UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Creation, refinement, and evaluation of conformational ensembles of proteins using the Torsional Network Model.

Máster Universitario en Bioinformática y Biología Computacional

Autor: Roncero Moroño, David Alejandro

Director: DEHOUCK, YVES

Segundo director: BASTOLLA, UGO

Ponente: MARTÍNEZ, GONZALO

Departamento de Informática

02/08/21

Universidad Autónoma de Madrid
ESCUELA POLITÉCNICA SUPERIOR

Presented by **David Alejandro Roncero Moroño**, student of the Master in Bioinformatics and Computational Biology .

This work has been carried out at the Bioinformatics department of *Centro de Biología Molecular Severo Ochoa* (CBM), which belongs to *Consejo Superior de Investigaciones Científicas* (CSIC).

Index

Abbreviations

Torsional Network Model	TNM
Nuclear Magnetic Resonance	NMR
Three dimensional	3D
Mass spectrometry	MS
Molecular dynamics	MD
All-atom Molecular Dynamics	AMD
Elastic Network Models	*D-loop*
Normal Mode Analysis	NMA
Principal Component Analysis	PCA
Gaussian Network Model	GNM
Anisotropic Network Model	ANM
Protein Data Bank	PDB
Artificial Neural Networks	ANNs
Root Mean Square Error	RMSD
Molprobity Score	MPscore
Angstroms	Å
Ramachandran percentage outside favored region	Rama_iffy
Rotamer percentage outliers	Rota_out
Clashscore	CS
Energy Minimization	EM
Optimized Potentials for Liquid Simulations	OPLS
Carbon alfa	Cα
Carbon beta	Cβ
Torsional Network Model Ensemble	TNM Ens
SIDEpro-improved Ensemble	SIDEpro Ens
Reference structures	Ref
Ensemble structures	Ens

1. Summary

One of the main limitations of structural bioinformatics lies in the difficulty of properly accounting for the dynamical aspects of proteins, which are often critical to their functional mechanisms. Among the tools developed to deal with this issue, the Torsional Network Model (TNM) relies on internal degrees of freedom (torsion angles of the protein backbone), and can give a description of the thermal fluctuations of a protein structure, as well as generate structural ensembles. However, the TNM is a coarse-grained model that cannot ensure that the newly created conformations are exempt from any structural defects. Therefore, the main hypothesis of this project is that TNM assembly process can be improved. The ability to generate high-quality structural ensembles describing the dynamical properties of a protein would indeed be highly valuable in various applications.

In this thesis, we create, evaluate and refine TNM ensembles from a set of reference protein structures defined experimentally (Levin et al., 2007). An approximation used in Bastolla and Dehouck, 2019, is developed: the evaluation is performed by Molprobity analysis, and the refinement is done by SIDEpro. Furthermore, a new approach is taken when refining the ensembles by Energy Minimization (EM).

The results show a potential improvement of the TNM ensembles when adjusting the target RMSD to the protein studied; point to a enhancement when using side-chain reconstructions , and to its combination with Energy Minimization as a way to optimize the structure quality. On the other hand, the pros and cons of the followed methodology are discussed, because the use of the available static-protein oriented measures and methods makes specially important to beware of their limitations when applied to the protein-dynamic oriented TNM.

Exploring further target RMSD values, adjusting them to specific protein dynamic simulations or replicating the same pipe-line in different data-sets are some of the proposals for future work. Furthermore, taking into account variables like the temperature, the flexibility of the protein, and the estimated optimal RMSD would be interesting for the next studies.

2. Introduction

2.1 Protein Dynamics

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function (Jumper et al., 2021). The development of experimental and computational methods to predict three-dimensional (3D) protein structures is crucial. Several experimental and computational techniques have been developed to get protein structures, such as nuclear magnetic resonance (NMR), X-ray crystallography, mass spectrometry (MS) or molecular dynamics (MD) simulations (Masrati et al., 2021).

Protein dynamics studies are essential to learn the mechanism underlying a protein biological role, such cellular signaling by protein-protein interactions and catalysis by enzymes. (Matsumoto, S. et al., 2021). However, one of the main limitations of the previous approaches is its inability to account for the dynamical aspects of proteins (Vakser IA., 2020). The thermal fluctuations and conformational changes are not considered. Methods such X-Ray crystallography provide little more that a purely static picture of their structures (Antunes et al., 2015). Over the years, several approaches have been carried out to use their physic-chemical properties to solve it. A potentially powerful strategy is describing the thermal fluctuations via an ensemble of static snapshots, which could easily be fed into any structure-based prediction tool (Miller et al., 2021).

One of the most promising methodology is the combination of experimental data from Nuclear Magnetic Resonance (NMR) with computational simulations of Molecular Dynamics (Lindorff-Larsen et al., 2005). But even beyond the requirement for specific experimental data, such approaches tend to be very computationally expensive and cannot be applied systematically to large protein data-sets.

2.2 Creating a dynamic protein model

The computational modeling of protein structure is essential for studying biological systems as an alternative to experimental structure determination methods (Zhang et al., 2009). One of the most used methods for the structure determination is the prediction based on homologous structures, and it grows as the available proteins do (Kryshtafovych et al., 2019). Another one is the use of inter-residue distances in order to do predictions (Marks et al., 2011).

These techniques generate a static model of the protein. Their inability to show the dynamic characteristics leaves to structural bioinformatics a challenge that must be overcome. All-atom molecular dynamics (AMD), the state-of-the-art simulation technique, enables cheap studies of small protein fluctuations (e.g., side-chain or very local backbone moves) in small protein systems (Pan et al., 2016). Nevertheless, many of protein systems are either too large to be effectively simulated using AMD or require very large supercomputer resources.

A little acceleration of AMD can be achieved when using structure-based models that provide quite valuable but limited insight into large-scale protein dynamics (Kmiecik et al., 2016). An important speed-up is possible by simplifying the protein description to a less complex level. It can be done by coarse-grained protein models (Orozco, 2014) and elastic network models (ENM) with normal-mode analyses (NMA) (López-Blanco JR and Chacón P, 2016). Both can be used as parts of multi-scale modeling methods merging computational tools of various resolutions from the low-resolution to atomic level (Stumpff-Kane et al., 2008).

2.2.1 The Normal-mode analysis and the Elastic Network Models

NMA is a fast and simple method to calculate vibrational modes and protein flexibility. The atoms are modeled as point masses connected by springs, which represent the inter-atomic force fields (Alexandrov en at., 2005). The ENM (Tirion, 1996) is a particular kind of NMA where a potential is set and only the atom pairs within a cutoff distance are considered. It uses uniform harmonic potentials for interacting atom (or residue) pairs instead of more complicated ones (Kmiecik et al., 2018).

The NMA method was used for studies of thermal fluctuations around the native structure in the early 80s (Go et al., 1983) (Karplus and Brooks, 1983). Low-frequency normal modes belong to the co-operative motions of bigger parts of the protein which are essential for its function, while high-frequency modes corresponding to small, isolated, non-co-operative vibrations are functionally meaningless. This allows to use the Principal Component Analysis (PCA) as an statistical approximation by taking into account just the low-frequency meaning normal modes. PCA can also be applied to structure snapshots from molecular dynamics or Monte Carlo simulations, to structures from NMR (Howe, 2001) or to the analysis of large sets (clusters) of proteins determined by X-ray crystallography (Yang et al., 2007).

The ENM was adapted to the Gaussian Network Model (GNM). It takes each residue as a single node with its co-ordinates usually given by the Cα atom. Nodes separated by less than a cut-off distance are connected by the same harmonic springs (Ma, 2005). The ones further that the cut-off distance are not connected at all. This model takes the assumption of Tirion (Tirion, 1996) that bonded and non bonded interactions can be summarized to a single uniform spring parameter. Even if it seems unrealistic, the NMA results are almost indistinguishable from those using complex energy-function potentials. This is due to the fact that any complicated potential function is very well approximated around its minimum with an harmonic function.

While the GNM assumes that fluctuations of residues around their mean positions are spherically symmetric, the Anisotropic Network Model (ANM) (Atilgan et al. 2001) takes the fluctuations around their starting points as an-isotropic and are represented by ellipsoids (Eyal et al., 2006). Due to the changes in the B-factor at different resolutions, the ANM is useful when dealing with very high resolution reference structures, while GNM represents a good approximation for most of the proteins in the Protein Data Bank (PDB) (Eyal et al., 2015).

ENMs have been combined with different input information and methods, such as, for example, data from atomic MD simulations (Mishra and Jernigan, 2018), Brownian simulations (Orellana et al., 2016) or structure-based models (Poma et al., 2018).

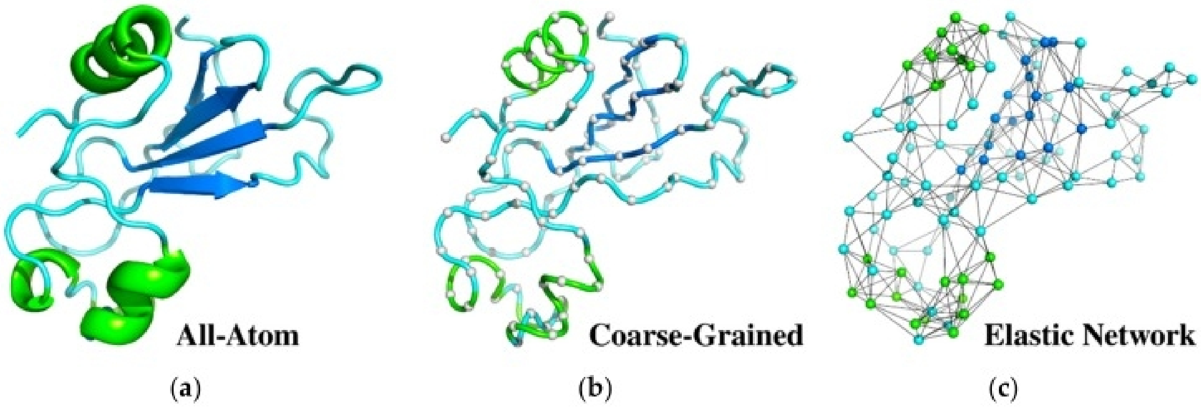


Figure 1: Different representations of the protein structure with PDB code 1a2p, used as reference in dynamics studies. (a) All-atom structure for MD simulations as a ribbon diagram; (b) coarse-grained model where gray spheres represent modes connected by a tube; (c) coarse-grained elastic network model with spring-type constraints and nodes in the positions of Cα atoms. There is a simplification of the model from left to right: in (a), all atoms are considered, which is proper of MD simulations; in (b) the model has been deeply simplified setting the modes which represent residues and links are conserved; in (c), only the Cα positions are stored, and the nodes are connected automaticaly according to the distance between them and the cutoff distance stablished. Taken from Kmiecik et al., 2018.

2.2.2 Using ensembles as an approach to understand the dynamics of a protein

Most of the previous systems use Debye-Waller factors to account for fluctuations in the structure, describing the motion of an individual atom as an isotropic Gaussian distribution of displacements about an average position. Despite its widespread utilization, there are limitations that rely on the B-factor (or Debye-Waller factor) as the sole parameter to describe the conformational variation of the structure. Plus, it has been suggested a misleading degree of accuracy due to the underestimation of the Root Mean Square Deviations (RMSD) from the average coordinates (DePristo et al., 2004). B-factors are commonly used to evaluate models because they are available from X-ray structures, but they are inaccurate (Dehouck and Mikhailov, 2013) .

Even if temperature factors can model the direction of protein motion, they are limited to these Gaussian distributions for describing the probability density function of each atom, and it is not able to get an-harmonic or multi-modal motion. (Levin et al., 2007). Structures can be described as ensembles of hierarchical conformational sub-states, representing them as a set of overlapping, non-interacting conformers accounting for a fraction of the total electron density.

The concept of the use of ensembles is three decades old (Kuriyan et al., 1991) but the concept was not deeply explored due to the computational expense of performing simulated annealing on systems with a lot of atoms, and the lack of high resolution data-sets. Such obstacles have been over-passed thanks to the increment in computer processing speeds and high resolution structures. Nowadays we can find many systems that create ensembles and refine them as a way to model a protein (Porter et al., 2019) (Karplus and McCammon, 2002) (Köfinger et al., 2019).

A large-scale assessment of the usefulness of ensemble refinement for extracting quantitative descriptions was carried out by Levin in 2007. It used the X-ray crystallographic data from a previous study (Furnham et al., 2006). Their results suggested that refinement with an ensemble of conformers can reduce the R-free factors and improve the estimation of the of motions of protein X-Ray structures (Levin et al., 2007). In this thesis, we use this data-set to perform our own analysis.

2.2.3 The Torsional Network Model

As an alternative NMA, the ANM uses the Cartesian coordinates of the Cα degrees of freedom (Atilgan et al., 2001). One of its unwanted features is that the modification of covalent bond lengths and bond angles has the same energetic cost as modifying soft degrees of freedom such as torsion angles, so ANM excitation do not keep the correct covalent geometry and secondary structure (Mendez and Bastolla, 2010). However, torsion angles offer the advantage that they allow us to represent all backbone atoms including C*β*.

In fact, at the beginning of the application of NMA to proteins, it was proposed that they should use only torsional degrees of freedom, that do not modify the covalent geometry (Go et al., 1983). The most similar approach had been a ENM based on pseudodihedral angles (Ma et al., 2006). The use of torsional degrees of freedom was reconsidered for modeling elastic deformations (Omori et al., 2009), but it had not been applied in the context of ENM.

The Torsional Network Model (TNM) is an ENM whose degrees of freedom are the torsion angles of the protein backbone. The proteins are represented by nodes connected to each other with elastic springs (Méndez and Bastolla, 2010) (*See **Figure 1, c)***. The torsion angles of the protein backbone are used as degrees of freedom, so the normal modes of motion can be computed analytically, and give a description of the thermal fluctuations of the structure. It presents several advantages: a) it does not need high requirements in terms of computer time, b) unlike most coarse-grained dynamical models, the parameters of the TNM have been optimized on a NMR data-set of conformational ensembles, to ensure that the predictions respect the cooperative aspects of residue motions (Dehouck and Mikhailov, 2013) and c) the amplitude of the thermal fluctuations from the B-factors in X-Ray structures can be estimated with a compatible approach with the TNM (Dehouck and Bastolla, 2017).

The TNM use the same effective potential of previous ENMs. It is minimally frustrated (Bryngelson and Wolynes, 1987), so all native interactions are at a local minimum and all the force constants are equal. Its normal modes allow us to accurately reconstruct the positions of more atoms than other Cartesian ENMs, and to define inter-residues interactions. Thus, it has a smaller computational cost without distorting the covalent geometry. A lower number of modes significantly contribute to the conformation changes of the protein. The thermal fluctuations predicted through the TNM correlate more strongly with the experimental data than in other models (Menez and Bastolla, 2010).

These results make the TNM a competitive method for modeling, with possible applications to flexible docking,for drug design and refinement of homology modeled protein structures. Nevertheless, being a coarse-grained model, the TNM cannot ensure that the new
conformations created from the computed normal modes, to represent structural fluctuations in the thermal ensemble, are exempt from any structural defects. In this thesis, we use the TNM as our

approach to the creation of model ensembles, and we try to gather information about its use and optimization, and its compatibility with some model-improving methods.

2.3 Evaluating a protein structure

When dealing with proteins, models may contain significant errors. Typical ones range from misplaced side chains, incorrect loop conformations or backbone distortions (Bordoli et al., 2009). The precision of a protein model determines its capability for applications. However, at the time of modeling, the quality of a model is unknown and has to be predicted as well. Bioinformatics offer a huge open-source availability of tools able to evaluate structures.

Scoring functions evaluate different structural features of protein models in order to generate a quality estimate. Usually, scoring functions are primarily designed to rank alternative models of the same protein sequence (Benkert et al., 2008; Eramian et al., 2008; McGuffin, 2008). Nevertheless, changes in model quality between different target proteins is normally larger than the variability within the ensemble of models generated by different prediction methods for a given protein (Battey et al., 2007).

