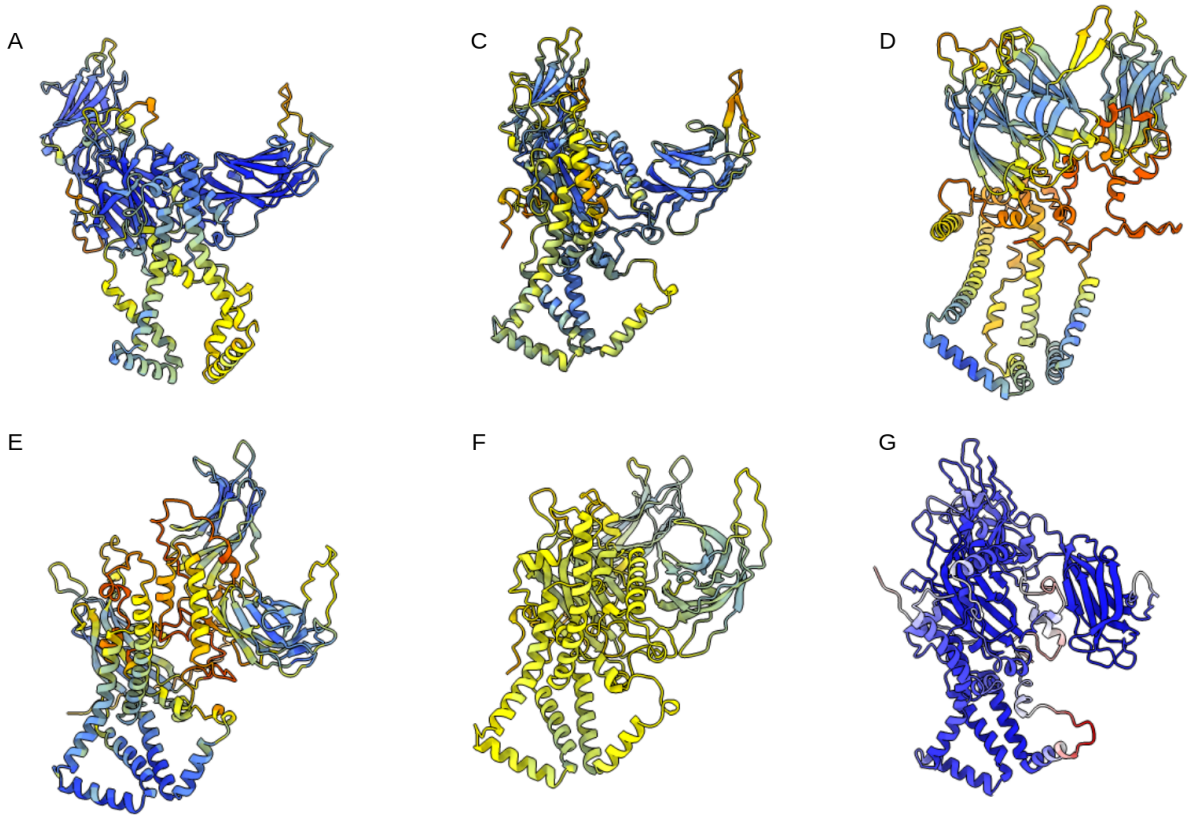
(a)

Figure 7: RSMD from all-atom simulation. Red for AF2, Navy Blue for OMEGA, and Gold for ESM



