Master 2 Physics Photonics Nanotechnology 2022-2023

Functional modes and thermal B-factor predictions for multigenic structural analysis predicted from Alpha Fold

Nicolas PETIOT

May 29, 2023

Supervisors: Dr. Adrien NICOLAÏ and Pr. Patrick SENET

NANOSCIENCES Department - Physics Applied to Proteins

Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS
Université de Bourgogne Franche-Comté / Faculté des Sciences et Techniques Mirande
9, Av. Savary - B.P. 47 870 21078. DIJON CEDEX - France

Abstract

Proteins are complex biomolecules that are critical to the functioning of living organisms. They are made up of sequences of amino-acids that fold into specific three-dimentional structures, which is highly related to their function. The process of protein folding is considered one of the most challenging problems in the field of biology and biochemistry, as it involves a delicate interplay of chemical and physical forces that determine the final shape of the protein. AlphaFold is a groundbreaking tool developped by DeepMind that uses artificial intelligence algorithms to pedict 3D structure of proteins based on their amino-acid sequence. This tool has the potential to revolutionize the study of enzymes as it provides fast and accurate way to predict the structure of molecules that have catalytic properties. The present work aim at using AlphaFold to study the catalytic properties of Glutathione Transferase (GST), especially from class δ and ε of *drosophilia melanogaster*, whith the ultimate goal of being able to design brand new sequences for enzymes with improved catalytic efficiancy.

In addition to using AlphaFold for the generation of 3D structures, molecular dynamics simulations can also be performed based on these structures. It allows to predict some of their potential behaviour and study various factors that may influence the function of the protein, such as thermal fluctuations or protein-ligand interactions. Simulationg these processes gives a deeper understanding of how proteins function and allows to identify areas for further investigation and improvements.

Contents

1 Introduction

Proteins are biological macromolecules that perform a large variety of functions in living cells comprising biochemical (enzymes), structural, mechanical, and signaling functions. They consist of chains composed of 20 different amino-acids. To perform their functions, proteins interact with small molecules referred to as ligands, which are able to bind to a protein with high affinity and specificity [1]. These protein/ligand interactions are crucial in biology, particularly in the context of drug design [2]. Since proteins interact with a broad range of drugs, it is of particular interest to study the mechanisms of binding of ligands to proteins and its impact on the structural dynamics to gain insights into (i) phenomena involved in the biological process and related to diseases [3] (misfolding, aggregation), and (ii) discovery, design, and development of new drugs [4]. The experimental structural data (e.g., X-ray crystallography, NMR, or cryo-EM) provide key structural information of the ligand-bound and ligand-unbound (APO) proteins [5]. Nevertheless, the static information is not always sufficient for understanding protein–ligand binding mechanisms, especially when pockets are highly flexible and contain several binding sites. Therefore, molecular dynamics (MD) and Normal Mode Analysis (NMA) are powerfull tools that provides a description of the dynamics and structures of protein–ligand systems with a high spatial and temporal resolution.

1.1 Glutathione Transferase

Glutathione transferases (GSTs) belong to a ubiquitous superfamily of enzymes that metabolize a broad range of reactive toxic compounds by catalyzing the conjugation of reduced tripeptide glutathione (γ-Glu-Cys-Gly; named GSH) to the electrophilic center of a second substrate [6–8], the reactivity of GSH being due to the thiol group SH of the cysteine residue. The conjugation reaction occurs spontaneously but GST accelerates it dramatically. This process of detoxification protects cells against damages caused by both exogenous and endogenous molecules. GSTs were first discovered in liver cells [9], and since then, they have been found to exhibit ligand-binding properties for a large variety of compounds, which are not always their enzymatic substrates [10]. Therefore, GSTs participate in diverse biological processes, making them multifunctional proteins. Moreover, GSTs are classified into three families according to their location in the cell: cytosolic, mitochondrial, and microsomal, which is not evolutively related to the two other classes [11]. First-discovered and most-abundant cytosolic GSTs are divided into 13 classes based on homology of their sequences.Members of the same cytosolic class have at least 40% of sequence identity, while members of different classes must have at most 25% of sequence identity. Even if they present a low homology with the cytosolic GST, mitochondrial GSTs can be considered as a particular class of GSTs (Kappa). Among the 42 GSTs identified in *Drosophila melanogaster*, δ and ε are the largest classes, with 25 members [12]. In their catalytic cycle, the GSH usually binds in a specific set of amino-acids called G-site and the hydrophobic xenobiotic in the so-called H-site. Interactions between insects and plant's chemicals lead to a major driving force in herbivorous insect evolution, hence this encourages the study of insect GSTs to understand how spontaneous mutations modify the stability, selectivity and the catalytic efficiencies of this enzyme superfamily.

1.2 AlphaFold

X-ray diffraction is a powerful experimental technique that have been used extensively to determine the three-dimentional structures of proteins. In this technique, a crystal of the protein is bombarded with X-rays, and the resulting diffraction pattern is used to determine the position of atoms within the protein. Over the years, X-ray diffraction experiments have played a pivotal role in determining the structures of tens of thousand of proteins, which are deposited in the Protein Data Bank (PDB). However, this experimental process can be time-consuming and technically challenging. Moreover, compared to the vast number of known protein sequences, the ensemble of solved structure is insignificant. In 2021, DeepMind [13] used machine learning approaches with the AlphaFold program. It uses computational models to predict the 3D structures of proteins based on it amino-acid's sequence with a high accuracy. In the field of de novo design of enzymes, AlphaFold has the potential to revolutionize the way we consider the design process, allowing to predict 3D structures that have not yet been experimentally characterized.

In addition to the initial AlphaFold program, DeepMind developped several other tools that have further expanded the capacities of protein structure predictions. One such tool is AlphaFold-multimer[14], which allows predictions for the structure of protein complexe such as homodimers. An other one is AlphaFill[15], which predict the positions of ligands, small molecules that bind to protein such as Glutathione. All together, these tools represent a major step forward in the field of protein study and will be at the root of the present work.