2.3.1 Global validation measure

Relative ranking of alternative models for the same protein is sometimes insufficient for getting its usefulness for applications. An absolute reliable quality estimation is crucial to use computational models (Schwede et al., 2009). The original tool ProSA (Sippl, 1993) was developed to evaluate experimental structures and to get the statistical significance by comparing its knowledge-based score to random structures made from the same sequence. But it cannot be used as a measure of an absolute model because it is highly dependent on the protein size. Eramian et al. (2008) used support vector regression to get the quality based on the comparison with similar models selected from a database of structure-pairs generated by the same method. Wang et al. (2009) calculated the agreement of a model with features predicted from the primary sequence as a measure using the SCRATCH suite (Cheng et al.,
2005). These scoring functions act on protein chains and cannot evaluate biological assemblies.

Benkert et al. (2008, 2009, 2010) created QMEAN, which analyses different geometrical features of proteins. The QMEAN score is compared to distributions obtained from high-resolution structures solved by X-ray crystallography. It can be used for estimate the absolute quality of a protein structure independently of protein size. Its QMEAN Z-score, estimates the similarity to the native protein of the model analyzed compared to experimental structures. That is, if the model evaluated is realistic or not according to their database. QMEAN has become a famous software for the protein global evaluation.

2.3.2 Local validation measure

The previous approached are focus on wide validation measures. None the less, no global quality can guarantee the absence of large local errors in a region of specific interest. Programs such PROCHECK (Laskowski et al., 1993), WHATIF (Vriend et al., 1990) and OOPS (Jones et al., 1991) tried to provide both statistical evaluations and flags local problem areas. Other validation utilities took aspects of the model-to-data agreement, like the R free value for X-ray (Clore and Garret, 1999), or RPF scores for NMR (Huang et al., 1999).

Molprobity is an actual widely used free-system of structure-validation for different structures. It gives a broad-spectrum solidly by the evaluation of model quality at global and local levels. It is build upon the work of the earlier systems cited above: PROCHECK, WHATIF and OOPS (Christopher et al., 2018). It can also perform some local corrections using a combination of previous free-code software and it offers a quality overview of the structure (Davis et al., 2007). It was pioneer in offering all-atom contact analysis and up-to-date, high-accuracy Ramachandran (Ramachandran et al. 1963) and Rotamer distributions at 1999. Besides, it can be applied to X-ray and NMR ensembles, which makes it really useful when dealing with several models of the same protein.

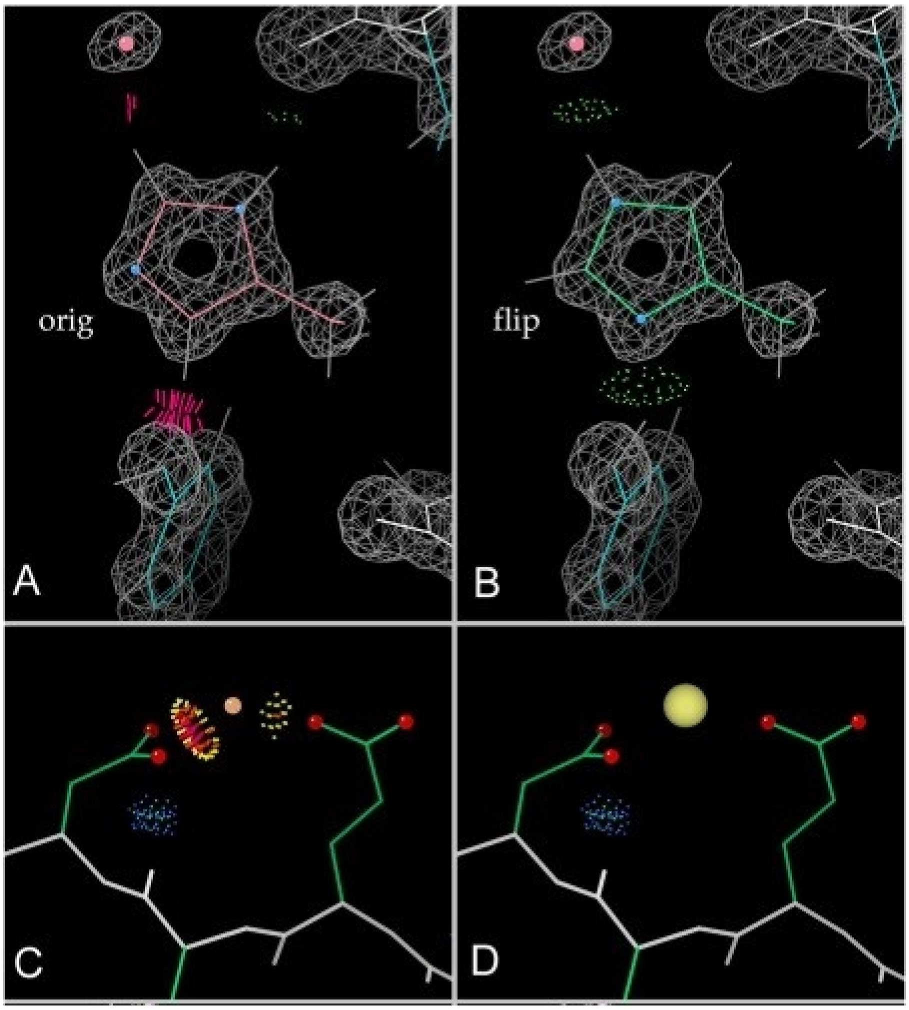
The program has suffered a lot of updates and improvements since its creation. The complementary Rotamer Rachamandran and the Cβ deviation criteria by 2003 (Lovell et al., 2003). The implementation of many Molprobity validations inside the Phenix crystallography package, the resulting improvements in clashscore, and Asn/Gln/His flips by 2010 (Chen et al., 2010). Some features like the simplification and improvement of the Methyl orientations adjustment, and improved all-atom contacts system called *OneDotEach*, the better-idealized output coordinates from asparagine, glutamine and histidine flips; and the redefinition of H-atom parameter sets were introduced by 2018 (Christopher et al., 2018).

The Molprobity usefulness became a standard, and its criteria has been adopted by the world-wide Protein Data Bank (wwPDB). Specifically, their parameters of clashscore, Ramachandran and Rotamers were included in the wwPDB. In this thesis, we use these ones to analyze the models. The Molprobity score is a single indicator of the model quality based in a weighted function of clashes, Ramachandran favored and Rotamer outliers, scaled and normalized so it approximates the resolution at which tat score would be average. If the Molprobity score is below the actual resolution, that means that the model has a good quality, because its combination of Clashscore, Ramachandran and Rotamers would be expected in a model of higher resolution. Far details about these parameters are explained in **4. Methods.**

2.4 Improving a protein model

The models generated by template-based modeling are similar to the homologous target sequence but cannot usually capture how sequence variations module the structure in detail. Due to differences in side-chains, these homology models have poor structure packing. (Heo et al., 2021) .On the other hand, models based on predictions may also deviate from their native structures, since current machine learning-based methods rely on co-evolutionary analysis (Senior et al., 2020).

Protein model refinement helps this issues. It tries to improve initial model qualities in terms of global structure arrangement, local structure packing and stereo-chemical properties (Read et al., 2019). Earlier methods looked for the refinement of local structures improving side-chain packing or hydrogen networks, while the global structure improved slightly with these local improvements (Bhattacharya and Cheng, 2013) *See **Figure 2***. Usually, the kind of refinement needed would depend on the method used to create the model in the first place. Many tools that create models have to be simplified or coarse-grained to deal with large search space, big data-sets, etc. This simplification would need to be compensated in order to improve the quality of the model.



clash
(H-bond, vdW)

Figure 2: example of one of the main measures of Molprobity, the clashes. (A): a histidine residue is clashing with two good H-bonds. The ideal situation would be after a flip of the Histidine (B). (C): a water clashes nearby carboxyl oxigens. A sodium ion (D) fits better this position. Molprobity allows to make the corrections in (B) and (D), but our interest is just the assesment. Conformations such (A) and (B) lower the quality of the model and are contemplated in the Molprobity Score. Adapted from Williams et al, 2018.

2.4.1 Prediction of protein side-chain conformations

In closely related proteins, the backbone often changes little and an accurate side-chain prediction can improve notably the quality of a model (Veenstra and Kollman, 1997). Also, in docking of ligands, changes in side-chain conformation are critical to the prediction of complexes (Meiler and Backer, 2006). Repacking is the recalculation of side-chain conformations. It needs to be very fast because many can be find in a model (Rohl et al., 2004).

Most side-chain prediction approaches are based on a Rotamer library, which is a statistical clustering of side-chain conformations in known structures that rotate around a single bond and are positioned in a local energy minimum (Dumbrack et al., 2002). They can be backbone-independent or backbone-dependent, if they are affected by backbone diedral angles and frequencies. Rotamer libraries classically catalog favored side-chain conformations by the mean χ values and standard deviations for each Rotamer. They are created by performing statistical analysis on a selected data-set of experimentally-determined models (Hintze et al., 2016). Another method is to use conformer libraries, which take plausible three-dimensional molecular structures (bigger than Rotamers, conformations) from known structures in form of Cartesian coordinates (Peterson et al., 2004)(Cole et al., 2018). After the repacking, in both methods, a scoring function evaluates the suitability of the sampled conformations.

One of the most extended software to perform prediction of side-chain conformation is SCWRL4 (Georgii en al., 2009). It uses a backbone-dependent Rotamer library which has been periodically updated, and short-range, soft van de Waals potentials as well as an an-isotropic hydrogen bond function. The scoring function samples angles around the mean values of the Rotamer library and averages the interaction energy of Rotamers of different side-chains. The result is a fast and accurate method to side-chain prediction.

The physics-based approach like the one from SCWRL4 is not suited for the coarseness of such models (Nagata et al., 2012). Specifically, they are quite sensitive to slight changes in atom-atom distances. So, some side-chains interactions end having much higher repulsive energies relative to the native structures. Knowledge-based energy functions from large training sets (Xiang et al., 2007) are more tolerant to discretization and sometimes capture subtle effects unseen by the physics-based methods.

SIDEpro (Nagata et al., 2012) is a software which uses this system. Furthermore, it uses artificial neural networks (ANNs) trained to compute an energy function by atom-atom distances. ANNs are trained with modified versions of PDB structures whose side-chains are set to the most accurate rigid Rotamer. SIDEpro begins by setting an initial probability of every Rotamer. using a backbone dependent library. Then, it iteratively updates these probabilities using the ANNs until all the Rotamer. probabilities converge. Next, the side-chains are set to the Rotamers with the highest probability. In the end, a post-processing clash reduction is applied. This SIDEpro approach surpasses SCWRL4 in speed and accuracy.

2.4.2 Molecular dynamics and energy minimization

Molecular dynamics (MD) simulations are usually applied to improve global structure (Feig and Mirjalili, 2016). MD ensemble averaging is the average taken over a large number of replicas of a system considered simultaneously. This has been found to be key to get better models which resemble experimental structures and to reduce inaccuracies in the force field and scoring functions (Mirjalili et al., 2013). MD ensemble averaging has become one of the most successful strategies for refining both global and local features. Before, refinement was used only on models build by homology because only them had the necessary initial qualities. Now, thanks to these MD improvements, it is possible to apply model refinement on models from other sources and improve their quality (Heo and Feig, 2020).

MD simulations predict the fluctuations of every atom in a protein over time using a general model of physics governing inter-atomic interactions (Karplus and McCammon, 2002). The force experienced by every atom is calculated based on the positions of other atoms and the energy function. The first MD simulation was performed by Rahman with a realistic potential for liquid argon (Rahman, 1964). Another simulation in liquid water was realized in 1974 (Stillinger and Rahman, 1974), and it was considered the first realistic MD simulation. The first protein simulations were carried out on bovine pancreatic trypsin inhibitor in 1977 (MacCammon et al., 1977).

An important step before beginning the MD is the Energy Minimization (EM). It ensures that the systems are at an energy minimum state, preventing for steric clashes or inappropriate geometry. (Panwar and Ashok, 2021). It tries to get to put the structure in a conformation with the lowest energy. It stops when the local energy minimum is reached, usually without getting to the global energy minimum (Fadlan and Nusantoro, 2021). However, this most stable conformer could be reached using suitable algorithms (Roy et al., 2015). Force fields used in the EM evaluate atomic interactions, including Wan Der Waals and electrostatic interactions, bond-stretching, bending, and torsion forces. The forced field is commonly determinate based on experimental data and by mechanical calculations based on laws of physics (Allinger, 1977).

The force field used in this thesis is the Optimized Potentials for Liquid Simulations (OPLS) (Jorgensen et al. 1996). The OPLS has become widely-used for simulations on Proteins. It uses potential functions with a partially united-atom model; sites for non-bonded interactions are placed on all non-hydrogen atoms and on hydrogens attached to hetero-atoms or carbons in aromatic rings. So, the hydrogens attached to aliphatic carbons are the only taken into account. Computation time is proportional to the total number of interaction sites squared. That makes the OPLS computationally attractive because the interaction sites are reduced being a all-atom representation.

3. Objectives

1. Definition of a computational pipeline for the creation of conformational ensembles of proteins.

2. Creation of raw conformational ensembles using the TNM.

3. Evaluation of the structural quality of the structures produced by the TNM.

4. Refinement of the structures produced by the TNM.

5. Comparison of the quality of the predictions when using a single (experimental) structure or the conformational ensembles.

4. Methods

4.1 Data-set for essays

In this thesis, the proteins from Levin et al., 2007 were chosen as the data-set. Levin et al. applied an automated ensemble refinement protocol to a sample of fifty crystal structures with a variety of sizes, resolutions and degrees of conformational flexibility, as well as two sets of simulated crystallographic data from molecular dynamic simulations. The reason we picked this data-set was because for each original structure, an ensemble of models had been created from the same experimental data, which fit our needs. A closer look to the original data can taken at ***Table 1***.

pdb1	*1q44*	*1q45*	*1q4r*	*1vjh*	*1vji*	*1vk0*	*1vk5*	*1vkp*	*2i3c*
pdb2	*2q3m*	*2q3o*	*2q3p*	*2q3q*	*2q3r*	*2q3s*	*2q3t*	*2q3u*	*2q51*
nmodel	16	16	4	16	16	8	16	8	16
Rmsd_avr	0.584526	0.428004	0.321884	0.45343	0.513187	0.540773	0.367215	0.40009	0.735026

pdb1	*1xmb*	*1xmt*	*1xq1*	*1xq6*	*1xri*	*1xy7*	*1xyg*	*1y0z*	*2a3q*
pdb2	*2q43*	*2q44*	*2q45*	*2q46*	*2q47*	*2q48*	*2q49*	*2q4a*	*2q4p*
nmodel	16	2	8	8	16	8	8	4	16
Rmsd_avr	0.363466	0.254826	0.330618	0.358668	0.523873	0.312711	0.350264	0.334263	0.943156

pdb1	*1z7x*	*1z84*	*1z8k*	*1z90*	*1ztp*	*1zwj*	*1zxu*	*2a13*
pdb2	*2q4g*	*2q4h*	*2q4i*	*2q4j*	*2q4k*	*2q4l*	*2q4m*	*2q4n*
nmodel	8	8	8	8	16	4	16	8
Rmsd_avr	0.334592	0.452785	0.283425	0.583902	0.608108	0.459244	0.453668	0.308212

pdb1	*2bdu*	*2be4*	*2bei*	*2exr*	*2f2g*	*2il4*	*2gu2*	*2h1s*
pdb2	*2q4t*	*2q4u*	*2q4v*	*2q4w*	*2q4x*	*2q4y*	*2q4z*	*2q50*
nmodel	4	16	8	16	8	4	16	8
Rmsd_avr	0.533203	0.660314	0.45253	0.414839	0.612617	0.373385	0.472364	0.642902

pdb1	*1vm9*	*1xfi*	*1xj5*	*1xm8*	*1ycn*	*1ydh*	*1ydw*	*1yvi*
pdb2	*2q3w*	*2q40*	*2q41*	*2q42*	*2q4c*	*2q4d*	*2q4e*	*2q4f*
nmodel	8	16	8	16	8	16	16	8
Rmsd_avr	0.338329	0.416611	0.488419	0.285021	0.585169	0.404876	0.547374	0.448589

pdb1	*2i3f*	*2i5t*	*2ab1*	*2amy*	*2atf*	*1vm0*	*1ybm*	*2a33*
pdb2	*2q52*	*2q53*	*2q4q*	*2q4r*	*2q4s*	*2q3v*	*2q4b*	*2q4o*
nmodel	8	4	16	16	16	4	16	8
Rmsd_avr	0.373457	0.304208	0.54933	0.561554	0.318254	0.266126	0.388566	0.392783

Table 1: group of tables with the Reference pdb codes (pdb1), the Ensemble pdb codes (pdb2), the number of models of each ensemble (n_models), and the RMSD average of these model ensembles with the reference structure; from Levin en al., 2007.