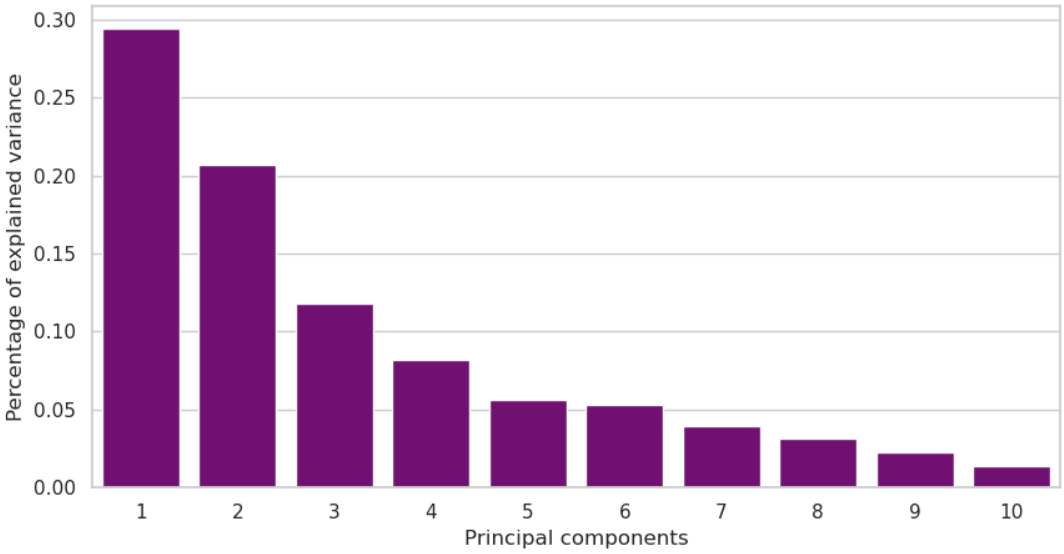
(a)

Figure 8: Evaluation of model predictions using the pLDDT score. The curves of different colors represent models predicted by various prediction methods: AlphaFold (AF, red), OmegaFold (OF, blue), TR (green), ESM (orange), and AlphaFold Multimer (AFM, purple)



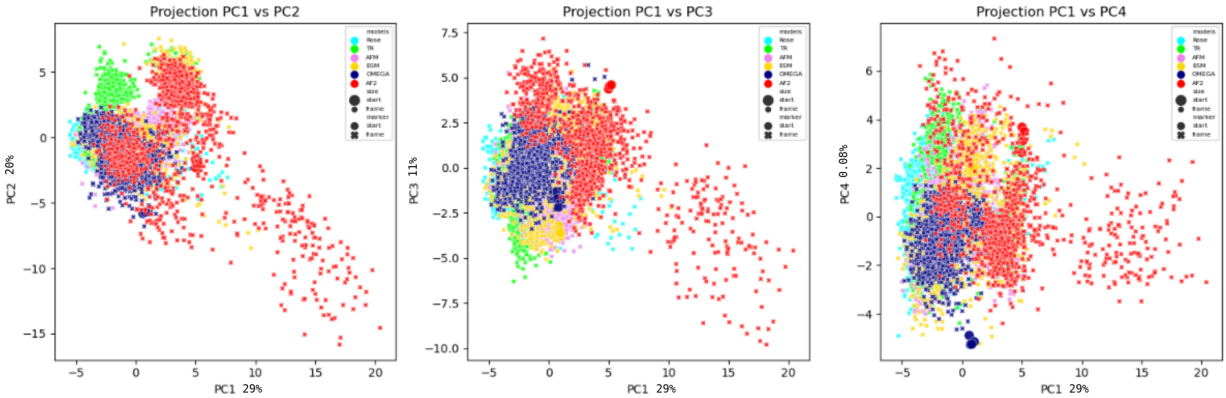
(a)

Figure 9: Panels A to F (AF, AFM, OF, ESM, TR): Structures are colored according to pLDDT scores; blue signifies areas with high confidence (reliable structural prediction), while red depicts areas with low confidence (uncertain structural prediction). Panel G displays the prediction made by RosettaFold, colored on the error estimate in the Å (RMSD) metric. This metric categorizes structural predictions into three levels of confidence: blue for very confident, red for less confident, and white for areas lacking confidence.