1.3 Goals

2 Materials and Methods

2.1 Multiple Sequence Alignment

Multiple sequence alignment (MSA) is a fundamental technique in bioinformatics used to compare and analyze the similarities and differences between multiple biological sequences. These sequences can be DNA, RNA, or protein sequences and can come from various species or different regions of the same genome. By aligning these sequences, it is possible to identify conserved regions that are important for function or evolution, as well as unique features that differentiate the sequences. In this study, we focus on a set of 25 GSTs sequences related to each other through evolution. Each sequence has a different length, making the alignment process especially useful. Through this analysis, we aim to identify regions of conservation and divergence between the sequences (see exemple below). Our first task was to use MSA to predict both the position of the interface of dimerization and the binding site of the set of GST sequence. The stability of the dimer structure is dependent on the interactions at this interface. Therefore, understanding the location and conservation of the dimer interface can provide insights into the stability of the dimer structure and the mechanisms of the biological activity. On the other hand, the binding site is the specific location on the enzyme where a substrate or ligand binds and interacts. The catalytic efficiency of the enzyme is related to the binding site because it determines the specificity and strength of the substrate-enzyme interaction. Therefore, understanding the location and conservation of the binding site is essential for elucidating the function and mechanism of action of the enzyme.

Each cell in the MSA matrix corresponds to a particular amino acid at a particular position in a particular sequence. We refer to the position in the MSA matrix as the MSA index. The MSA index allows us to compare the amino acid residues at each position across all the sequences in the alignment. We focused on highlighting residues that are known to be part of the dimer interface or binding site. By doing so, we can compare these residues across all sequences and identify any conserved or variable regions.

2.2 Anisotropic Network Model

As seen in the introduction, the AlphaFold program allows for the prediction of protein structures with remarkable accuracy. These structures have the potential to be used in a wide range of applications, including molecular dynamics simulations. In this work, we aim to use AlphaFold-predicted structures as input for the Anisotropic Network Model (ANM) to study protein dynamics. Specifically, we will compare the results obtained from the ANM simulations using AlphaFold structures with those obtained using experimentally determined structures. More precisely, we will use the ANM to predict the thermal B-factors of the studied structures[16], which are important indicators of protein flexibility and stability. The insights gained from this study will contribute to the ongoing efforts to develop computational tools for protein structure analysis and facilitate a deeper understanding of protein dynamics and function. To achieve these goals, it is necessary to provide a detailed description of the

ANM and its underlying mathematical principles. The ANM is a widely used method for studying the collective motions and dynamics of proteins based on their structure. It models the protein as a network of connected nodes and springs (representing covalent and non-covalent interactions between them).

2.2.1 Theory

Let $\vec{r_i}$ being the position of the node i and M_i it's mass. In the ANM, each node is assumed to be at the bottom of an harmonic potential, since interactions are modeled by connections between nodes, the force matrix is obtained by computing the mass-weighted Hessian matrix

$$\hat{H}_{ij} = -\frac{\Gamma_{ij}\gamma}{\sqrt{M_i M_j}} \frac{\vec{R_{ij}}\vec{R}_{ij}^T}{R_{ij}^2} \tag{2.1}$$

where γ is the spring constant used to model interactions between nodes, $\vec{R_{ij}} = \vec{R_j} - \vec{R_i}$ and Γ is the contact matrix. In the case were $i = j$, the force matrix is computed so that the self interacting term is the response to all the applied forces.

$$\hat{H}_{ii} = -\sum_{j \neq i} \hat{H}_{ij} \tag{2.2}$$

Γ_{ij} is the contact matrix that allow or not two nodes to be interacting with each other. The normal modes and eigenfrequencies are given by the diagonalization of the mass-weighted Hessian matrix.

$$\hat{H}\vec{e}_k = \tilde{\omega}_k^2 \vec{e}_k \tag{2.3}$$

It is important to make some remarks at this point, first in the equation (2.1) $\hat{H}_{ij}$ is actually a three by three matrix. It means that the ovearall $\hat{H}$ will be a square matrix of dimention $d = 3N$ with N the number of considered nodes. Since the diagonalization algorythms have a complexity of $O(d^3)$, it means that the computation time will highly depends on the choosen set of nodes. Second, the masses of the nodes are expressed in g.mol^{-1} to avoid numerical errors in the diagonalization. Expressing γ in kcal.mol^{-1}.Å^{-2}, it means that the eigenvalues $\tilde{\omega}_k^2$ are expressed in kcal.g^{-1}.Å^{-2} which is directly proportional to a squared frequency $\omega^2 = 4.184 \times 10^{26}$ s^{-2}. Thermal B-factors of the node i is directly proportionel to the mean squared fluctuations of the node's position $\sigma^2(\vec{R_i})$ and can be computed from the normal modes using the following

$$\beta_i = \frac{8\pi^2}{3} \frac{k_B T}{M_i} \sum_k \frac{|\vec{e}_{k,i}|^2}{\tilde{\omega}_k^2} \tag{2.4}$$

where $\vec{e}_{k,i}$ contains the elements of $\vec{e}_k$ related to the node i. β_i is expressed in m^2 but will always be converted in ^{2} because of conventions and usual order of magnitude.

Ever since the beginning, we were talking in a very abstract way about "nodes". In this work, we did consider the amino-acid's center of mass (COM). When available, we compared the predicted values of B-factors with X-ray based measurement in term of pearson correlation.

$$\mathcal{R} = \frac{\sum_i (\beta_i - <\beta>_i)(B_i - <B>_i)}{\sqrt{\sum_i (\beta_i - <\beta>_i)^2} \sqrt{\sum_j (B_j - <B>_j)^2}} \tag{2.5}$$

The values of $\mathcal{R}$ are between -1 and 1 and gives the linear correlation between predicted and measured B-factors. A coefficient of 1 is associated to a perfect correlation wherease a coefficient of 0 means that there is basically no links between prediction and experiment. Eventually, a negative value of $\mathcal{R}$ means that there is a correlation with opposite sign and would be interpreted as a result even worst than no correlation. Note that we need to consider the same ensemble of points for β_i and B_i.