The original X-ray structures are named as “Ref” (Reference), and we work mainly with them. One set is made by an automated ensemble refinement protocol from the original Ref files, and they are classified as “Ens” (Ensembles). Every conformation is compound by a variable number of models: 2, 4, 8 or 16 and receive a different specific name. The last set consist on the biological assemblies of the protein, that is, their functional unit, but it is not used in this thesis.

In Levin et al., 2007, the coordinate sets from each simulation were aligned to the original structure and used to calculate structure factors, which were then averaged to produce a single set of reflections. Conventional single conformer models with isotropic temperature factors were fitted to the simulated data. These models were then used as starting structures for the automated refinement of one-, two-, four-, eight-, and sixteen conformer models against the structure factors calculated from the simulations, using a combination of torsion angle simulated annealing (Rice and Brunger, 1994) and standard maximum likelihood refinement.

The refinement procedure used by Levin et al., 2007 is similar to that described in Wilson and Brunger (Wilson and Brunger, 2000), where each atom is given an individual temperature factor, all conformers are given equal fixed occupancies, and the initial separation of the conformers is achieved by torsion dynamics simulated annealing. This data-set was taken from the wwPDB, where all the structures are stored, and then adapted and expanded according to the needs of the thesis. To do so, a list of all the pdb codes used in Levin et al., 2007 was taken from the paper. It was parsed and then used as key-names to retrieve the correspondent PDB files from the wwPDB.

4.2 Creation of raw conformational ensembles using the Torsional Network Model

The TNM (Méndez and Bastolla, 2010) has been previously developed and is readily available. It gives as output the normal modes of motion of a protein, in terms of variations of backbone torsion angles. Modified structures of the target protein will be constructed by randomly selecting combinations of normal modes (in such away that the energies of the ensemble of conformations respect the Boltzmann distribution), and rebuilding each structure from the modified values of the backbone torsion angles. An important input parameter is the amplitude of the modeled fluctuations, i.e. the root mean square deviations (RMSD) of the coordinates with respect to the initial structure. The TNM software was given by the original authors. It also was adapted to run it from a Jupyter Notebook using Python language. The TNM receives as input the reference structure, and a standard file with several parameters which determine the way the ensembles are generated. We set most of the parameters from Bastolla and Dehouck, 2019 when choosing energy and torsion values options.

None the less, we changed a parameter that had not been optimized yet, the “THERM_ENS rmsd”. It is an important input parameter which determines the amplitude of the modeled fluctuations, i.e. the RMSD of the coordinates with respect to the initial structure. The number set is the target RMSD in Angstroms (Å) used for the TNM to create de models . So, the TNM software will construct the ensembles in a random iterative way trying to keep the general RMSD close to the number chosen. *See **Figure 3.*** The RMSD values to perform the models were: 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2. Other parameter that was modify is “THERM_ENS nconfo“, which determines the number of conformations generated by protein, slightly different from each other. The number chosen for all the essays was 30, as it allows for statistical significance. The last parameter modify was THERM_ENS seed, which was set to 0. This allows the experiment to be reproducible. The TNM ensembles and results were build in a branched folder structure adapted to the Levin et al, 2007 data-set. For the 8 RMSD values, 30 models were made for each one of the 50 proteins. The total number of TNM models was 12000.

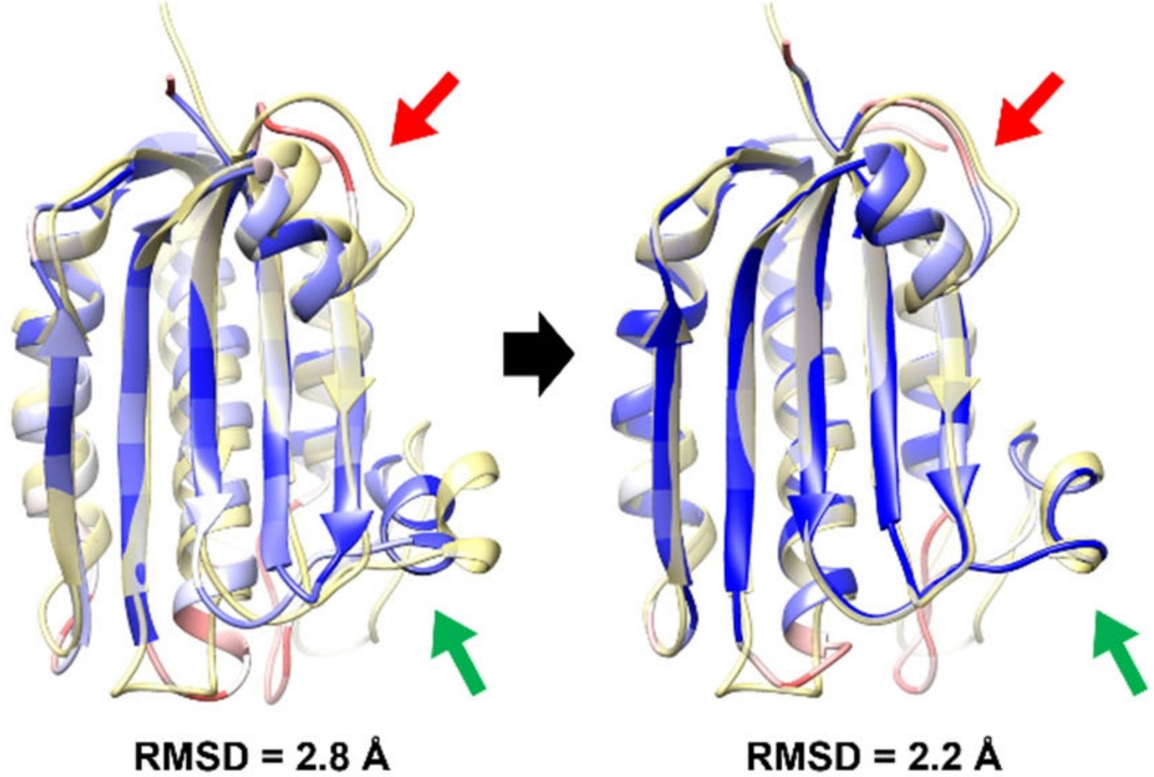


Figure 3: representation of the same protein structure when it is simulated with a higher RMSD and a lower RMSD. The semi-transparent structure is the native or reference structure, while the coloured is the model generated. When a lower RMSD is selected, the model simulated is better adapted to the reference structure whereas if the RMSD Is higher, the model is more deviated from the original. It can be easily observed on the upper right loop.

4.3 Molprobity: quality evaluation of the TNM Ensembles.

The criteria to assess the quality of the ensembles generated, as well as all the later modifications in order to improve the models, is Molprobity (Chen et al., 2010). The wwPDB approves several parameters as suitable measures for protein quality: the clashscore (CS), Ramachandran conformations outside favored region (Rama_iffy) and Rotamer outliers (Rota_out). These are summarized by the Molprobity score (Mpscore), which is the main reference we use to judge the overall quality of a structure in this thesis. The Mpscore is calculated as follows:

MPscore = 0.426 *ln(1+clashscore) + 0.33 *ln(1+max(0, rota_out|-1)) + 0.25 *ln(1+max(0, rama_iffy|-2)) + 0.5

- ***Clashscore:*** is defined as the number of unfavorable all-atom steric overlaps ≥ 0.4Å per 1000 atoms
- **Rota_out:** is the percentage of side-chain conformations classed as Rotamer. outliers. It uses an updated Rotamer library.
- **Rama_iffy:** The φ and ψ torsion angles, which are defined for the backbone N-Cα and Cα-C bonds, are known to take up only certain values due to steric considerations (Ravikumar et al., 2021). The Ramachandran plot classifies the angles in fully allowed, partially allowed and

disallowed regions. Rama_iffy is the percentage of backbone Ramachandran conformations outside the favored region.

The coefficients were derived from a log-linear fit to crystallographic resolution on a filtered set of PDB structures, so that a model's MPscore is the resolution at which its individual scores would be the expected values. Thus, lower MPscores are better. Each component of the formula is a local indicator which is normalized to the same scale to generate the Molprobity Score, which is comparable to the resolution of a structure. Besides de Mpscore, we study these components separately to get a better look to the changes in the structure by using directly the percentages of outlier Rotamers and Ramachandran and the clashscore. These are calculated taking into account the total number of Rotamers/Ramachandran recognized and calculating a percentage of the disallowed conformations.

Molprobity is usually run from the web server. None the less, it is not advisable nor allowed to use it this way when dealing with a significant number of Proteins (bulk study). There is a local instance of Molprobity for this purpose available at HTTP://molprobity.biochem.duke.edu. It was installed and adapted to be run on a Jupyter Notebook written in Python language. A lot of features from Molprobity are eliminated when it is used this way. Graphics, detailed analysis, and other features are just reachable from the web-server. However, the Molprobity score and other measurements are still available in order to do a general analysis.

The main command used to get the data has been “oneline-analysis”, which provides a .txt file as output with a line for every model presented in the input file. Each line contains the parameters that are needed: Clashscore, Rotamer outliers, Ramachandran outliers and the Molprobity score. This part has been undoubtedly the most time and computational expensive, as “overnight computations” were needed in order to analyze all the models generated.

4.4 Refinement of the models generated by the TNM

In order to asses the potential of the combination of TNM ensembles with other model-improving methods, we used SIDEpro (Nagata et al., 2012). An instance was downloaded and installed on the computer from HTTP://sidepro.proteomics.ics.uci.edu/ . It was adapted to be run from a Jupyter Notebook in Python language. SIDEpro takes as input a pdb file. After a quick computation, it returns another pdb file whose side-chains have been optimized changing their coordinates. The output pdb files have a standard format and could be use as input for other post-refinement processes.

4.4.1 Energy Minimization

The second method to refine the TNM output was the Energy Minimization process using the Optimized Potentials for Liquid Simulations (OPLS) force field. GROMACS was chosen for this function. It is a free software available from HTTP://www.gromacs.org/ , and can be implemented in Bash. It allows to do Molecular Dynamics (MD) simulations, but our interests were the EM. To do this process, the external tutorial HTTP://www.mdtutorials.com/gmx/lysozyme/01_pdb2gmx.html was followed. The process was automated. In order to do this, several steps had to be taken due to incompatibilities and errors. The right way to run it on all the TNM structures was combining a Bash script (with the modified steps from the tutorial) and parsing integrated with

Python. The script had to be modified every time it was run on a protein in order to work. A copy of the model script (used to create every script for the proteins) can be found on the **Annex.**

First, the water molecules from the original are removed. This is done with a simple grep command. Then, we use the pdb2gmx GROMACS module. It generates three files: the topology, a position restraint file, and a post-processed structure file containing important information for the next steps (atom types, charges, angles…). Next, an aqueous system is simulated: a cubic box is defined and then is filled with water using a solvable module.

A mdp file containing parameters is needed, the same from the tutorial is used, as it is suitable for a great range of proteins if the OPLS force field is employed. It is assembled and an atomic-level description of the system is obtained. The resulting file is passed to the GROMACS command genion, which process the topology to reflect the removal of water molecules and addition of ions. The solvable, electro-neutral system is now assembled. Finally, to reduce steric clashes or inappropriate geometry, the structure is relaxed through the EM using the GROMACS MD engine.

4.5 Parsing the data-set

An efficient way to read all the data from Levin et al., 2007 was needed. Biopython is a set of freely available tools for biological computation written in Python by an international team of developers. It is a distributed, collaborative effort to develop Python libraries and applications for Bioinformatics (HTTP://biopython.org/). Its package Bio.PDB allows to parse and manipulate pdb files easily, and also access to information like the number of atoms, residues, chains, and models present in a file.

For most utilities, it iterates through the coordinates and collect the data that it needs. It also allows to manually iterate through the atoms and residues, so it is easy to access any point. Bio.PDB was used to read the pdb files and to get all the data. Knowing the number of residues and atoms was crucial to know that the downloaded structures were the right ones, and also to check that the data-set was in most part coherent. This was specially important when checking for the number of residues of every model in the ensembles from Levin et al., 2007.

4.6 Integration and adaptation of all the previous software in a Python package

As it has been said several times, all the previous programs were adapted to be run from a Jupyter Notebook in Python. The external programs (Molprobity, SIDEpro, GROMACS) were not rewritten in any way, but their features were properly called from the Python script. The reason was the integration of all the methods in the same place, so the parsing of the files was combined with the different programs and automatized.

An initial script imports some common modules needed. It also establish some global paths that are used in all the classes. The second initial script “build” is done to build all the folders needed in the package, following a branched distribution. It must be run every time the program is run in a new computer so all the paths exist when the other utilities are run. Still, there are local computations inside the functions to check the existence of these paths and to create them in case they do not.

The whole program is written around the classification of the input files. They can be “Ref files”, if they come from an unique trusted structure (just one model inside the files); “Bio files”, if they correspond to biology assemblies (just one model); “Ens files”, if they are ensembles made from one of the Ref files (they can contain from two to sixteen models); and the “TNM files”, which contain thirty TNM ensembles. Its kind is usually specified as an argument in the input or inherited from another function, as the way to treat the files varies according to it. In this thesis, we have worked essentially with the Ref files and the TNM files.

The functions of the Python program created are distributed in classes. It can be seen in the **Annex**. Once the coding was finished, it was converted to a Python class and each main class was set as a module. The main classes created were:

- ***Strdata:*** to load the structure, get the number of atoms, residues, chains and models, select specific sets of atoms, calculate the RMSD of a structure and get the coordinates of specific sets. This class uses a lot of functions from Bio.PDB, and text file parsing. Besides, a function to calculate manually the RMSD from the coordinates was developed in order to check the results from Bio.PDB.

- ***Protein:*** to check the integrity of the structures and to throw a summary output about the Protein loaded. It calls Strdata and uses text file parsing.
- ***Molprobity:*** to call several functions from Molprobity. It gives several outputs to assess the quality of a structure, such as clashscore or the Mpscore. Besides, a function for calculate manually the Mpscore was developed using the formula *(See **4.3**)*, in order to check that it is right.
- ***SIDEpro****:* to call SIDEpro and optimize the PDB files according to its kind (Ref, Ens, Bio, or TNM).
- ***Torsional_network_model:*** to call the TNM program and create TNM ensembles from the Ref files. Also offers some output specific for the TNM.
- ***GROMACS:*** to call a bash script with specific instructions for GROMACS to perform the Energy Minimization process.

4.6.1 Graphics and statistic analysis in R language

R is a programming language and free environment for statistics and graphics. It is widely used among data miners for the develop of statistics and data analysis (Ihaka and Gentleman, 1996). All the data collected was adapted to be read by R language. The Rstudio interface was used to ease the process. Some basic features from R were employed, and besides, the ggplot2 package was chosen to do clear graphics.

The main statistic test used was the Welch’s t-test. It is an adaptation of Student’s t-test, but it is more reliable when the two samples have unequal variances, which happens in a lot of data-sets of this thesis. It allows to test the null hypothesis that two populations have equal means. Conventionally, it is considered that a p-value lower than 0.05 α (level of significance) brings to refuse the null hypothesis and accept the alternative hypothesis: there is a statistically significant difference in the means. This allows us to check if the data we observe is really different or if it can be considered just random deviations. Usually, the “factor” of the means is gathered, in order to make comparisons. This factor is simply calculated by dividing the second data-set mean by the first data-set mean.