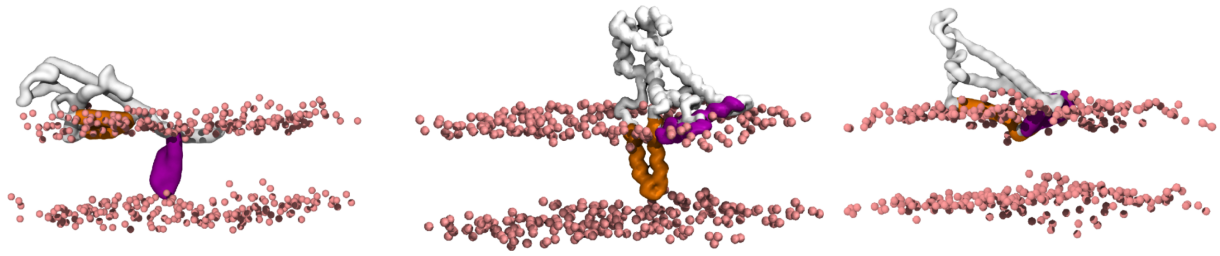
(a)

Figure 10: Variance explained by each Principal Component (PC) on CG simulations).



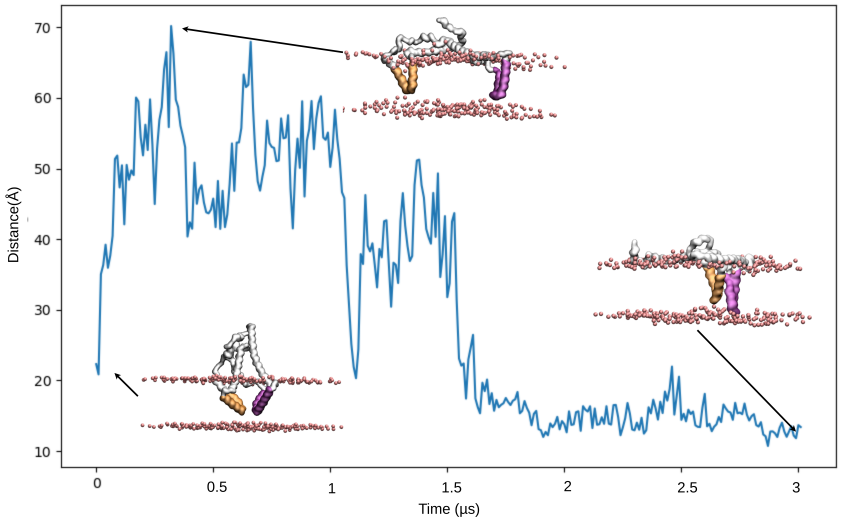
(a)

Figure 11: Projection of the first principal components (PC1 against PC2, PC3 and PC4). Each point represents an observation. Colors represent different models: Aqua for Rose, Lime Green for TR, Violet for AFM, Gold for ESM, Navy Blue for OMEGA, and Red for AF2



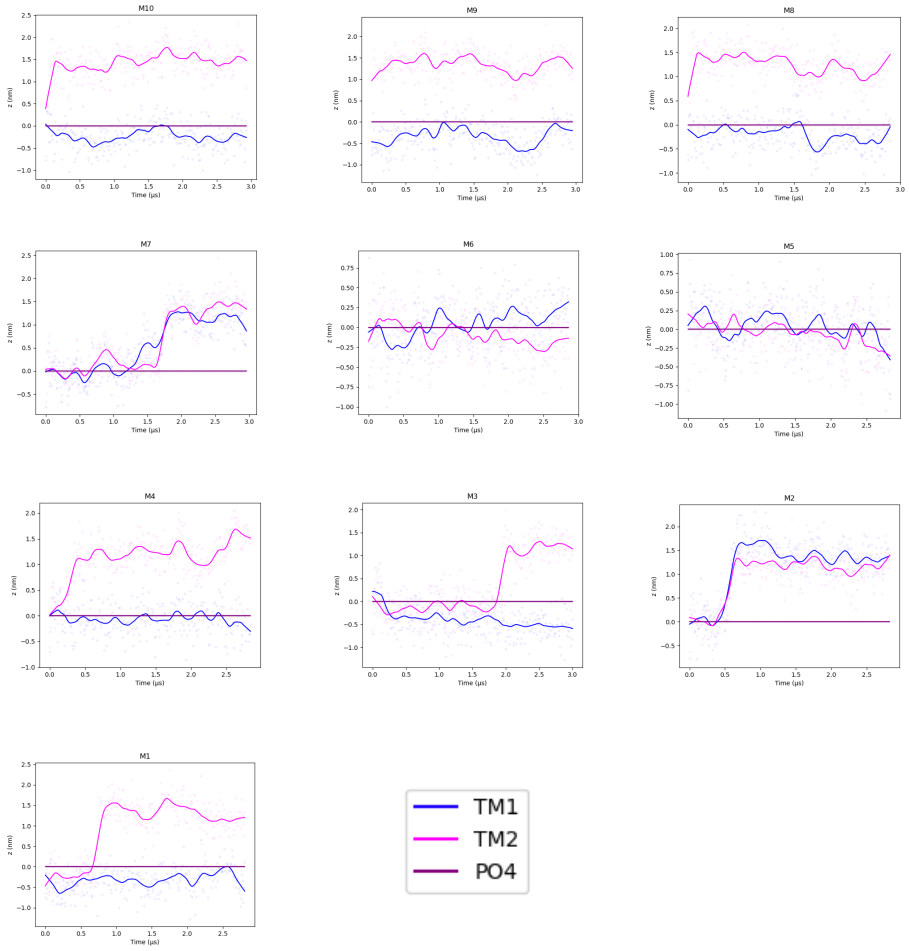
(a)