The thermal fluctuation of the nodes is not the only informations one can get from normal modes, indeed fluctuations of the $\vec{R}_{ij}$ vector can also be an interesting metric to analyse as it represents the relative motion of nodes within a given mode.

$$d_{ij} = k_B T \sum_k \frac{1}{\omega_k^2} \left(\frac{\vec{e}_{k,j}}{\sqrt{M_j}} - \frac{\vec{e}_{k,i}}{\sqrt{M_i}} \right)^2 \tag{2.6}$$

Such metric is especially interesting to study as it provide insights about the fluctuations of a mode with respect to another and allow to identify pairs of nodes involved in specific functional modes.

2.2.2 Parametrization

In the Anisortopic network model as presented above, we consider two parameters namely the contact matrix Γ_{ij} and the spring constant γ. In order to get predictions that are physically relevant, it is necessary to compute the good parameters. Let us first consider the contact matrix. Given two nodes i and j, we will consider $\Gamma_{ij} = 1$ if the distance $|\vec{R}_{ij}|$ is smaller than the cut-off R_c. It is then necessary to find a good numerical value for R_c. We know that for high distances, the interactions between the nodes are supposed to be small if not nulls. To add this kind of interactions, we look for R_c as small as possible. We also know that the eigenvalues are associated to frequencies for the collectives modes. The six firsts modes should be global transations and rotations with a null frequency associated. In the case where R_c is too small, some other modes will have null eigenfrequencies with no physical justifications. R_c is then the smallest value that gives exactly 6 null eigenvalues for the Hessian. Eventually, in the equation (2.1) γ is a constant, it means that $\tilde{\omega}_k^2 \propto \gamma$. From the equation (2.4), one can show that β_i is inversly proportional to γ which is just a scaling factor. It can then be computed from comparitions with experimental measurments of thermal B-factors using least squared methods.

2.3 Molecular Dynamics

Molecular dynamics (MD) has emerged as a powerful computational tool for simulating molecular systems with exceptional precision. By numerically integrating the equations of motion for atoms and molecules, MD provides detailed information about the dynamic behavior and interactions within these systems. With its ability to capture the atomic-level motion and thermodynamics, MD offers insights into the structural changes, energetics, and properties of molecules under various conditions. Compared with ANM, MD stands out for its superior precision in capturing the dynamic behavior of molecular systems. For the purpose of this work, the amount of system to be studied is far too big (25 structures ×3 states in the catalytic cycle) and cannot be considered as an option. But considering it's accuracy, one can consider a sample of GSTs that will be simulated via MD.
MD consider interactions between atoms in a large variety of form. First, non-bonded interactions are considered with a lehnard-jones potential and electrostatic interactions.

$$V_{\text{non-bonded}}(\vec{r}_i) = \sum_j 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{R_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}} \right)^6 \right] + \sum_j \frac{q_i q_j}{4\pi\epsilon_0 R_{ij}} \tag{2.7}$$

Interactions between bonded atoms are also described by the following.

$$V_{\text{bonded}}(\vec{r}_i) = \sum_{\text{bonds}} K_l (l - l_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{tortions}} K_\varphi [1 + \cos(n\varphi + \delta)] \tag{2.8}$$

The ensemble of parameters for those interactions are optimized and used by MD algorithms by the so called force fields. Since the potential consider interactions between all the atoms of the system, this kind of simulation is called All Atoms MD, in opposition to coarse grained MD that we wont discuss here. Finally forces of interactions are obtained taking the gradient of the potential and the simulation is achieved by integration of the Newton's equation :

$$m \frac{d^2 \vec{r}_i}{dt^2} = -\vec{\nabla} V(\vec{r}_i) \tag{2.9}$$

From the time serie obtained, one can compute the associated Thermal B-Factors as mean squared fluctuations of the atom's position.

$$\beta_i = \frac{8\pi^2}{3}\left(\left\langle \vec{R}_i^2 \right\rangle - \left\langle \vec{R}_i \right\rangle^2\right) \tag{2.10}$$

Comparisons between such factors computed from MD and from ANM gives again extra informations about the precision of the considered models.