4.6.2 Workflow

A consistent method has been followed during all the work. *See* ***Annex*** for more details. This can be summarized in:

1.- Reading and parsing the Reference structures (***Strdata*** and ***Protein*** classes)

2.- Generating TNM ensembles with the right parameters (***Torsional_network_model***)

3.- (Optional) Refining the TNM ensembles with side-chain reconstruction or EM (***SIDEpro***, ***GROMACS***)

4.- Evaluating the TNM ensembles (***Molprobity***)

5.- Analyzing with statistics and graphs

5. Results

5.1 Check for basic features of the data-set

The number of managed structures was high, so the first step was ensuring that the data was whole and reasonable, to prevent future errors and guarantee the reliance of the thesis. Several checks were set when parsing the structure of each pdb file. For most cases, everything seemed right, but we found out that in one protein the number of residues mismatch between the ensemble and the reference file. This would have been a problem when analyzing the data, even if the difference was one residue. This is due to the way Bio.PDB calculates the RMSD: it goes throw all the coordinates and calculates the deviation *(See **4.5**)*. If there is a mismatch between the number of coordinates, it will throw an error.

There was a mistake in just one couple, so it was chosen to solve it manually. Their pdb names were 2il4 for the Ref file, and 2q4y for the ensemble. The "ok_counts" check was 0, which meant that there was a mismatch in residues or chains between the ensemble and the model. We iterated through both pdb lines and found the error. While the reference file had 90 residues, the ensemble models had 91. In fact, there was a proline residue at the beginning of the models which seemed to be duplicated. To be sure, we checked the wwPDB data. The discordance was the same in the data bank. All pointed that it was a duplicate Proline, so it was taken away. The file was restored with a text editor, so the structure was integral again. As it was a duplicated residue at the beginning of the sequence, we simply deleted it and no other concerns were taken. During the post-analysis, we checked that the modified structures had still sense.

pdb1	pdb2	nmodel	nchain	nres	ok_counts	ok_ens	ok_bio	Rmsd_avr	Ref_MP	Ens_MP_avr
2il4	2q4y	4	1	90	0	1	1	-1.000000	1.392000	0.000000
2il4	2q4y	4	1	90	1	1	1	0.373385	1.392000	1.540750

Table 2: data before the correction (upper line) and after the correction (lower line). As the ok_counts is 0, there is some missmatch between the Ensemble and the Reference files. Then, the program is not able to calculate the RMSD average nor the MP score due to this error. However, the reference Molprobity score is still calculated. Once the mistake is corrected and the analysis is run again (lower line), the program is able to calculate both parameters.

Other checks were the RMSD and the Mpscore. We calculated manually the RMSD using its formula and the coordinates of the PDB files. Also, the specific Molprobity parameters and the Mpscore were manually calculated using its formula to double-check the results.

# Atom_id: CA	Avr: 0.2661					
Model	RMSD	Resolution	CS	Rota_out	Rama_iffy	Mpscore
ref	0	1.8	0.877	0.013	0.001	1.864
0	0.3103	1.8	0	0.397	0.237	1.279
1	0.2322	1.8	0	0.333	0.106	1.046
2	0.2384	1.8	0	0.333	0.18	1.075
3	0.2836	1.8	0	0.151	0.106	0.829

Table 3: output from the RMSD-MPscore analysis from 1vm0 (Reference) and 2q3v (Ensemble). On the header, the atoms that were used to calculate the RMSD (Cα) and the RMSD average.

Besides the checking, it allowed to fully understood the way the quality indicators were obtained and their significance. The code can be seen in the ***Annex.*** All the checks were right, so no more concerns were taken.

5.2 The Torsional Network Model

Knowing the data-set to be consistent allowed to start the analyses. The first step was to study the change of the structure quality when creating the TNM ensembles. An ensemble of 30 models was created from the Ref files of the data-set. For now, the parameters used in one of the original TNM papers (Bastolla and Dehouck, 2019) were taken, in order to study the general behavior of the TNM.

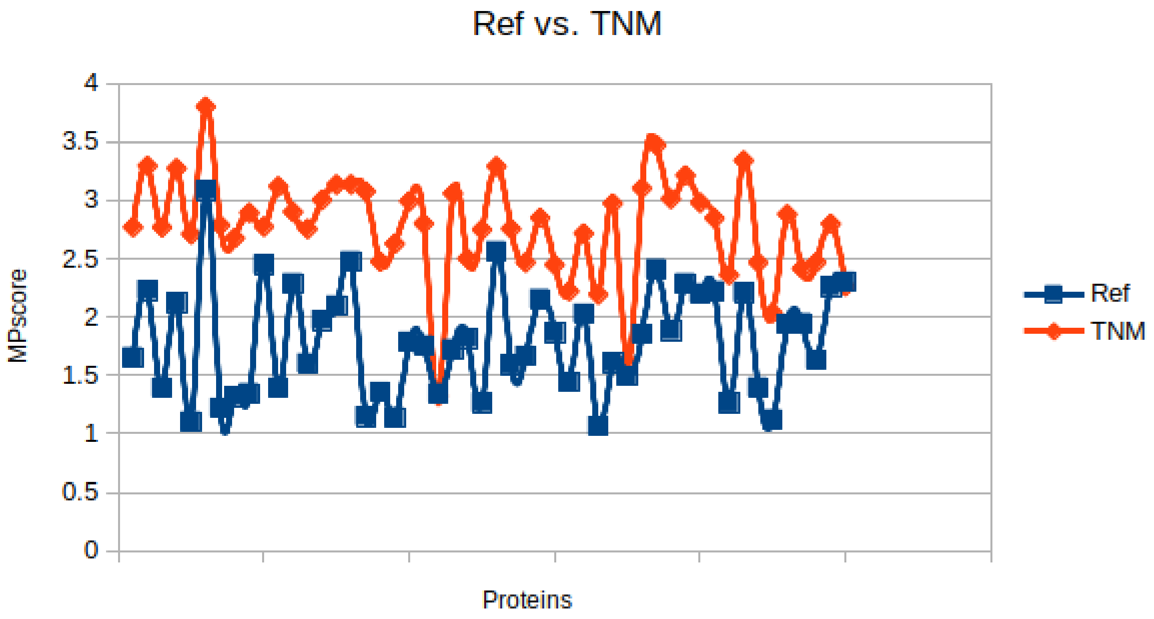


Figure 4: schematic graph representing the MPscore of the Ref files (Blue) and the TNM files (orange). The different proteins are represented on the same vertical line, so the blue-ref points have its TNM ensemble MPscore average represented exactly above.

A broad approach was chosen by doing the average of the MPscores of the TNM ensembles. Then, it was represented on the same graph grouped by Protein (***Figure 4***). This quick graph supports the results from the original TNM paper (Mendez and Bastolla, 2010). It suggests that the mean quality of the individual models gets worse when using the TNM. Plus, it seems that there is a "regular quality lost".

The distribution of the data can be seen in ***Figure 5.*** The median and the quartiles point to the TNM modeling as a definitive loss of quality. In order to know if it is statistically significant, a Welch t-test was performed. The null hypothesis is that there is no difference in the means.

Welch Two Sample t-test:

```
data:  Ref_tnm$Ref and Ref_tnm$TNM
t = -10.83, df = 97.761, p-value < 2.2e-16
mean of x mean of y
 1.788860  2.771862
```

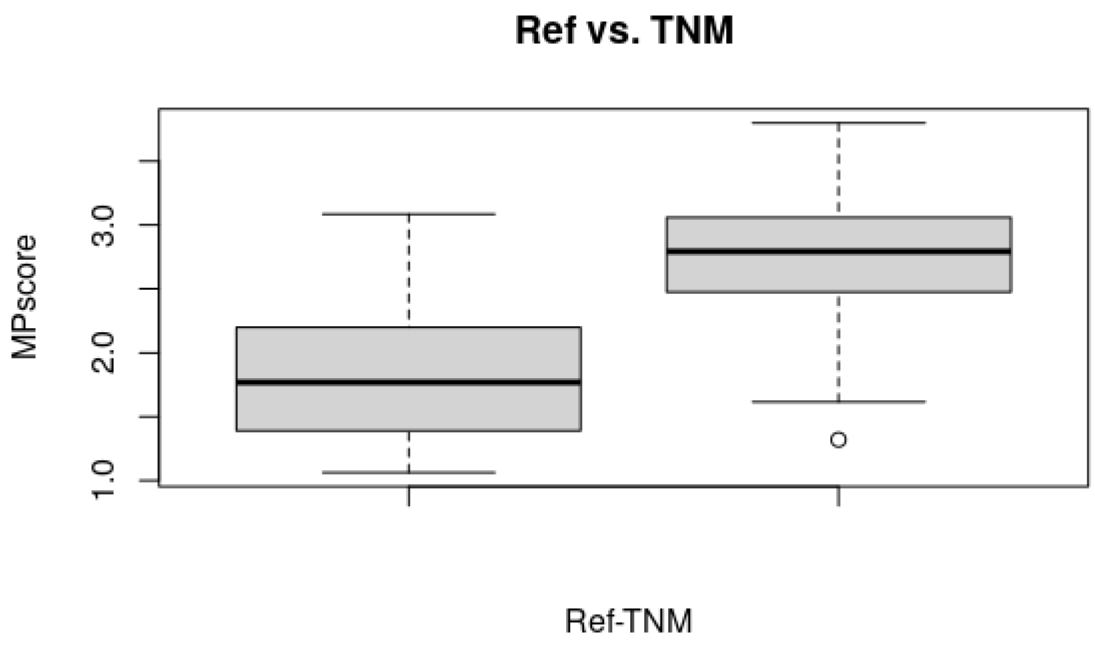


Figure 5: boxplot representing the MPscore of the Ref structures (left) and the TNM structures (right).

The difference of means is 0.982 MPscore; with a factor of 1.549 of MPscore increment for the TNM ensemble in relation to the Reference. The p-value is 2.2e-16 << 0.05 α (level of significance); so it can be affirmed that the Ref-TNM condition affects the Mpscore with statistically significance.

5.2.1 Local distortions

By the Mpscore, we can know that the average quality is reduced, but more concrete features can be studied. Next step was to check the changes in Clashscore (CS), Ramachandran no-favored percentage (Rama_iffy) and Rotamer outliers percentage (Rota_out) produced by the TNM, which are components of the MPscore. The 30 TNM ensembles with the default parameters (Bastolla and Dehouck, 2019) were taken. Then, the Ref clashscores, Ramachandran and Rotamers were extracted and normalized following the Molprobity method.

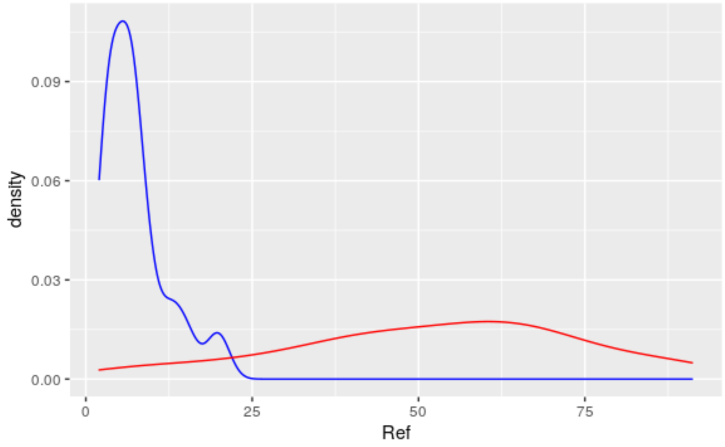


Figure 6: curve density with the CS of the reference structures (blue) and the TNM ensembles (red).

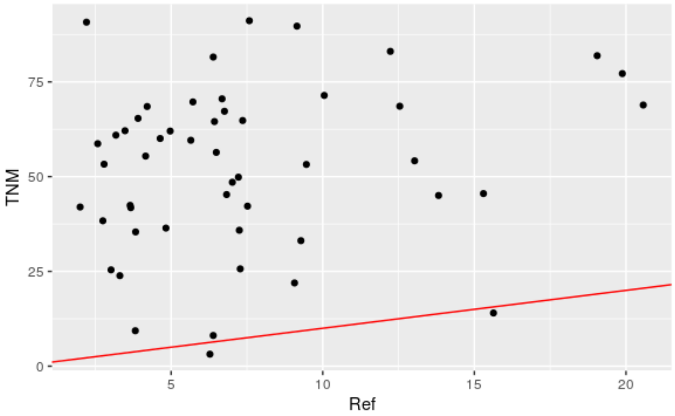


Figure 7: dispersion graph for the Clashscore in the Ref and the TNM. For lower values in Ref (x-axe), higher values in TNM (y-axe) are related. The red line is the diagonal reference line x = y

PDB_code	Ref	TNM
2a33	6.39	81.545
2i3c	5.65	59.586
1xm8	4.21	68.495
2ab1	6.43	64.524
1z8k	2.79	53.278

Table 4: first 5 clashscores of the reference proteins vs. the average clashscore of the TNM ensembles. The first 5 proteins are shown. All the proteins showed this patron of increase in TNM ensembles.

In the *clashscore* instance, its rise in the TNM ensembles was striking. In ***Table 4***, a rising of the order of 19 times can be seen in the 1z8k protein. In ***Figure 6*** and ***7***, the difference between the Ref clashscores and the one from TNM ensembles can be visually appreciated: the clashscore shoots up in the TNM ensembles. The possible reasons for this will be seen at the discussion. To be absolutely sure that this difference is statistically significant, a Welch t-test is performed. The null hypothesis is that there is no difference in the means:

```
data:  ref_tnm_CS$Ref and ref_tnm_CS$TNM
t = -14.254, df = 53.377, p-value < 2.2e-16
mean of x mean of y
  7.41840  52.47539
```

The difference factor in the CS mean was 7.08 times higher for the TNM files. The p-value 2.2e-16 is consistently minor than the convention 0.05 α. The null hypothesis is rejected and the alternative hypothesis is accepted. So it can be assured that the TNM process affects the CS.

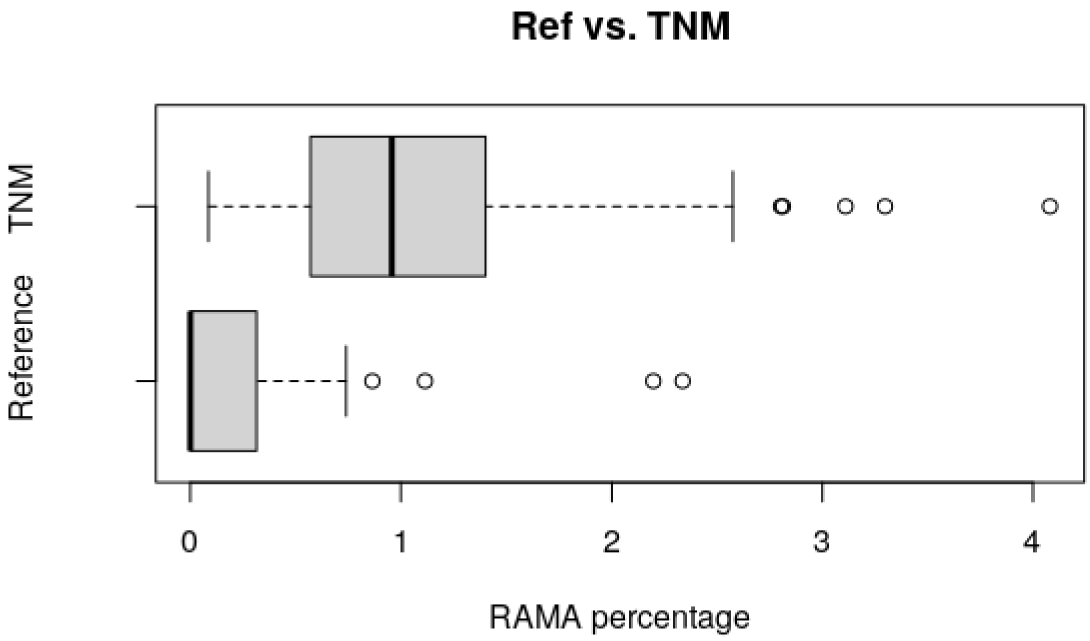


Figure 8: boxplot with the distribution of the Ramachandran percentages. The upper graphic is the TNM ensembles set of data, while the one below is the Reference.