Figure 12: Left: HP2 domain emerging from the membrane. Center: HP1 domain emerging from the membrane. Right: Both HP1 and HP2 domains emerging from the membrane. These emergences from the membrane are observed in some replicas, regardless of the system in CG simulations. For our analysis, we chose not to consider these instances. Other predictive tools, such as PSIPRED[41] or DREAMM,[42] indicate that this domain remains inside the membrane. When one or both domains emerge, they never re-enter the membrane during the simulation.



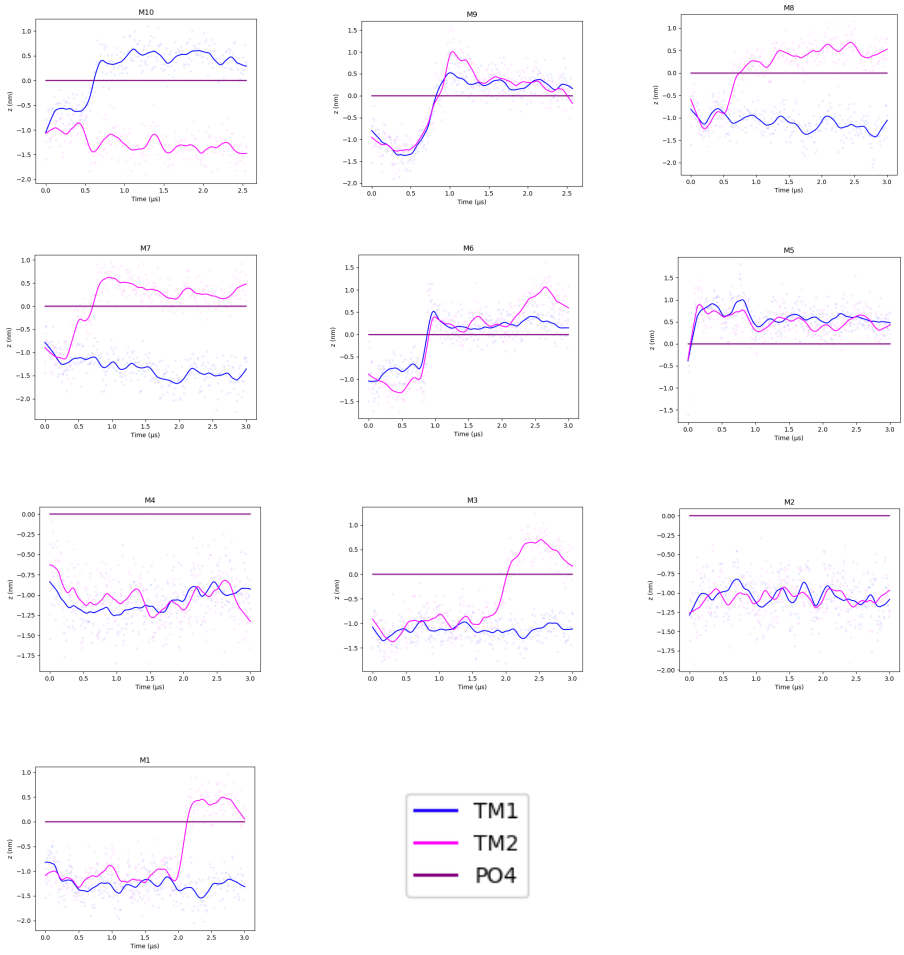
(a)

Figure 13: Distance between HP1 and HP2 domains throughout an AlphaFold CG simulation.