3 Results and Discussion

3.1 Sequences and Static Structures

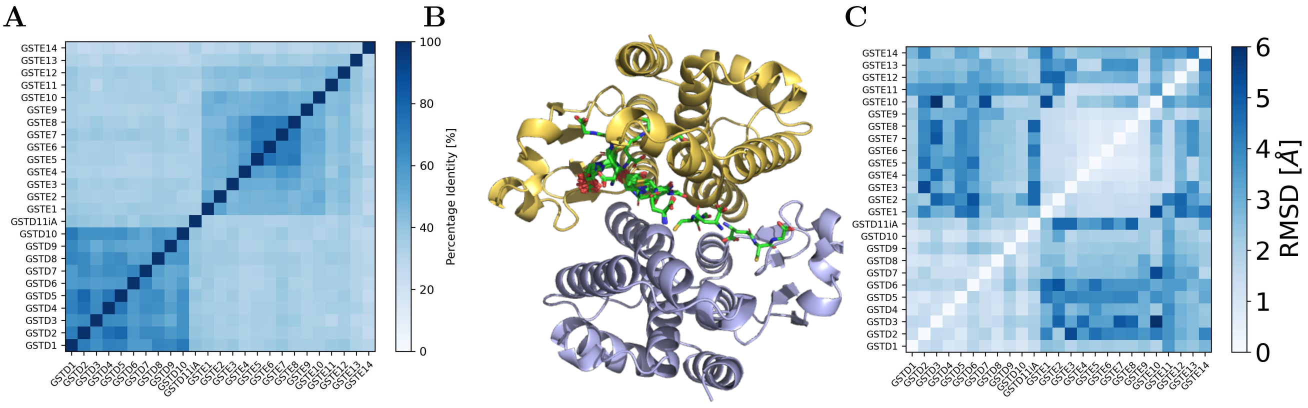
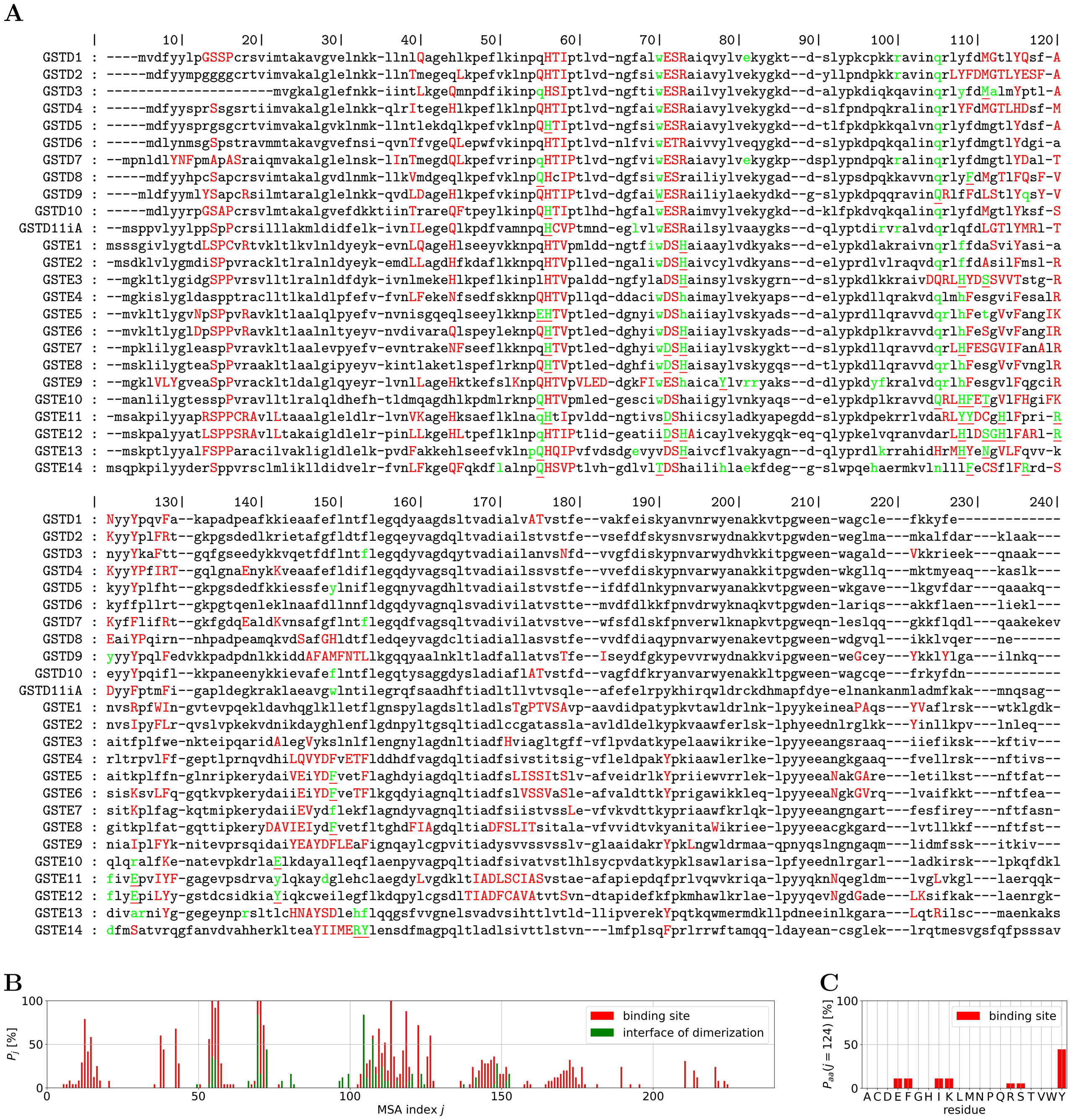


Figure 3.1: Sequences and structures associated to the selected *Drosophila Melanogaster*'s GSTs

The first step of our methodology was to study the sequences of the GSTs of interest. We computed their Multiple Sequence Alignment as well as the percent identity matrix associated (see Fig. 3.1 pannel A). This allows to identify regions of high/low conservation as well as clusters of identity. For instance, it is clearly visible that among the class δ, the GSTs are self similar from a sequence point of view. In contradiction, it seems that among the class ε, the GSTs E5, E6, E7, E8 are self similar but the other ones seems much more different. The sequences were then used as a base for the AlphaFold program to predict the 3D structure (pannel B) and pairs of structures were compared using the Root Mean Squared Deviation (i.e. geometrical distances between atoms, pannel C) and once again in the associated matrix, one can identify two main clusters for class δ and ε but here the E10 gives much higher values than we might have expect. This is due to a much longer sequence in the terminal part that make the RMSD higher.