The next parameter to be analyzed was the Rama_iffy. It is a measure of the backbone, as it points to the local structures with very unlikely angles. The results obtained can be seen at ***Figure 8.*** In the Reference structures the percentage of the Ramachandran outliers is really low. In fact, 29 out of 50 had 0 Ramachandran outliers, and the others had very low Ramachandran percentages and were below 1%, except three of them: 1ydw, 1xj5 and 2be4. To get a statistical reference to judge if the condition Ref-TNM influence the Ramachandran percentage, a t-test was run. The null hypothesis affirms that there is no difference. The p-value is 8.452e-09, which is lower than the reference value 0.05 α . We accept the alternative hypothesis: Ref-TNM condition does influence the Rama_iffy. The mean factor difference was 4.87 times higher Rama_iffy in the TNM.

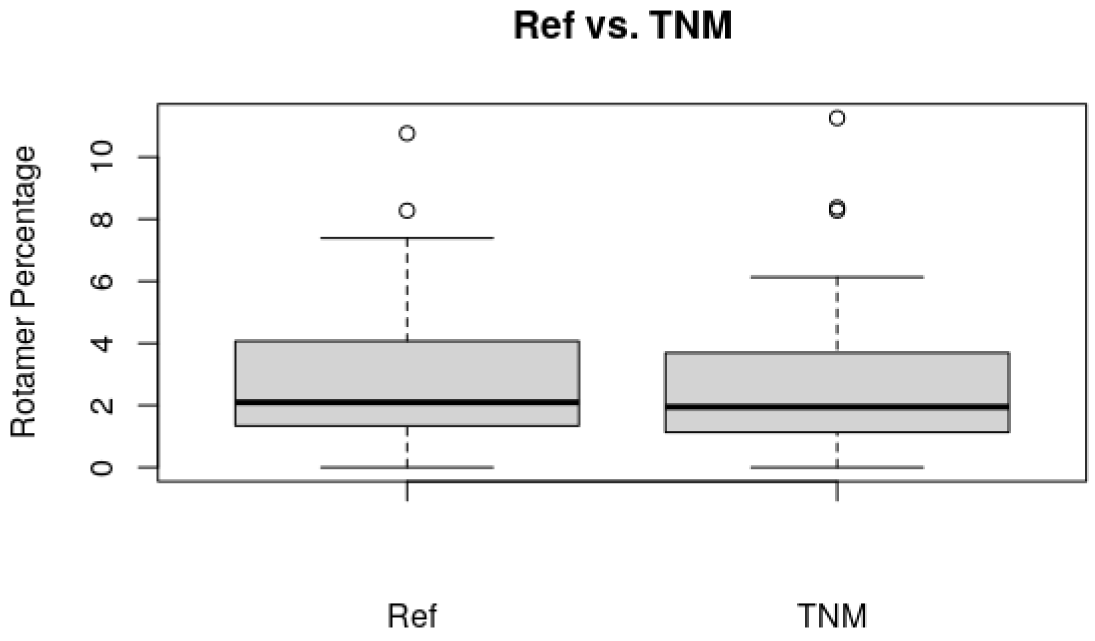


Figure 9: boxplot with the distribution of the Rotamer outlier percentages. The graphic on the left is the Reference dataset, while the one on the right is the TNM ensemble dataset.

At first sight, the TNM ensembles does not seem to have a strong effect in the Rotamer percentage of outliers ***See Figure 9.*** None the less, the only way to ensure that this is true is performing a statistical test. A t-test with the null hypothesis that the condition Ref-Ens does not affect the Rotamer percentage is run, and we get a p-valor of 0.7184. It is higher than the reference value (0.7184 >> 0.05α). So we accept the null hypothesis and it can be affirmed that the condition does not affect the Rota_out. The possible reasons will be seen at **6. Discussion**.

5.2.2 Quality of the TNM Ensembles according to the target RMSD

In the previous point, we have seen that the TNM ensemble process causes a general decrease in the quality of the model. Also, the local measures of CS, Rota_out and Rama_iffy were seen, as these are components of the Mpscore indicator. The CS and Rama_iffy rose in the TNM ensembles, while the Rota_out stayed the same.

During the TNM assembly process, one of the main parameters that can be changed is the target RMSD. As previously said, it determines the Root mean square deviation from the reference structure that the modeler will target while performing. Still, that does not mean that the RMSD between the ensemble and the reference would be the target, but a close one. This is because during the translation from angle coordinates to Cartesian coordinates, a lineal approximation is performed, leaving room for a small variation in the RMSD.

Changing the RMSD target will influence the ensembles created, as it limits or expands the freedom that atoms can fluctuate from the reference structure. To study the effects of this, we performed 30 ensembles for every RMSD value to guarantee an adequate sample size: 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2. For each 30 ensembles, we did the average.

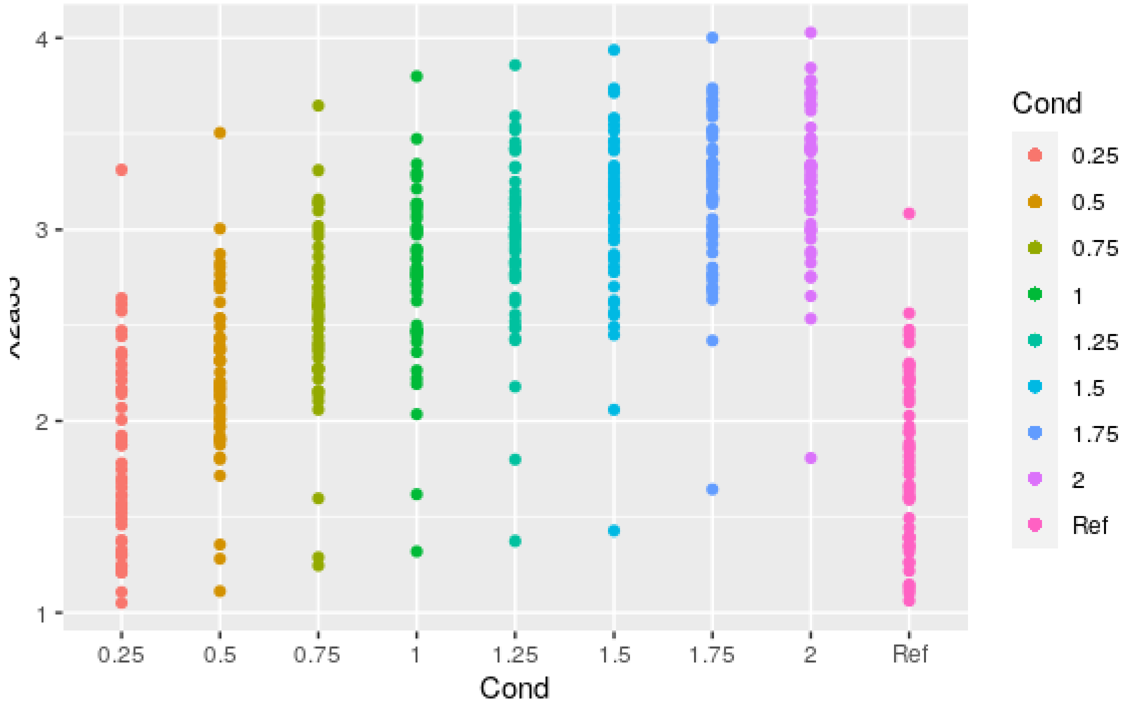


Figure 10: scatter plot representing the relation between the Mpscore (y-axe) and the target RMSD used for the assembly (x-axe). Every point is a single protein. When increasing the target RMSD, the MPscore seems to rise in a lineal way.

See Figure 10. A general trend can be seen in the graph: the higher the RMSD target, the higher the Mpscore (that is, worse quality of the model). On the last RMSD targets (x1.75) (x2) an overlap in the Mpscore is almost reached, which could indicate "an incoming plateau". This can be seen easily on the isolated upper group, which corresponds to the reference protein structure *1ydw.*

As the relation between conditions seems to be lineal, we perform a lineal regression coupled analysis. For this purpose, we choose the x1 condition (the original ensembles) and the x0.5 condition. ***See Figure 11***. The lineal model represented (red line) has the following output:

```
          Coefficients:   Estimate  Std. Error    t value Pr(>|t|)
(Inter)       -0.26127      0.16323   -1.601      0.116
X1            0.90338       0.05816    15.531        <2e-16 ***
Multiple R-squared:  0.834,     Adjusted R-squared:  0.8306
```

If there was no-relation between the Mpscore and the target RMSD condition, we would have similar Mpscores for both x1 and x0.5. The regression line would be y = x, and the coefficients of the model would be 0 and 1. But this is not happening: for every Mpscore point in x1, the ensembles have 0.903 in x0.5. The p-value is <2e-16, which is way lower than the 0.05α convention. The R-squared is relatively high (0.834 on 1), which means that the data fits well the lineal model. We can assure that the x1-x0.5 condition affects the Mpscore.

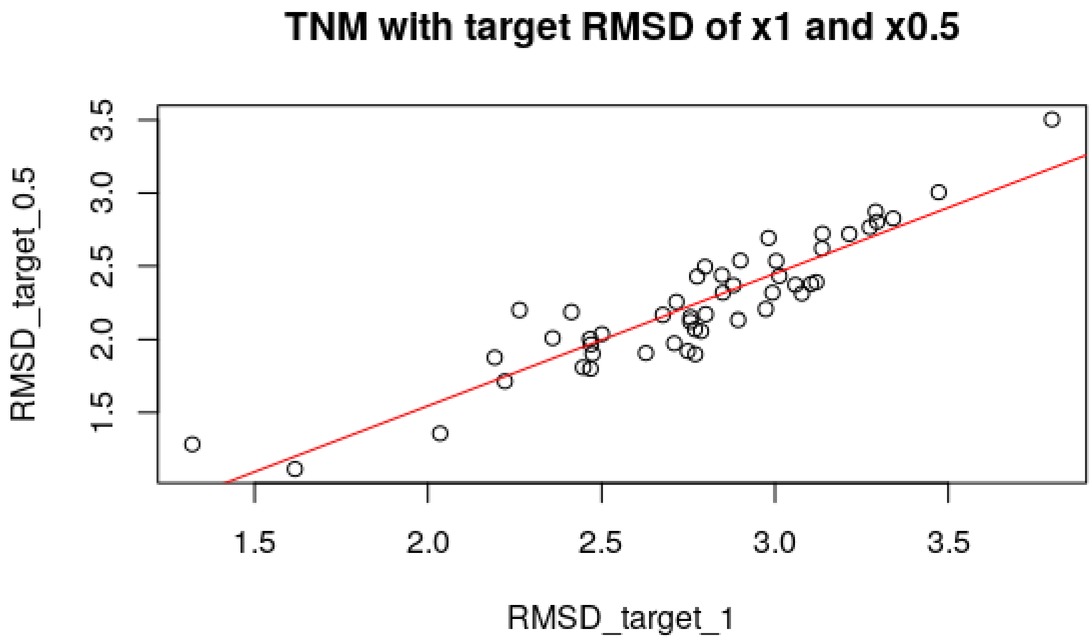


Figure 11: dispersion graph of MPscore from TNM ensemble x1 and TNM ensemble x0.5. The regression line of the lineal model is represented by the red line.

An apparent overlap between the Mpscores of the highest target RMSD values had been observed. To see the distribution of the data, we made a box-plot with all the conditions (***Figure 12***). The data distributions effectively seemed to get stuck when getting to high target RMSD values (1.5, 1.75 and 2). Besides, this graph also allows to see that the Mpscores of the X0.25 condition have a similar distribution than the Reference ones.

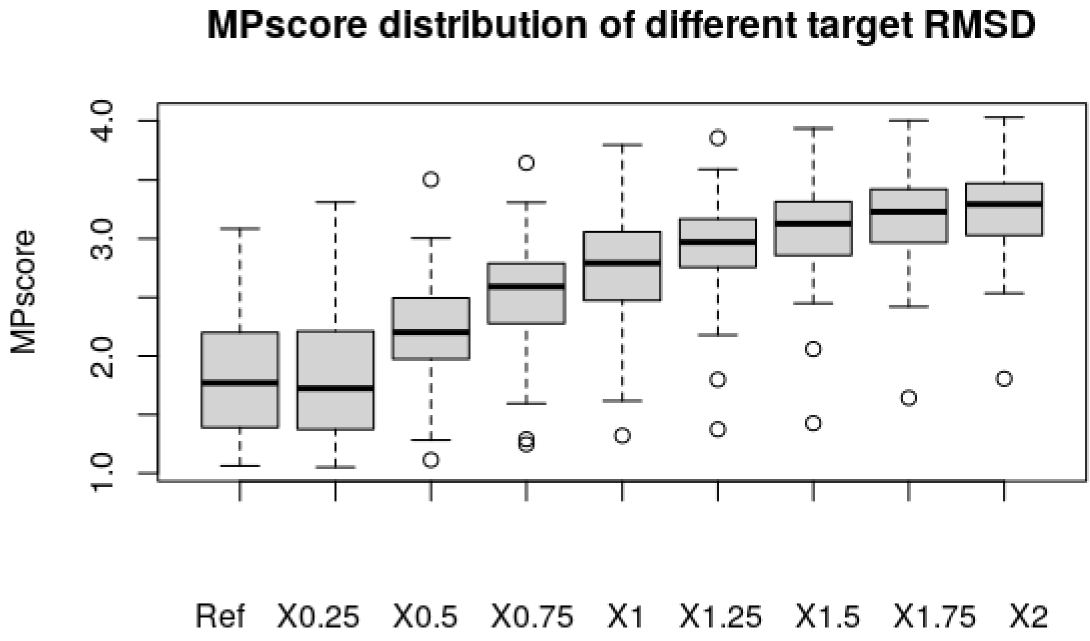


Figure 12: boxplot with the MPscore distribution of the Ensembles with all the different target RMSD. The reference MPscore is showed on the left.

A further study is performed by making Welch t-test analyzes to the adjacent conditions to find the point where “a single 0.25 step” on the target RMSD affects the Mpscore mean. The analysis are summarized by the p-value:

X2_X1.75 = 0.292
X1.75_X1.5 = 0.2192 X1.5_X1.25 = 0.1205 X1.25_X1 = 0.07374
X1_X0.75 = 0.01257 X0.75_X0.5 = 0.0009884 X0.5_X0.25 = 1.805e-05

Only on those couples where the p-value is lower than 0.05 α, we can reject the null hypothesis and assure that there is a statistical significant difference in the means. The higher “0.25 step” where a statistically significant change of Mpscore mean can be seen is on the X1_X0.75 one. From there, the means seem to be more different every time, because the p-values continue to decrease. Nevertheless, this does not mean than the higher target RMSD conditions do not have a difference with lower conditions. For example, if we perform the same t-test study with X2 and X1.5, a p-value of 0.02463 is obtained, which is already lower than the 0.05 α limit and sets the difference in condition as significant.

Finally, to confirm that the differences found between the Ref Mpscore distribution and the X0.25 are not significant, a final Ref-X0.25 test is performed. A p-value of 0.7631 is obtained: 0.7631 >> 0.05 α; so the null hypothesis is accepted and no significant difference in the means is considered.

All this results suggest that the model quality improves (Mpscore decreases) when lowering the target RMSD, and the model quality gets worse when elevating the target RMSD. At least from a static perspective, as it will be discussed at **6.** However, this influence is not uniform, as the MPscore gets stuck when rising the RMSD from X1.5; the model does not get much worse. On the other hand, it also gets stable when lowering the RMSD, as the TNM ensembles atom coordinates get really close to the ones of the reference structure.