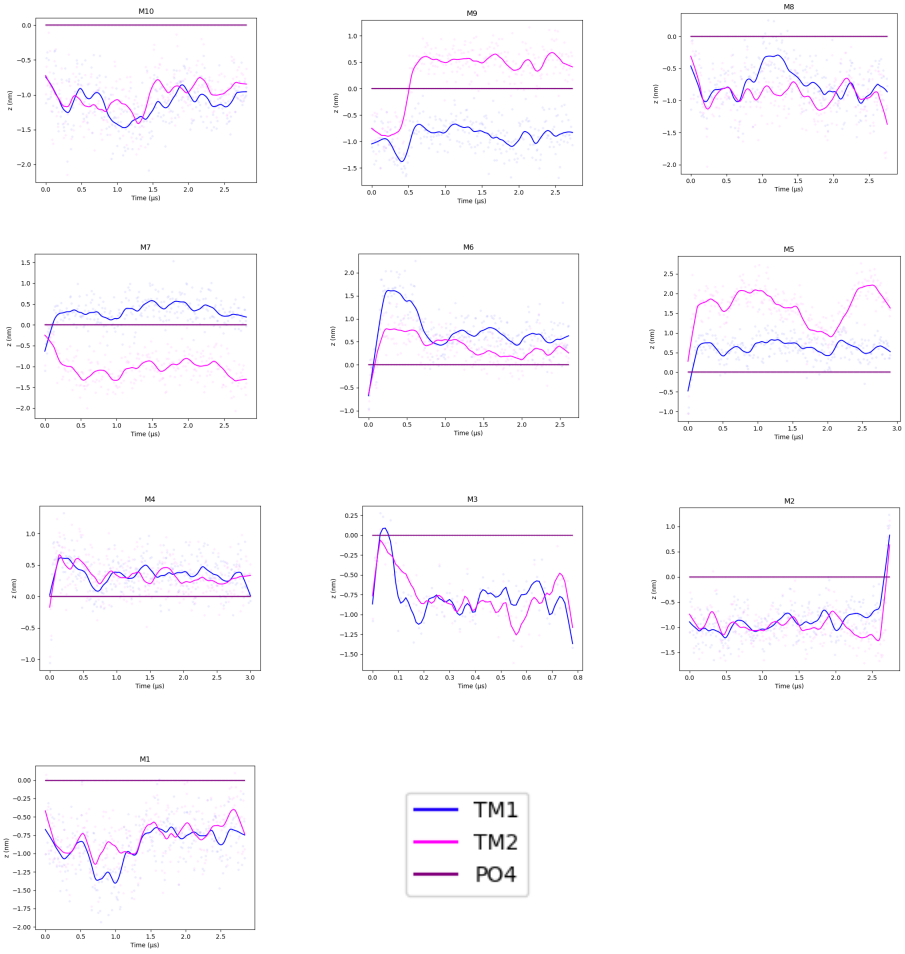
(a)

Figure 14: Distance of the center of mass along the Z-axis for the HP1 and HP2 domains of the AF model