The information about the structures can be completed by an information about the position of the Glutathione. As mantioned earlier, the program AlphaFill allows to make such precisions and this gives 40 different positions of Glutathion-like ligands in the GSTD1 (i.e. ligands that are chemically close to Glutathione), those positions are represented in the 3D structure (pannel B). Distances between atoms allowed us to determine from these data the residues that are involved in the protein-ligand binding as well as the dimerization of the structure. Indeed, two atoms that are closer than 3Å were considered as in contact. This information can be computed for all 25 structures and projected on the MSA matrix computed before. This gives the following representation (see Fig. 3.1), where we computed the probability of a given residue to be in the binding site or in the interface of dimerization. From this information, we are able to compute the conservation of any residue in the binding site / interface of dimerization. Here, we give an illustration in the case of the residue 124, which have a high degree of conservation among all the studied GSTs and have been identified as a part of the binding site in %.



```
           |        10|       20|       30|       40|       50|       60|       70|       80|       90|      100|      110|      120|
   GSTD1 : ----mvdfyylpGSSPcrsvimtakavgvelnkk-llnlQagehlkpeflkinpqHTIptlvd-ngfalwESRaiqvylvekygkt--d-slypkcpkkravinqrlyfdMGtlYQsf-A
   GSTD2 : -----mdfyympggggcrtvimvakalglelnkk-llnTmegeqLkpefvklnpQHTIptlvd-ngfsiwESRaiavylvekygkd--d-yllpndpkkravinqrLYFDMGTLYESF-A
   GSTD3 : ---------------------mvgkalglefnkk-iintLkgeQmnpdfikinpqHSIptlvd-ngftiwESRailvylvekygkd--d-alypkdiqkqavinqrlyfdMalmYptl-A
   GSTD4 : -----mdfyysprSsgsrtiimvakalglelnkk-qlrItegeHlkpeflklnpQHTIptlvd-ngfaiwESRaiavylvekygkd--d-slfpndpqkralinqrlYFdMGTLHDsf-M
   GSTD5 : -----mdfyysprgsgcrtvimvakalgvklnmk-llntlekdqlkpefvklnpQHTIptlvd-ngfsiwESRaiavylvekygkd--d-tlfpkdpkkqalvnqrlyfdmgtlYdsf-A
   GSTD6 : -----mdlynmsgSpstravmmtakavgvefnsi-qvnTfvgeQLepwfvkinpqHTIptlvd-nlfviwETRaivvylveqygkd--d-slypkdpqkqalinqrlyfdmgtlYdgi-a
   GSTD7 : --mpnldlYNFpmApASraiqmvakalglelnsk-lInTmegdQLkpefvrinpqHTIPtlvd-ngfviwESRaiavylvekygkp--dsplypndpqkralinqrlyfdmgtlYDal-T
   GSTD8 : -----mdfyyhpcSapcrsvimtakalgvdlnmk-llkVmdgeqlkpefvklnpQHcIPtlvd-dgfsiwESrailiylvekygad--d-slypsdpqkkavvnqrlyFdMgTlFQsF-V
   GSTD9 : ----mldfyymlYSapcRsilmtaralglelnkk-qvdLDageHlkpefvkinpQHTIPtlvd-dgfaiWESRailiylaekydkd--g-slypkdpqqravinQRlfFdLStlYqsY-V
  GSTD10 : -----mdlyyrpGSAPcrsvlmtakalgvefdkktiinTrareQFtpeylkinpQHTIptlhd-hgfalwESRaimvylvekygkd--d-klfpkdvqkqalinqrlyfdMgtlYksf-S
GSTD11iA : --msppvlyylppSpPcrsilllakmldidfelk-ivnILegeQlkpdfvamnpqHCVPtmnd-eglvlwESRailsylvaaygks--d-qlyptdirvralvdqrlqfdLGTlYMRl-T
   GSTE1 : msssgivlygtdLSPCvRtvkltlkvlnldyeyk-evnLQageHlseeyvkknpqHTVpmldd-ngtfiwDSHaiaaylvdkyaks--d-elypkdlakraivnqrlffdaSviYasi-a
   GSTE2 : -msdklvlygmdiSPpvrackltlralnldyeyk-emdLLagdHfkdaflkknpqHTVplled-ngaliwDSHaivcylvdkyans--d-elyprdlvlraqvdqrlffdAsilFmsl-R
   GSTE3 : --mgkltlygidgSPPvrsvlltlralnldfdyk-ivnlmekeHlkpeflkinplHTVpaldd-ngfylaDSHainsylvskygrn--d-slypkdlkkraivDQRLHYDSSVVTstg-R
   GSTE4 : --mgkislygldaspptraclltlkaldlpfefv-fvnLFekeNfsedfskknpQHTVpllqd-ddaciwDShaimaylvekyaps--d-elypkdllqrakvdqlmhFesgviFesalR
   GSTE5 : --mvkltlygvNpSPpvRavkltlaalqlpyefv-nvnisgqeqlseeylkknpEHTVptled-dgnyiwDShaiiaylvskyads--d-alyprdllqravvdqrlhFetgVvFangIK
   GSTE6 : --mvkltlyglDpSPPvRavkltlaalnltyeyv-nvdivaraQlspeyleknpQHTVptled-dghyiwDSHaiiaylvskyads--d-alypkdplkravvdqrlhFeSgVvFangIR
   GSTE7 : --mpklilygleaspPvravkltlaalevpyefv-evntrakeNFseeflkknpqHTVptled-dghyiwDSHaiiaylvskygkt--d-slypkdllqravvdqrLHFESGVIFanAlR
   GSTE8 : --msklilygteaSpPvraakltlaalgipyeyv-kintlaketlspeflrknpQHTVptled-dghfiwDSHaisaylvskygqs--d-tlypkdllqravvdqrlhFesgVvFvnglR
   GSTE9 : --mgklVLYgveaSpPvrackltldalglqyeyr-lvnlLageHktkefslKnpQHTVpVLED-dgkFIwEShaicaYlvrryaks--d-dlypkdyfkralvdqrlhFesgvlFqgciR