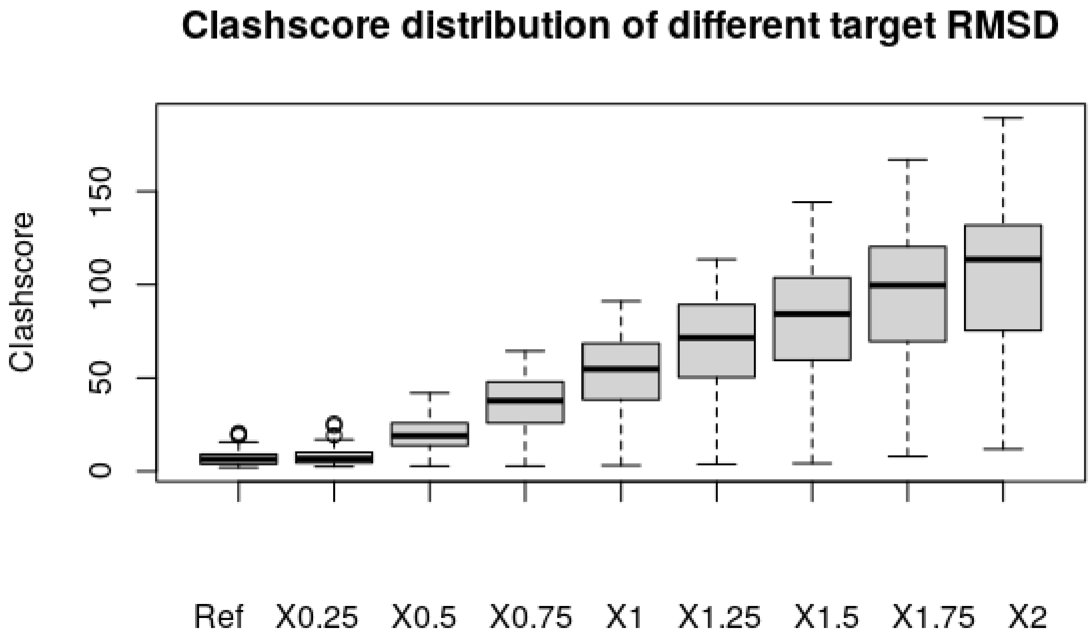


Figure 14: boxplot with the Clashscore distribution of the Ensembles with all the different target RMSD. The reference Clashscore is showed on the left.

In order to make a closer analysis, the local parameters Rota_out, Rama_iffy and CS are studied when changing the target RMSD condition. In the case of the clashscore (CS), ***See Figure 14***, its rise is also proportional to the target RMSD. The CS distribution is very close to the reference one in the x0.25 condition, and it gets higher when increasing the target RMSD. Again, it seems to have a lineal behavior. We perform several statistical t-test where the null hypothesis is that there are not statistical meaningful differences between the means (***Table 5***).

	Ref-X0.25	X0.25-X0.5	X0.5-X0.75	X0.75-X1	X1-X1.25	X1.5-X1.75	X1.75-2
Mean factor	1.12	2.4	1.79	1.460	1.280	1.140	1.110
P-value	0.3817	1.28E-11	8.668E-09	2.745E-05	0.003251	0.09678	0.1727

Table 5 : several Welch t-test performed on the CS data from the TNM ensembles with different target RMSD and the Ref structures.

The level of significance is α = 0.05, so when comparing it to the p-value we can know if the difference in the mean is statistically significant. Even if there is a slight difference in CS between the Ref and the X0.25 target RMSD condition (factor = 1.12), it is not significant. In the target RMSD variations, we can see an evolution in the p-values (related to the variation of the mean factors). First, In the 0.25 target RMSD steps, the factor increases and it becomes significant (0.25-0.5; 0.5-0.75;0.75-1; 1-1.25). The 0.5-1 target RMSD range has the highest variation. From target RMSD = 1, this tendency breaks and seems to develop into a plateau without great CS change. This data agrees with the obtained in the Mpscore. Again, even if the higher target RMSD conditions do not show significant variation between adjacent conditions, they show it when compared with lower target RMSD: X2-X1.5 test has the p-value = 0.002884 < 0.05 α .

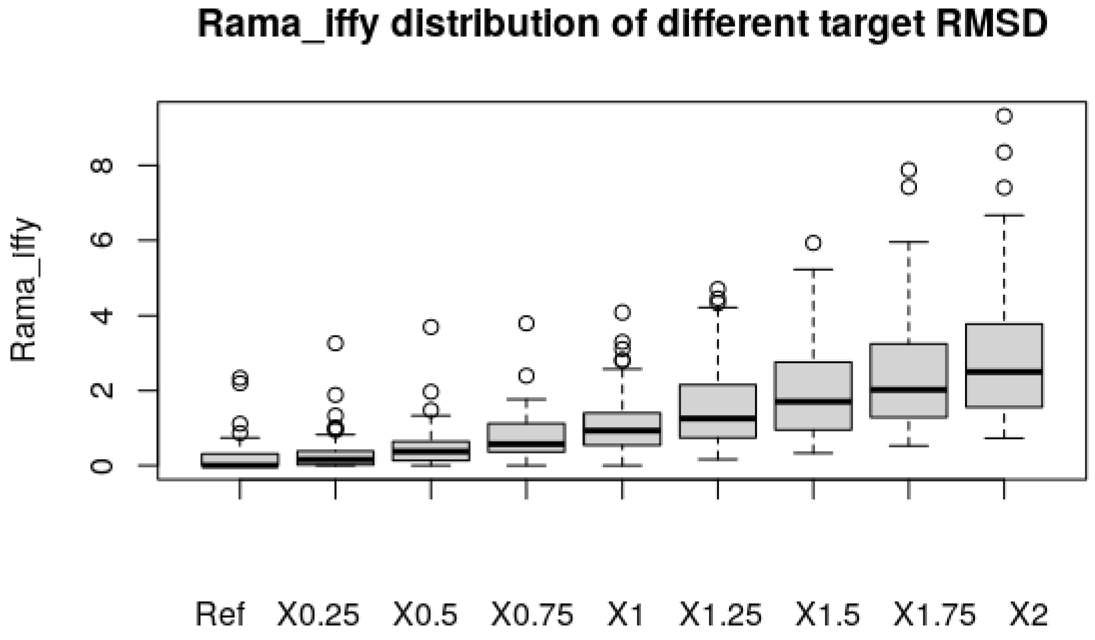


Figure 15: boxplot with the Rama_iffy distribution of the Ensembles with all the different target RMSD. The reference Rama_iffy is showed on the left.

Next, the Ramachandran percentage of conformations outside favored region is studied. ***See Figure 15***. The Rama_iffy distribution evolves following the same dynamics than the Mpscore and the CS, and again, the lower target RMSD conditions resemble more the data from the reference structure. **In *Figure 16***, it can be seen how the lower target RMSD ensembles have more lower Rama_iffy values, and how the density curve is displaced towards the right (higher Rama_iffy values) when increasing the target RMSD condition.

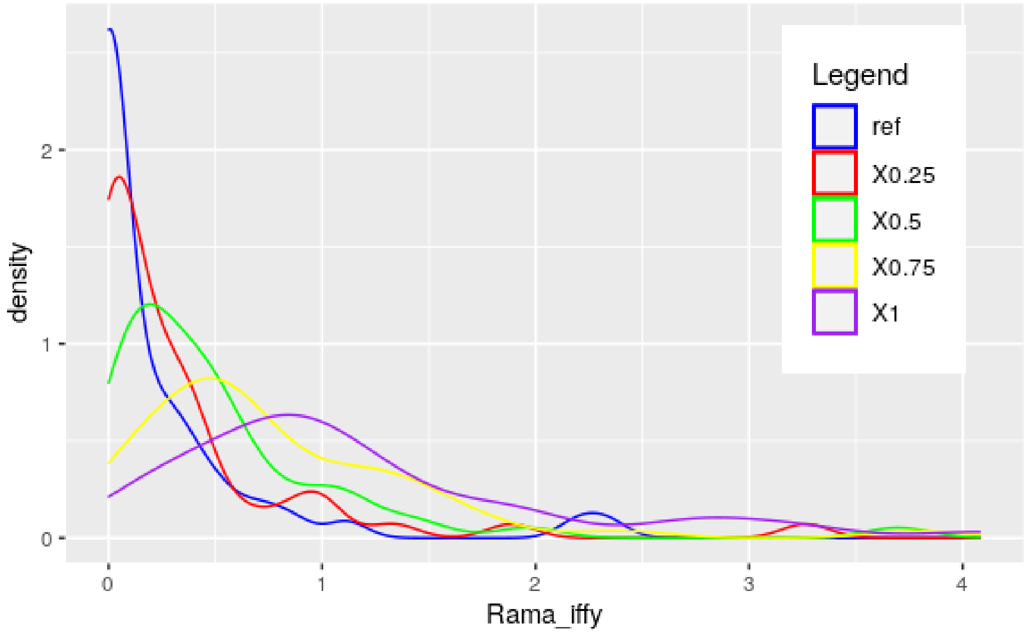


Figure 16 : density curve with the RAMA iffy of several Target RMSD conditions.

The last local parameter is the percentage of outlier Rotamers (Rota_out). As it has been seen in **5.2.1**; the Rota_out parameter does not get involved when the TNM modeling takes places, so, theoretically, we should not appreciate any important difference in Rota_out when changing the target RMSD condition. None the less, it is possible that changing the fluctuations of the modeler affected somehow to the Rotamers.

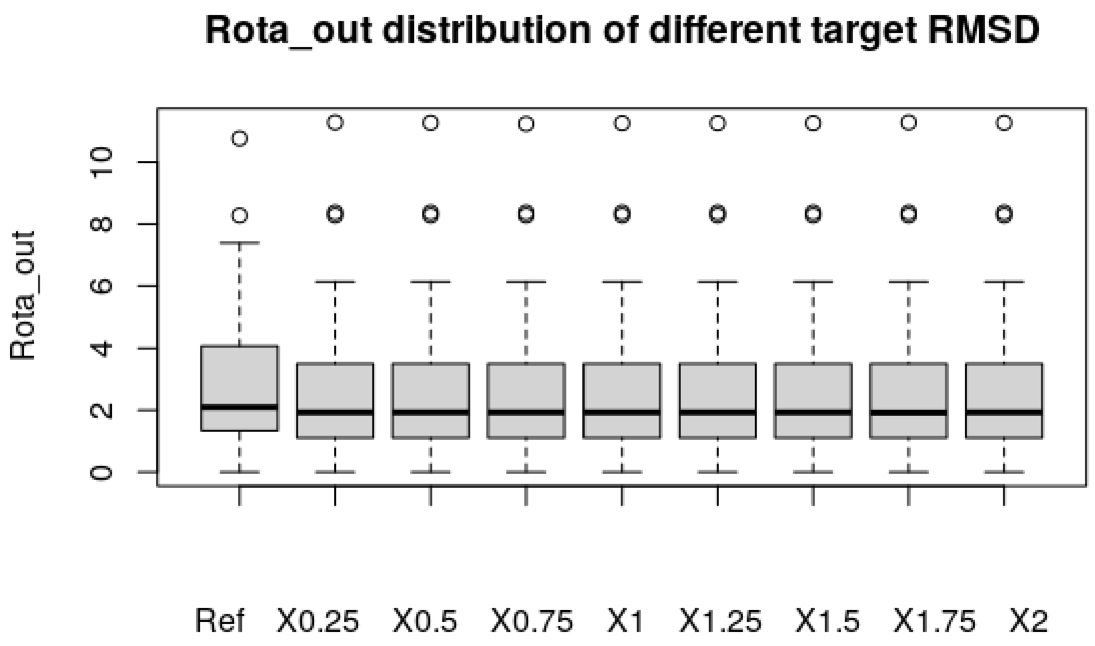


Figure 17: box-plot with the Rota_out distribution of the Ensembles with all the different target RMSD. The reference Rota_out is showed on the left.

***See Figure* 17**. As suspected, the Rota_out does not seem affected at all by the target RMSD condition. In fact, the different Rota_out values of the TNM ensembles are almost the same values, with differences of some decimals due to stochastic variations. None the less, we perform a couple of t-test to be able to sustain this with statistical support:

Ref-X0.25 test: p-value = 0.5485 > 0.05; the null hypothesis is accepted, there is no statistical significant difference between the means.

X0.25-Ref test: p-value = 0.9996 > 0.05; the null hypothesis is accepted, there is no statistical significant difference between the means.

The Mpscore variation observed between sets of ensemble with different target RMSD conditions are explained by the variation in the clashscore and the Ramachandran percentage of outliers from the favored region. Further possible causes, reasons and consequences of these results are considered in **6. Discussion**.

5.3 Refinement of the structures produced by the TNM

Once observed the MPscore behavior according to the setting of target RMSD values; the next objective was to improve even more the quality of the models. SIDEpro was one of the programs chosen for this task. SidePro allows to optimize the side-chains of the structure, giving as output a new file which can be analyzed with MP. SidePro should constitute a good method of refinement for the TNM ensembles, as it does not position side-chains.

The Mpscores of the TNM ensembles created with the standard parameters (RMSD = 1), and the same structures after being refined with SidePro were compared. The data-sets will be call from now on: Ref, for the original struture; TNM_RMSD=1, for the TNM ensembles in standard condition; and SP for these same ensembles after their refinement with SIDEpro.

See Figure 18. The SIDEpro refinement seems to have decreased the Mpscore as expected (the quality has risen). However, a t-test is performed in order to know if this difference is significant.

```
data:  tnm1_sp1$ori and tnm1_sp1$imp
t = 5.5252, df = 93.742, p-value = 2.939e-07
mean of x mean of y
 2.771862  2.220677
```

The data means have a factor of 0.8 of decrease MPscore for the SIDEpro-refined structures. The p-value is 2.939e-07, much lower than 0.05, so this difference is statistically significant. We can affirm then that the SIDEpro process has an effect in the Mpscore.

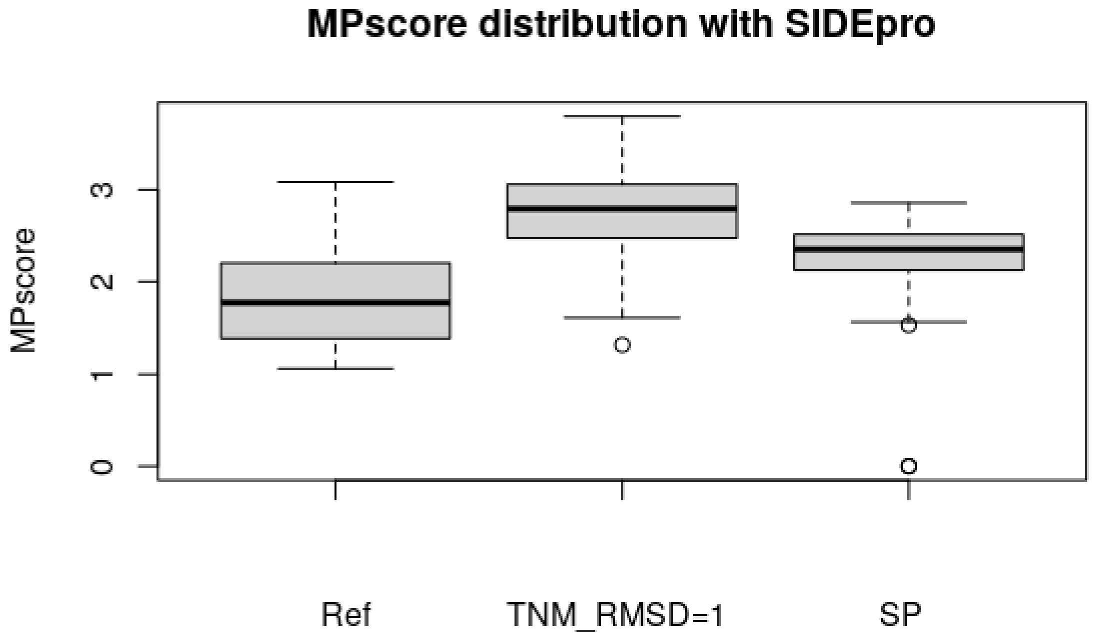


Figura 18: MPscore distribution of the TNM ensembles before (left) and after (right) SIDEpro side-chain refinement.

TNM modeler does not position the side-chains, while SIDEpro is a software specialized in optimizing these side-chains. So, taking a look to the local components, specially the Rota_out should be interesting. This data is extracted and represented.