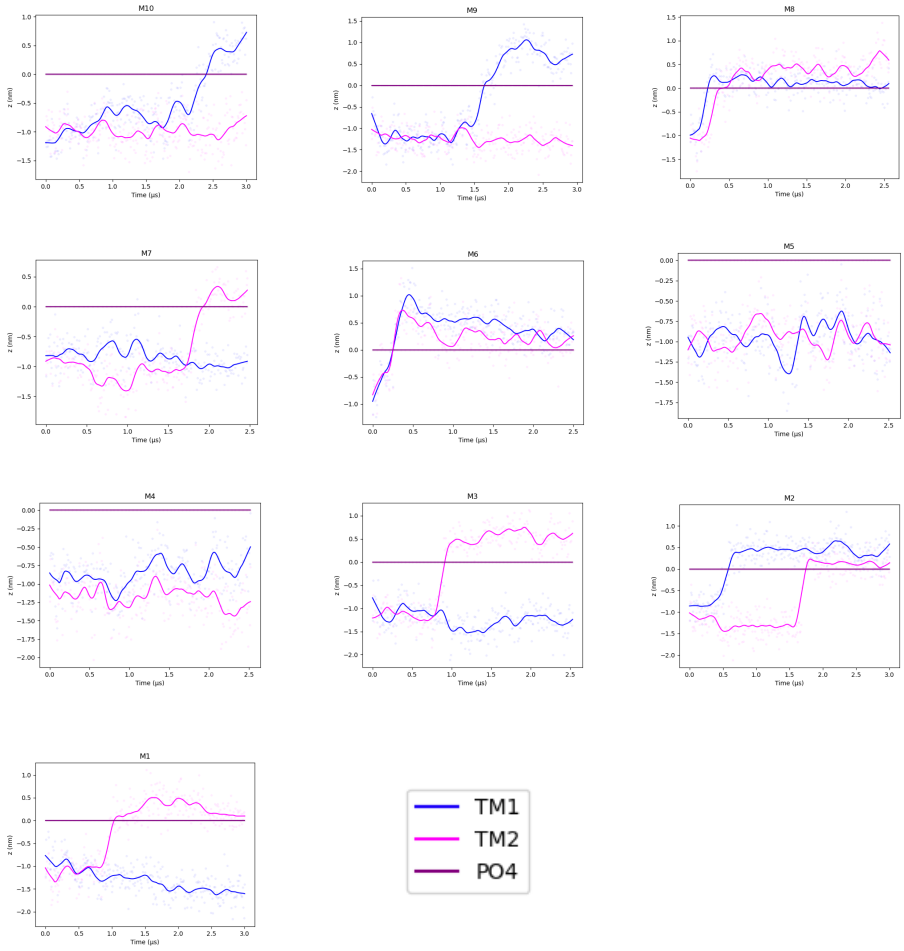
(a)

Figure 15: Distance of the center of mass along the Z-axis for the HP1 and HP2 domains of the AFM model



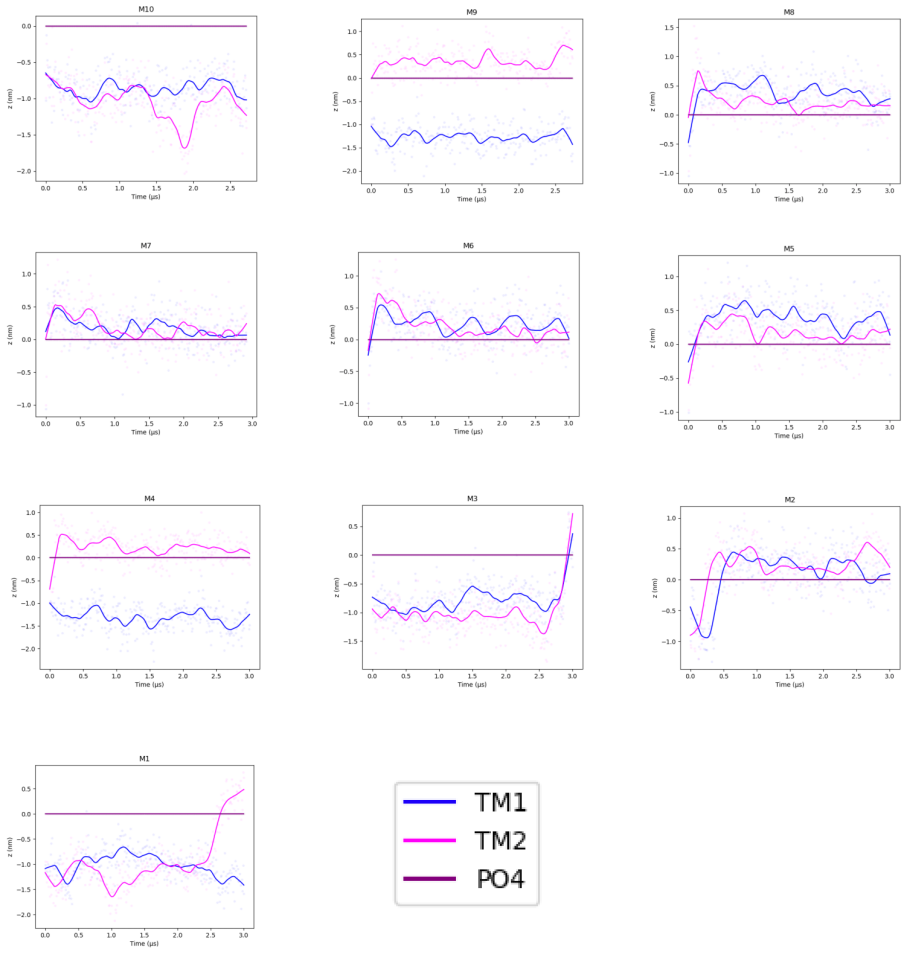
(a)