  GSTE10 : --manlilygtesspPvravlltlralqldhefh-tldmqagdhlkpdmlrknpQHTVpmled-gesciwDShaiigylvnkyaqs--d-elypkdplkravvdQRLHFETgVlFHgiFK
  GSTE11 : -msakpilyyapRSPPCRAvlLtaaalgleldlr-lvnVKageHksaeflklnaqHtIpvldd-ngtivsDShiicsyladkyapegdd-slypkdpekrrlvdaRLYYDCgHlFpri-R
  GSTE12 : --mskpalyyatLSPPSRAvlLtakaigldlelr-pinLLkgeHLtpeflklnpqHTIPtlid-geatiiDSHAicaylvekygqk-eq-qlypkelvqranvdarLHlDSGHlFARl-R
  GSTE13 : --mskptlyyalFSPParacilvakligldlelk-pvdFakkehlseefvklnpQHQIPvfvdsdgevyvDSHaivcflvakyagn--d-qlyprdlkrrahidHrMHYeNgVLFqvv-k
  GSTE14 : msqpkpilyyderSppvrsclmlikllldidvelr-fvnLFkgeQFqkdflalnpQHSVPtlvh-gdlvlTDShailihlaekfdeg--g-slwpqehaermkvlnlllFeCSflFRrd-S

           |       130|      140|      150|      160|      170|      180|      190|      200|      210|      220|      230|      240|
   GSTD1 : NyyYpqvFa--kapadpeafkkieaafeflntflegqdyaagdsltvadialvATvstfe--vakfeiskyanvnrwyenakkvtpgween-wagcle---fkkyfe------------
   GSTD2 : KyyYplFRt--gkpgsdedlkrietafgfldtflegqeyvagdqltvadiailstvstfe--vsefdfskysnvsrwydnakkvtpgwden-weglma---mkalfdar---klaak---
   GSTD3 : nyyYkaFtt--gqfgseedykkvqetfdflntflegqdyvagdqytvadiailanvsNfd--vvgfdiskypnvarwydhvkkitpgween-wagald---Vkkrieek---qnaak---
   GSTD4 : KyyYPfIRT--gqlgnaEnykKveaafefldiflegqdyvagsqltvadiailssvstfe--vvefdiskypnvarwyanakkitpgwden-wkgllq---mktmyeaq---kaslk---
   GSTD5 : kyyYplfht--gkpgsdedfkkiessfeylniflegqnyvagdhltvadiailstvstfe--ifdfdlnkypnvarwyanakkvtpgween-wkgave---lkgvfdar---qaaakq--
   GSTD6 : kyffpllrt--gkpgtqenleklnaafdllnnfldgqdyvagnqlsvadivilatvstte--mvdfdlkkfpnvdrwyknaqkvtpgwden-lariqs---akkflaen---liekl---
   GSTD7 : KyfFlifRt--gkfgdqEaldKvnsafgflntflegqdfvagsqltvadivilatvstve--wfsfdlskfpnverwlknapkvtpgweqn-leslqq---gkkflqdl---qaakekev
   GSTD8 : EaiYPqirn--nhpadpeamqkvdSafGHldtfledqeyvagdcltiadiallasvstfe--vvdfdiaqypnvarwyenakevtpgween-wdgvql---ikklvqer---ne------
   GSTD9 : yyyYpqlFedvkkpadpdnlkkiddAFAMFNTLlkgqqyaalnkltladfallatvsTfe--Iseydfgkypevvrwydnakkvipgween-weGcey---YkklYlga---ilnkq---
  GSTD10 : eyyYpqifl--kkpaneenykkievafeflntflegqtysaggdysladiaflATvstfd--vagfdfkryanvarwyenakkltpgween-wagcqe---frkyfdn-----------
GSTD11iA : DyyFptmFi--gapldegkraklaeavgwlntilegrqfsaadhftiadltllvtvsqle--afefelrpykhirqwldrckdhmapfdye-elnankanmladmfkak---mnqsag--
   GSTE1 : nvsRpfWIn-gvtevpqekldavhqglklletflgnspylagdsltladlsTgPTVSAvp-aavdidpatypkvtawldrlnk-lpyykeineaPAqs---YVaflrsk---wtklgdk-
   GSTE2 : nvsIpyFLr-qvslvpkekvdnikdayghlenflgdnpyltgsqltiadlccgatassla-avldldelkypkvaawferlsk-lphyeednlrglkk---Yinllkpv---lnleq---
   GSTE3 : aitfplfwe-nkteipqaridAlegVykslnlflengnylagdnltiadfHviagltgff-vflpvdatkypelaawikrike-lpyyeeangsraaq---iiefiksk---kftiv---
   GSTE4 : rltrpvlFf-geptlprnqvdhiLQVYDFvETFlddhdfvagdqltiadfsivstitsig-vfleldpakYpkiaawlerlke-lpyyeeangkgaaq---fvellrsk---nftivs--
   GSTE5 : aitkplffn-glnripkerydaiVEiYDFvetFlaghdyiagdqltiadfsLISSItSlv-afveidrlkYpriiewvrrlek-lpyyeeaNakGAre---letilkst---nftfat--
   GSTE6 : sisKsvLFq-gqtkvpkerydaiiEiYDFveTFlkgqdyiagnqltiadfslVSSVaSle-afvaldttkYprigawikkleq-lpyyeeaNgkGVrq---lvaifkkt---nftfea--
   GSTE7 : sitKplfag-kqtmipkerydaiiEVydflekflagndyvagnqltiadfsiistvssLe-vfvkvdttkypriaawfkrlqk-lpyyeeangngart---fesfirey---nftfasn-
   GSTE8 : gitkplfat-gqttipkeryDAVIEIydFvetfltghdFIAgdqltiaDFSLITsitala-vfvvidtvkyanitaWikriee-lpyyeeacgkgard---lvtllkkf---nftfst--
   GSTE9 : niaIplFYk-nitevprsqidaiYEAYDFLEaFignqaylcgpvitiadysvvssvsslv-glaaidakrYpkLngwldrmaa-qpnyqslngngaqm---lidmfssk---itkiv---
  GSTE10 : qlqralfKe-natevpkdrlaElkdayalleqflaenpyvagpqltiadfsivatvstlhlsycpvdatkypklsawlarisa-lpfyeednlrgarl---ladkirsk---lpkqfdkl
  GSTE11 : fivEpvIYF-gagevpsdrvaylqkaydglehclaegdyLvgdkltIADLSCIASvstae-afapiepdqfprlvqwvkriqa-lpyyqknNqegldm---lvgLvkgl---laerqqk-
  GSTE12 : flyEpiLYy-gstdcsidkiaYiqkcweilegflkdqpylcgsdlTIADFCAVAtvtSvn-dtapidefkfpkmhawlkrlae-lpyyqevNgdGade---LKsifkak---laenrgk-
  GSTE13 : divarniYg-gegeynprsltlcHNAYSDlehflqqgsfvvgnelsvadvsihttlvtld-llipverekYpqtkqwmermdkllpdneeinlkgara---LqtRilsc---maenkaks
  GSTE14 : dfmSatvrqgfanvdvahherklteaYIIMERYlensdfmagpqltladlsivttlstvn---lmfplsqFprlrrwftamqq-ldayean-csglek---lrqtmesvgsfqfpsssav
```