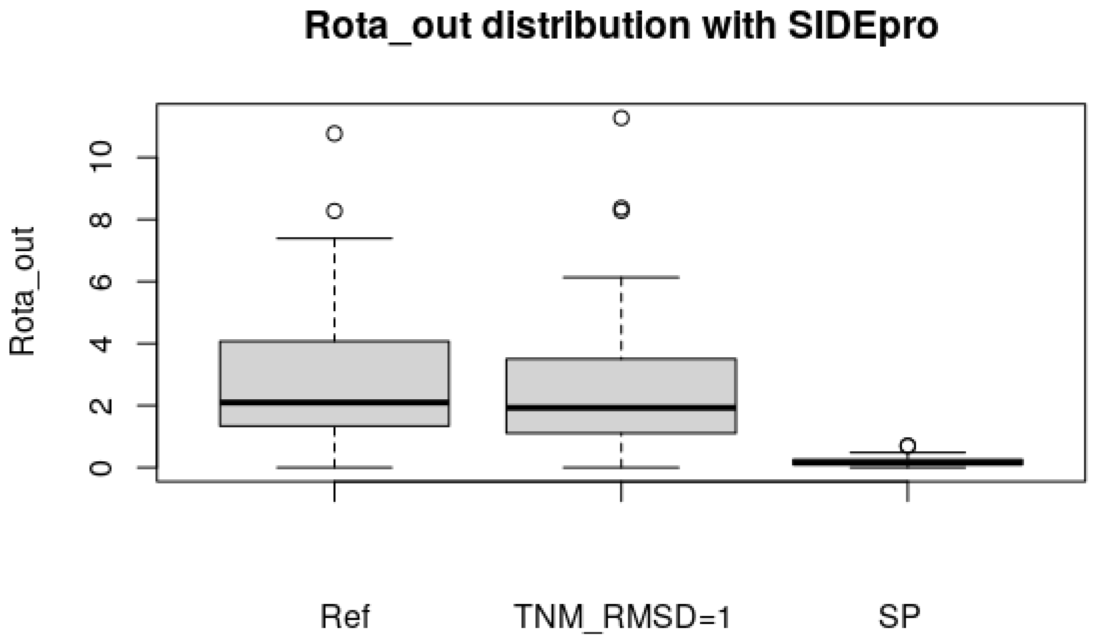


Figure 19: boxplot with the Rota_out (y-axe) of the TNM ensembles before the SIDEpro refinement (TNM_RMSD=1) and after (SP). The reference structure Rota_out is on the left

Figure 19, it can be seen how the percentage of Rotamer. outliers (Rota_out) dramatically decreases after the SIDEpro refinement. On ***Figure 20*** , the change in the data can be easily appreciated. In the original data-set (blue line) the Rota_out extends from 0% to 6%, and a great amount of the data is found between 0% and 3%. After the refinement (red line), most of the Rota_out are located close to 0; with a light pick on 0.75%. The Rota_out have been strongly decreased. To have a numeric statistical measure, a t-test is performed. The mean factor is 13,4 times more Rota_out in the TNM raw ensembles than in the SIDEpro improved: and a p-value of 1.352e-09 << 0.05; so we consider it statistically significant.

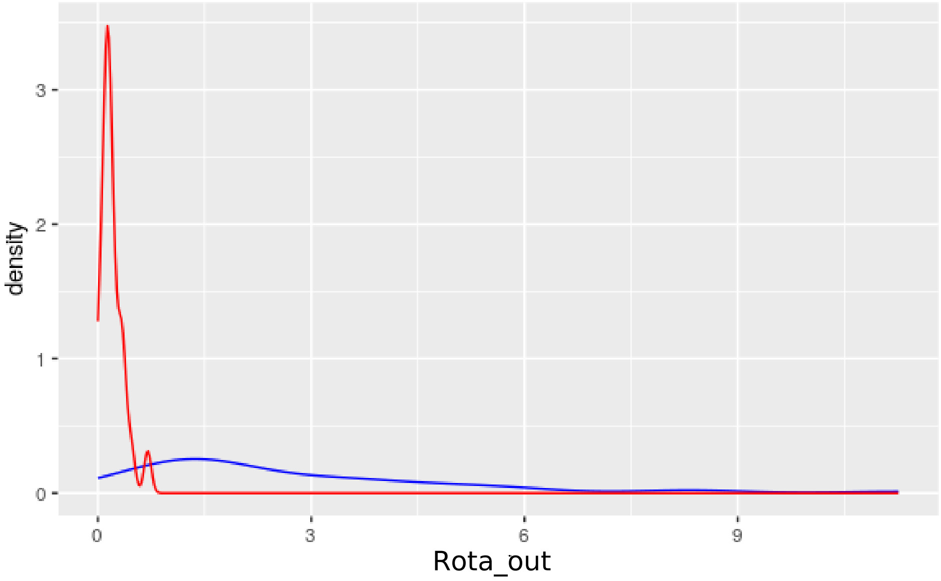


Figure 20: density curve with the Rota_out on the x-axe; and the density or in the y-axe. The blue line represents the original TNM ensemble while the red line represents the same after SIDEpro enhancement.

Next, we continue analyzing the other two components: Ramachandran percentage outside favored region (Rama_iffy); and Clashscore (CS). As the Rama_iffy is a measure of the backbone conformations, and SIDEpro is a side-chain software, it should not have been affected. ***See Figure 21.*** As it was expected, the SIDEpro refinement seems to have little influence in the Rama_iffy component. Still, a slight variation can be observed. The emergence of a fifth outlier point (above the Imp box) is particularly interesting. To see the meaningfulness of this variation, a t-test is performed. The p-value 0.7938 >> 0.05, so the difference is not statistically significant.

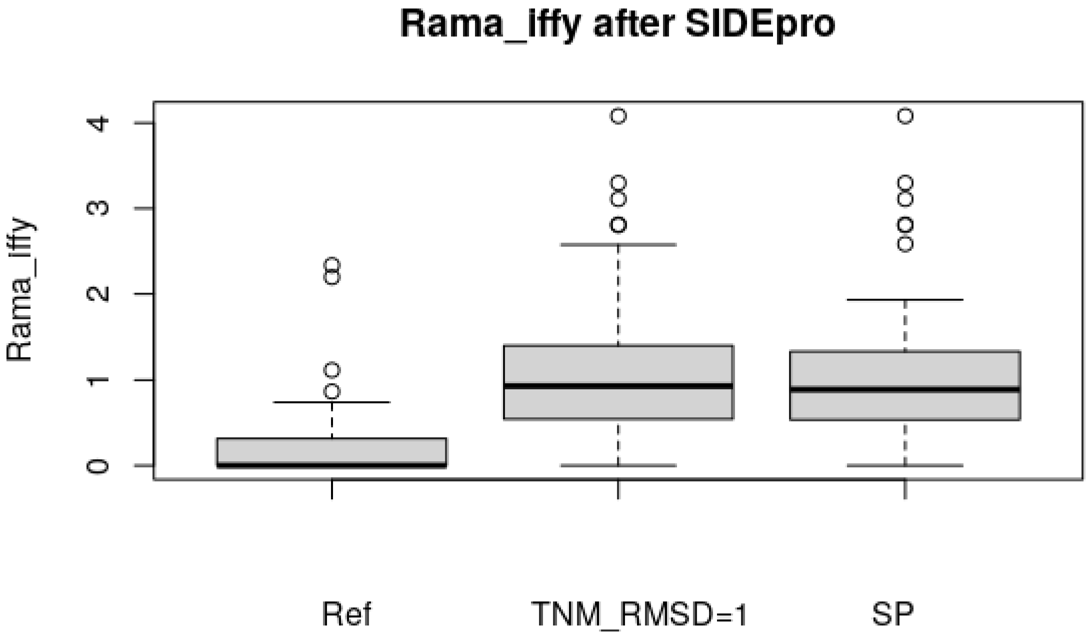


Figure 21: boxplot with the Rama_iffy (y-axe) of the TNM ensembles before the SIDEpro refinement (TNM_RMSD=1) and after (SP).

Finally, the CS is studied. The side-chains have been re-positioned during the SIDEpro improvement, so a change in the CS is expected. During this Rotamer. optimization, probably the number of overlaps between atoms have varied. None the less, the question is if the SIDEpro software, which uses a Rotamer. library to act, optimizes the Rotamers at the expense of the clashscore. ***See Figure 22***.

In fact, it does not increase the CS (number of overlaps > 0.4 ºA for 1000 atoms), but it decreases it. This is probably due to that SIDEpro, besides the Rotamer. library, uses a system to avoid steric clashes. The significance of this data is checked with a t-test: the factor is 1.8 times more CS in the raw TNM ensembles than in the SIDEpro enhanced, and the p-value is 8.303e-09 < 0.05 α; so the change in CS with the SIDEpro condition is statistically significant.

The SIDEpro process produces a drop of the Mpscore, that is, an overall improvement of the quality of the Ensembles. This Mpscore decline is due to the changes in the Rota_out component and the CS component, but non in the Rama_iffy component. The consequences of these results will be considered in **6. Discussion**.

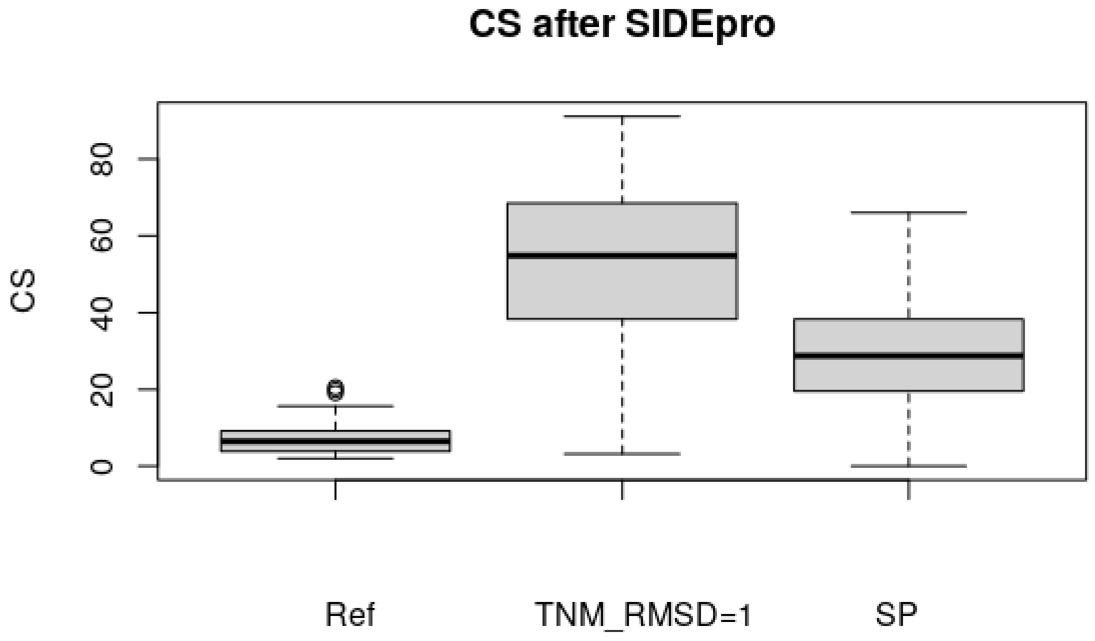


Figure 22: boxplot with the CS (y-axe) of the TNM ensembles before the SIDEpro refinement (TNM_RMSD=1) and after (SP).

5.3.1 Using a force field to minimize the energy and improve the quality

After the SIDEpro enhancement, the refinement of our ensembles is continued by using a simple function of energy minimization. It is done in GROMACS, and we start from the output structures of SIDEpro, in order to know how much can we improve the structure with both methods combined. The same approach is followed: generating new structures optimized and then analyzing them by the Molprobity criteria.

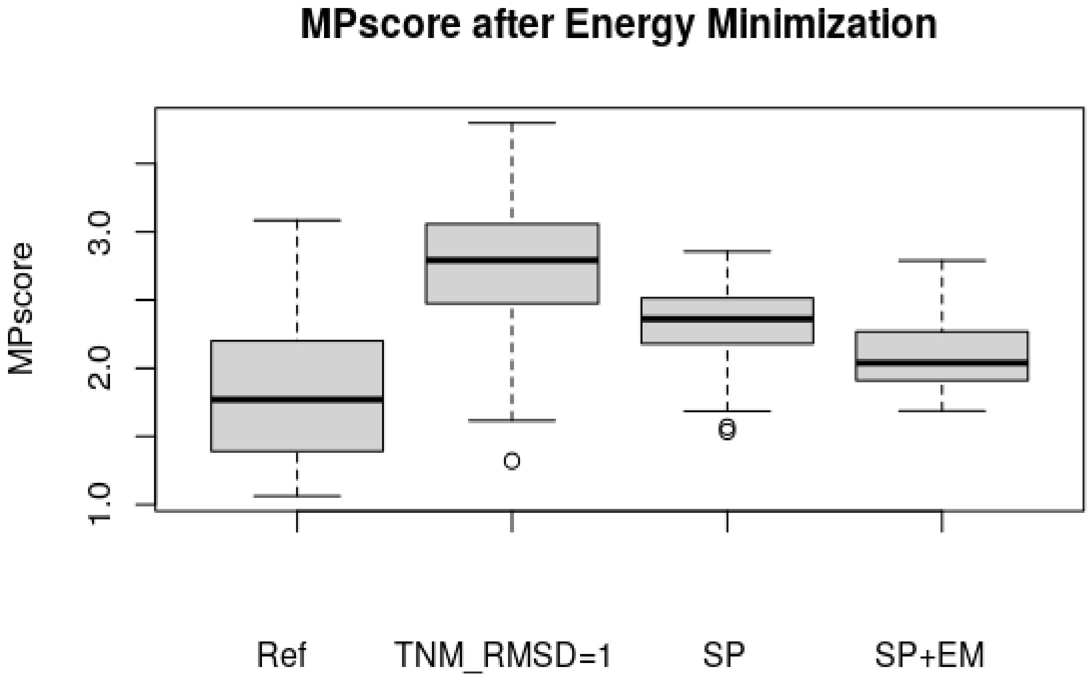


Figure 23: boxplot with the MPscore of the SIDEpro improved TNM ensembles (SP) and the same structures after minimize its energy with GROMACS (SP+EM). On the left, the reference structure and the TNM_RMSD=1 ensembles MPscore distribution are displayed

Figure 23, it can appreciated a decrease of the Mpscore distribution after the EM is performed. This means that the average quality of the models improve. A t-test to confirm the suspicions is carried out.

```
data:  tnm_SP_minimized$SP and tnm_SP_minimized$MIN
t = 3.2288, df = 94.165, p-value = 0.001712
mean of x mean of y
 2.307534  2.118551
```

The factor change of the means is 1.09 MPscore for the EM refined. It is not a great improvement compared to previous steps. None the less, it is enough to be statistically significant; p-value = 0.001712 < 0.05 α level of significance. Regardless the results, the low mean factor suggests that the changes on the local parameters would be small.

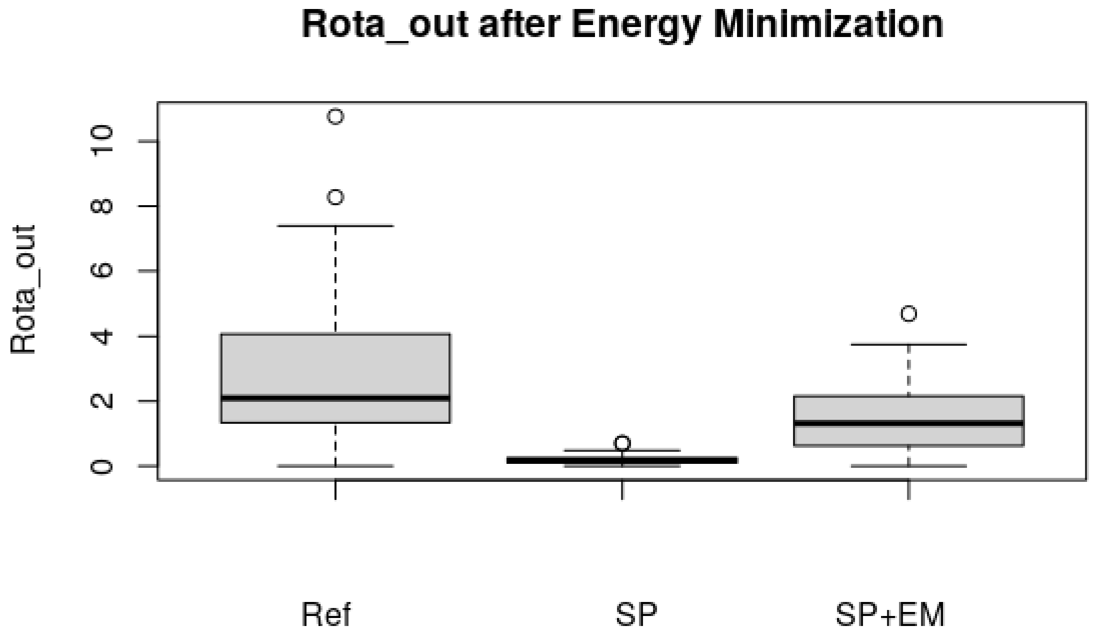


Figure 24: boxplot with the Rota_out of the SIDEpro improved TNM ensembles (left) and the same structures after minimize its energy with GROMACS (right).