Figure 16: Distance of the center of mass along the Z-axis for the HP1 and HP2 domains of the RF model



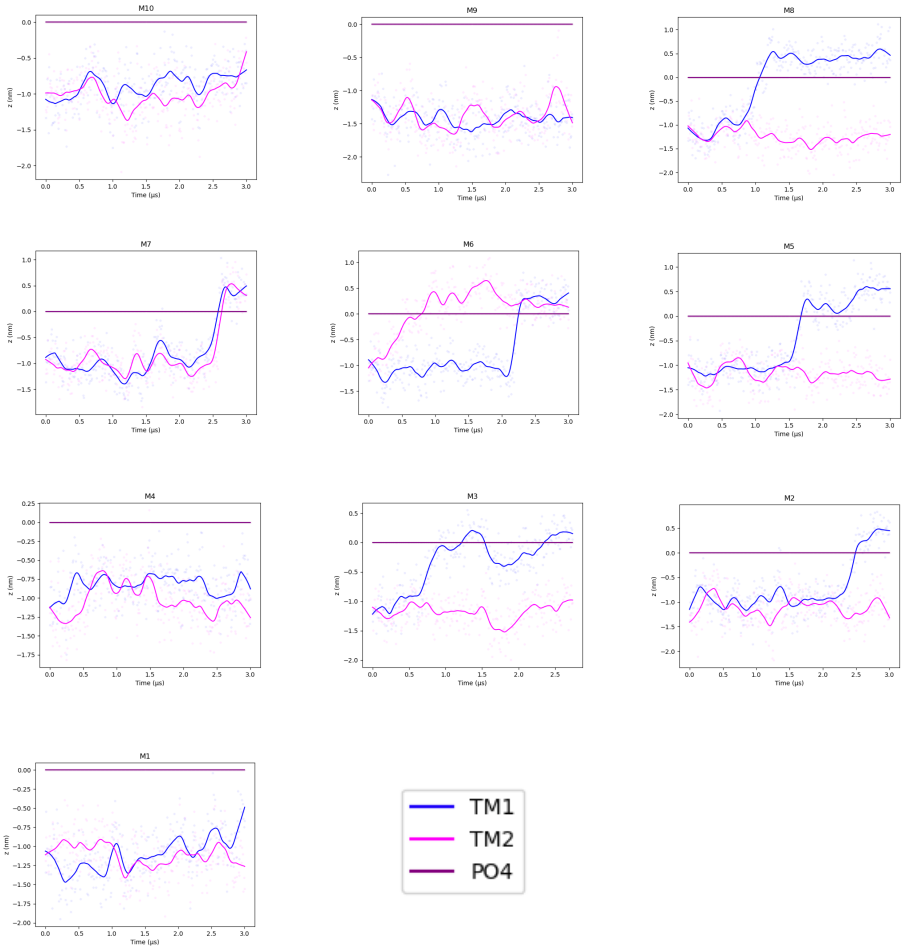
(a)

Figure 17: Distance of the center of mass along the Z-axis for the HP1 and HP2 domains of the TR model