Figure 3.2: Indentification of residues in the binding site and of the interface of dimerization

3.2 Dynamics from Normal Modes

As explained in the introduction, this present work not only cares about the informations that have been extracted from the static predictions of the AlphaFold and AlphaFill programs but also about the dynamics of the dimers. In this section we will present the next step of our methodology with the Anisotropic Network Model, starting with the parametrization.

3.2.1 Parametrization

The parametrization step is needed to make sure that the predictions of the model are physically relevant. From the amino-acid's center of mass, it is very simple to compute the mass-weighted Hessian (see eq. 2.1). As presented before, a first step is to make sure that the cut-off R_c is such that the eigenvalues of the Hessian are non-null. Taking the exemple of the GSTD1, we computed $\tilde{\omega}_k^2(R_c)$ for the modes 5, 6, 7 and 8.

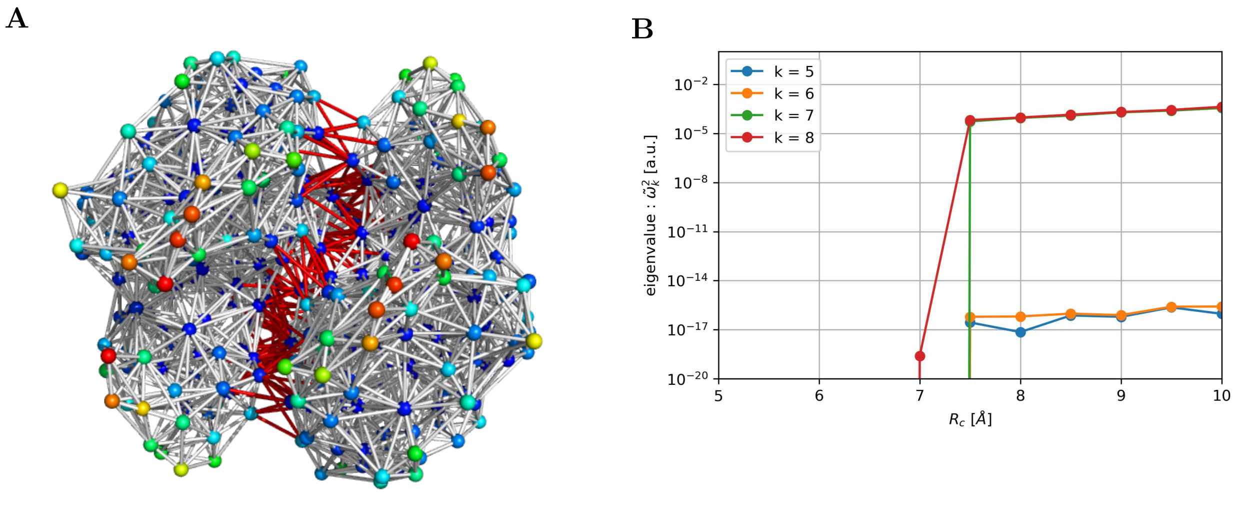


Figure 3.3: Coarse grained representation & cut-off parametrisation of GSTD1 structure

In the figure 3.2.1, it is clearly visible that for $R_c = 7.5$, the eigenvalues for $k \geq 6$ are no longer nulls. It is then convenient to use this value of cut-off for this structure. Note that later, such computations will be performed for all 25 structures in order to have a correct representation of the structures' topology. Those eigenvalues are computed for $\gamma = 1$ kcal.mol^{-1}.Å^{-2}, but as we have seen before, $\tilde{\omega}_k^2 \propto \gamma$. It is now time to compute the thermal B-factors to be able to compute the optimal γ value.

3.2.2 Predictions

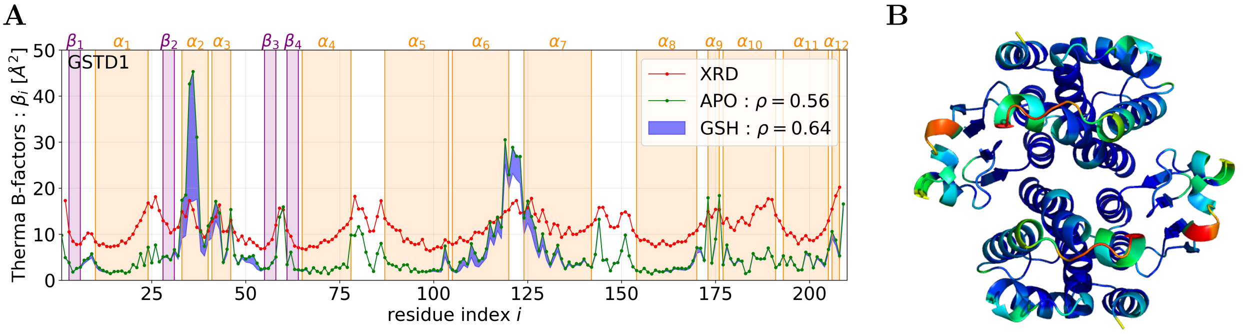


Figure 3.4: Thermal B-factors predicted from ANM-COM and comparisions with XRD measurments

The GSTD1 structure already have been studied and a X-ray diffraction based measurment of the Thermal B-factors have been done. This curve is represented in red in the figure (3.2.2). The predictions of the ANM-COM are done first taking $R_c = 7.5$Å and $\gamma = 1$kcal.mol^{-1}.Å^{-2}, then using least squared methods to fit the predictions on the measurments, one can compute a γ value of 14kcal.mol^{-1}.Å^{-2}. Eventually, it is possible to add 6 nodes that corresponds to the amino-acids of the GSHs predicted from AlphaFill. In the case of the GSTD1, the program gives 40 different positions. We computed the thermal B-factors for those 40 structures (protein + ligands) and plotted the range of B-factors predicted in blue in the figure (3.2.2). From these results, there is several remarks to be done. First in the APO configuration, one can clearly see that the position of the predicted spikes matches the position of the XRD measurments, even though their relative amplitude doesn't seems to be perfect. The computation of the pearson correlation gives 56%. Then, taking into account the GSHs, it appears that in the regions where the ANM gives the highest predicted B-factors, the ligands introduce a bigger rigidity and reduces the predictions in the regions of index $33-52$ and $103-138$. The computed pearson correlation for the GSH was achieved taking for each residue the minimal value of B-factors within the range of 40 predictions and gives an increase of 8% comared to the APO configuration. It is also possible to look at the conservation of the residues in this range the same way we did in the previous section. Let us for instances consider the residue 34. Projected on the MSA matrix, this gives an index of 39 which gives the following histogram.

For this specific index, the GSTs usually have either Leucines or Tyrosines amino-acids with a high probability or Isoleucine, Phenylalanine, Methionine or Valine with much smaller probability. All those residues have hydrophobic side chains and the main difference that can be done is the presence or abscence of aromatic rings in the amino-acid's side chain.