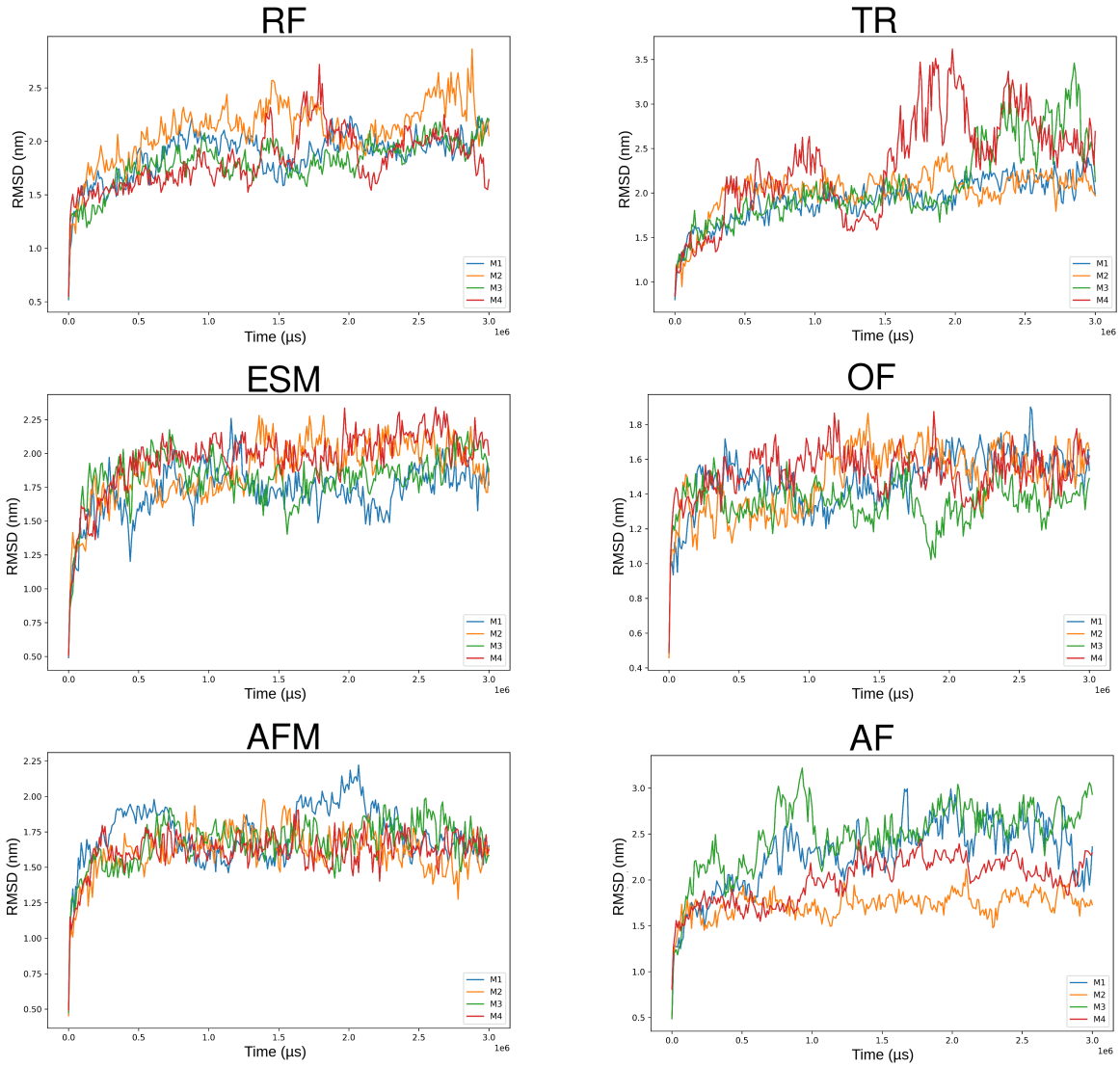
(a)

Figure 18: Distance of the center of mass along the Z-axis for the HP1 and HP2 domains of the OF model



(a)

Figure 19: Distance of the center of mass along the Z-axis for the HP1 and HP2 domains of the ESM model



(a)

Figure 20: RMSD for CG simulations.