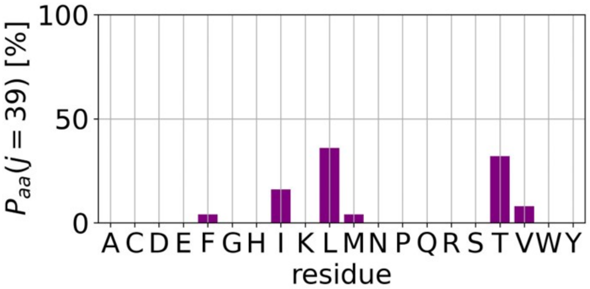


Figure 3.5: Amin-acid conservation for the 39th residue of the MSA matrix

So far, we have seen a methodology that allows, starting from a protein sequence to determine the associated structure and flexibility using ANM. It is now possible to extend this methodology to the set of 25 GSTs. The parameter γ is set to be 14 kcal.mol^{-1}.Å^{-2} and the R_c cut-off will be systematically computed for each structure. Eventually, using a projection on the MSA matrix, it is possible to build the following representation.

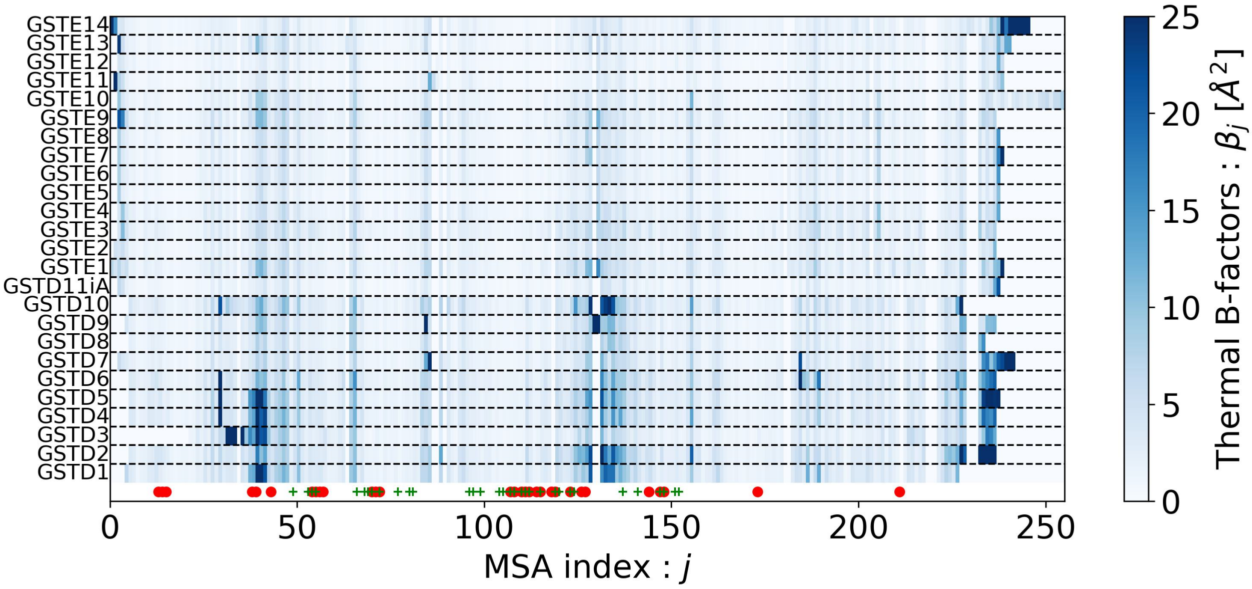


Figure 3.6: ANM-COM prediction for the thermal B-factors of the 25 GSTs considered

3.3 Dynamics from Molecular Dynamics

3.4 Comparison between Structures

Bibliography

(1) Du, X.; Li, Y.; Xia, Y.-L.; Ai, S.-M.; Liang, J.; Sang, P.; Ji, X.-L.; Liu, S.-Q. *International journal of molecular sciences* **2016**, *17*, 144.

(2) Li, L.; Koh, C. C.; Reker, D.; Brown, J.; Wang, H.; Lee, N. K.; Liow, H.-h.; Dai, H.; Fan, H.-M.; Chen, L., et al. *Scientific reports* **2019**, *9*, 7703.

(3) Silva, J. L.; Vieira, T. C.; Gomes, M. P.; Bom, A. P. A.; Lima, L. M. T.; Freitas, M. S.; Ishimaru, D.; Cordeiro, Y.; Foguel, D. *Accounts of chemical research* **2010**, *43*, 271–279.

(4) Payandeh, J.; Volgraf, M. *Nature Reviews Drug Discovery* **2021**, *20*, 710–722.

(5) Chakraborti, S.; Hatti, K.; Srinivasan, N. *International Journal of Molecular Sciences* **2021**, *22*, 6830.

(6) Mannervik, B. *Adv Enzymol Relat Areas Mol Biol* **1985**, *57*, 357–417.

(7) Armstrong, R. N. *Chemical research in toxicology* **1997**, *10*, 2–18.

(8) Hayes, J. D.; Flanagan, J. U.; Jowsey, I. R. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 51–88.

(9) Combes, B.; Stakelum, G. S., et al. *The journal of clinical investigation* **1961**, *40*, 981–988.

(10) Axarli, I. A.; Rigden, D. J.; Labrou, N. E. *Biochemical Journal* **2004**, *382*, 885–893.

(11) Oakley, A. *Drug metabolism reviews* **2011**, *43*, 138–151.

(12) Gonis, E.; Fraichard, S.; Chertemps, T.; Hecker, A.; Schwartz, M.; Canon, F.; Neiers, F. *Insects* **2022**, *13*, 612.

(13) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. *Nature* **2021**, *596*, 583–589.

(14) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J., et al. *BioRxiv* **2021**, 2021–10.

(15) Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A. *Nature Methods* **2022**, 1–9.

(16) Atilgan, A. R.; Durell, S.; Jernigan, R. L.; Demirel, M. C.; Keskin, O; Bahar, I. *Biophysical journal* **2001**, *80*, 505–515.