See figure 24. The Rota_out distributions goes notably higher after doing the EM; which means that more Rotamers are considered as outliers. A statistic test is performed:

```
data:  tnm_SP_minimized_ROTA$SP and tnm_SP_minimized_ROTA$MIN
t = -7.3966, df = 49.734, p-value = 1.485e-09
mean of x mean of y
0.2026014 1.4446531
```

The factor in means is 7.22 times Rota_out higher after the EM process in relation to the ensembles after the SIDEpro refinement, and the significance is high: p-value = 1.485e-09 < 0.05 α level of significance. The EM process affects negatively to the results, as it rises the number of Rotamer outliers. Probably, the GROMACS software has generated worse rated Rotamer conformations when trying to minimize the global energy of the structure.

Next, we analyze the Ramachandran percentage outside favored region (Rama_iffy) . ***See Figure 25.*** The Rama_iffy distribution seems to have changed a little. The firsts and third quartil have become wider, so does the interquartile range, as well as the maximum value. Another t-test is performed (*see below*).

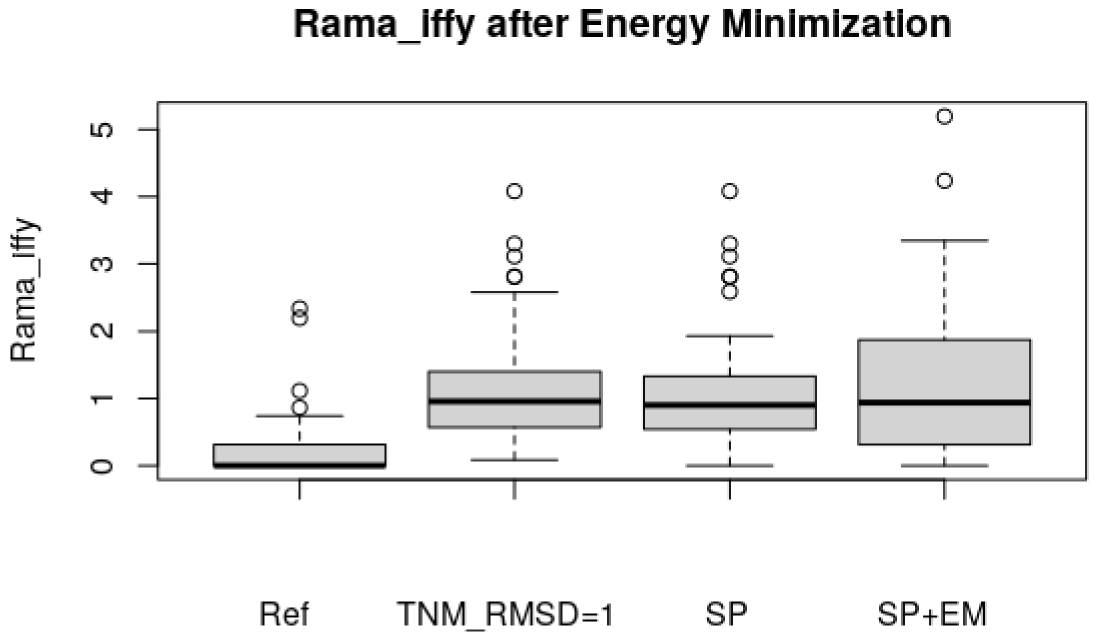


Figure 25: boxplot with the Rama_iffy of the SIDEpro improved TNM ensembles (left) and the same structures after minimize its energy with GROMACS (right).

```
data:  tnm_SP_minimized_RAMA$SP and tnm_SP_minimized_RAMA$MIN
t = -0.31314, df = 90.783, p-value = 0.7549
mean of x mean of y
 1.131456  1.196898
```

The Rama_iffy means seems to be pretty similar in both cases. Its factor is 1.05, with a p-value = 0.754 > 0.05 α. The null hypothesis is accepted and no-differences are considered statistically significant. None the less, it seems that the variance has been increased on the ***Figure 25***. To check statistically if this difference of variances is meaningful, a F-test is performed. We take the populations as normal, and the null hypothesis is that they have the same variance.

```
F test to compare two variances

data:  tnm_SP_minimized_RAMA$SP and tnm_SP_minimized_RAMA$MIN
F = 0.61326, num df = 48, denom df = 48, p-value = 0.09358
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.345924 1.087192
sample estimates:
ratio of variances
        0.6132583
```

The ration of variances is 0.613; but the p-value = 0.09 > 0.05 α, so it is not considered as statistically significant. The null hypothesis is accepted and we cannot affirm that there is a difference in the variances. So finally, we take the EM process as not statistically influential on the Rama_iffy parameter.

The EM process improved the Mpscore (by making it lower). None the less, the Rota_out rose (which should make the Mpscore higher and a worse model) and the Rama_iffy stayed the same. As the Mpscore is determined by these two parameters and the clashscore (CS). Therefore, it is expected that the decreasing of the Mpscore in the EM condition is explained mainly by the CS. ***See Figure 26***. The CS distribution has fallen down after the EM process. A statistical test is performed to understand its magnitude.

```
Welch Two Sample t-test

data:  tnm_SP_minimized_CS$SP and tnm_SP_minimized_CS$MIN
t = 9.3532, df = 56.758, p-value = 4.283e-13
mean of x mean of y
 29.59910  11.00633
```

We find a tighter variance and a much lower mean. The mean factor is 2.69 times CS higher the SIDEpro ensemble in relation to the EM output; and the p-value = 4.283e-13 << 0.05 α shows a substantial significance. These results explain the decreasing of the Mpscore regardless the Rota_out and the Rama_iffy.

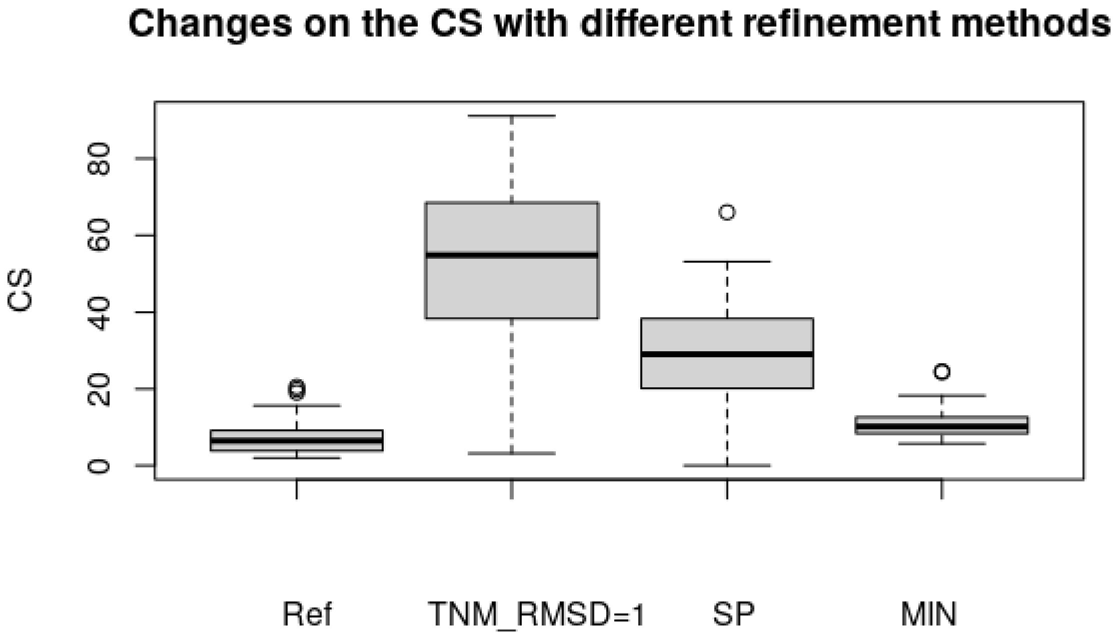


Figure 26: boxplot with the CS of the SIDEpro improved TNM ensembles (left) and the same structures after minimize its energy with GROMACS (right).

6. Discussion

The deciphering of mechanisms that guide conformational changes in proteins are best described with torsional angles, which are well suited to evaluate the influence of each residue to the conformational changes. The Torsional Network Model (TNM) (Mendez and Bastolla, 2010) allows to translate from Cartesian coordinates, which are the typical output of structure determination methods, to angle coordinates. Thus, it offers an efficient way to represent protein structures and their fluctuations. In this thesis, we create, evaluate and refine TNM ensembles from a set of reference protein structures defined experimentally (Levin et al., 2007). An approximation used in Bastolla and Dehouck, 2019, is developed: the evaluation is performed by Molprobity analysis, and the refinement is done by SIDEpro. Furthermore, a new approach is taken when refining the ensembles by Energy Minimization (EM).

The statistically supported results obtained during this thesis showed: (i) the TNM assembly process produce an increase on the Mpscore and thus, a reduction of the overall structure quality (***point 5.2***); (ii) the local parameters affected are the Ramachandran percentage outside favored region (Rama_iffy), and the Clashscore (CS); while the percentage of Rotamer outliers (Rota_out) is not significantly influenced (***point 5.2.1***); (iii) the quality of the TNM ensembles increases with lower target RMSD and decreases with higher target RMSD (***point 5.2.2***); (iv) the SIDEpro reconstruction produce a descent in the Mpscore, therefore a gain in model quality, sustaining previous results (Bastolla and Dehouck, 2019) ; and (v) the EM process improves the general quality of the model by lowering the average Mpscore, thanks to a great drop in the CS, but reverts partially the Rota_out decrease reached by SIDEpro (***point 5.3.1***). A summary of important results can be seen at ***Table 6***.

	Ref - TNM_RMSD=1	**RMSD=1 – RMSD=0.5**	**TNM_RMSD=1 – SP**	**SP – SP+EM**
MPscore	1.549	0.903	0.800	0.920
CS	7.080	0.158	0.552	0.371
Rama_iffy	4.87	0.448	*No_significant*	*No_significant*
Rota_out	*No_significant*	*No_significant*	0.074	7.220

Table 6: summary table with the parameter factor means obtained from the original condition (header, left side of the hyphen) to the final ensemble condition (header, right side of the hyphen); (Original - Final). The first column shows the parameter analyzed. If the factor is lower than 1, it means that the score has decreased; if it is higher, it has increased. By multiplying the Original score by the factor, the final score is obtained. "No_significant" points to the relations that were not statistically significant. Ref = Reference, RMSD=1 = TNM ensembles with target RMSD =1, SP = TNM ensembles RMSD=1 after SIDEpro, SP+EM = TNM ensembles RMSD=1 after SIDEpro and Energy Minimization.

During the ***TNM assembly process,*** the average Mpscore increment in relation to their Reference structures and therefore, the loss of quality of the TNM ensembles, is due to the increment of the Rama_iffy and the CS. The TNM modeler translates the Cartesian coordinates from the Reference structure to angle coordinates, and then, the pdbs are reconstructed to Cartesian from the angle coordinates in order to be compared. During this process, small local variations of torsion angles can induce large global Cartesian displacements since an approximate match of the Cartesian coordinates can be purchased with different values of torsion angles. Thus, deviations from the original coordinates are carried out. These uncontrolled deviations are probably the cause of the quality loss of the models: the backbone gets Ramachandran-non-favored conformations and the number of overlaps grow, rising the clashscore.

However, this is a mandatory limitation of the TNM. As it uses less degrees of freedom (the torsion angles), information about the structure is lost during the assembly-reconstruction process, so it is rebuilt with these local variations. On the other hand, this same approach allows to identify important sets of residues that contribute to conformational changes and reveal new options for the design of altered functional dynamics and grants a computational space save (as the information stored is less).

The CS shoots up in the TNM ensembles, in fact, it is 7 times higher than in the reference structure. This huge rise was not expected, because it is high even taking into account possible dynamic protein movements. It points to a TNM aspect which should be reviewed and improved. It could be done by adding an overlap control or an energy function when translating the coordinates to avoid this event. The Rama_iffy grows 4.87 times, which is also a great loss. None the less, this problem is more difficult to solve, as the Ramachandran conformations are set by a library which only contemplates static conformations. Maybe, solving the CS problem would indirectly contribute to sort out this trouble too. The reason why Rota_out is not affected is simple: the TNM does not position the side chains, so they are found in the same state as in the Reference structure.

The target RMSD determines the “freedom of movement” that the TNM can take to position atoms in relation to the reference structure. The higher it is, the more the ensemble coordinates will vary from the Reference. This can limit the deviations that the TNM produces during the reconstruction to Cartesian coordinates, as it constraints the translation possibilities from the torsion angles to the Cartesian coordinates. It explains why the lower the overall target RMSD is, the lower the Mpscore and higher the quality. When limiting the target RMSD, the deviations generated by the TNM are being limited too.

The target RMSD 0.25 condition and the Reference have statistically the same quality. Because of the constraint in the reconstruction, the ensembles created are more or less the same structure than the Reference, as it can be confirmed in the CS (***Figure 14***), the Rama_iffy (***Figure 15***), and the Rota_out (***Figure 17***). It is highly probable that, if lower than 0.25 target values were used, the same results would be obtained. None the less, this question would need to be further explored in future studies.

On the other hand, when the TNM has more freedom to translate the angle coordinates, more and more variations are included in the ensembles. Therefore, when increasing the target RMSD, as the variations increases, also do the deviations, which rises the Mpscore (See progression of Mpscore with the target RMSD, ***Figure 13***). The higher target RMSD values 1.5, 1.75, 2; seem to head towards a plateau. None the less, we do not have enough data to affirm this. This question, as the previous one, would be interesting for future studies.

Lowering the target RMSD parameter could seem like a good choice to improve the TNM ensemble quality. However, abusing of this is useless. Shall we remember that the objective of the TNM is studying and explaining the dynamics of a protein. If the ensemble creation is constraint to be almost equal to the reference structure, no fluctuation would be taken into account and the use of the TNM would be non-sense. Another factor which has been completely ignored is the temperature. In order to catch the fluctuations produced by the thermal conditions, certain RMSD is necessary.

The data collected though, indicates than an adequate target RMSD can make the difference between a good quality ensemble and a poor quality one. Therefore, looking for an optimal RMSD which enables certain mobility from the reference structure, while keeping the ensemble quality, should be key for TNM studies. This target RMSD should adapt to the sample protein and the size of the dynamics regarded. For example, the target RMSD in an allosteric protein assembly should be way lower than in a transporter-protein, as the last one suffers a greater replacement during its dynamics.

The factors of the RMSD=1 and RMSD=0.5 target conditions (***Table 6***), show a consistent descent of the MPscore when lowering the RMSD. The CS drops down with a factor of 0.158, which means that a loss of the 84.2% CS has taken place from the higher RMSD condition to the lower. This decrease is as shocking and abrupt as the increase of CS in the Ref-TNM step. This points to the choice of an adequate RMSD target as a vital parameter to avoid excessive overlaps in the TNM ensembles; and solves, at least partially, one of the weaknesses we had found in the TNM assembly process. At the same time, the Rama_iffy is also reduced by a factor of 0.448, which also marks the target RMSD as a crucial parameter for keeping the integrity of the backbone.

The SIDEpro refinement is appropriate for the TNM assembly process. The reason is that TNM does not position side chains, so a great improve could be reached with this approach. In ***Figure 18***, we can see that the SIDEpro refined ensembles have a lower Mpscore (by a factor of 0.8) than the original TNM ensemble and therefore, the quality is better. Still, it does not get to the same quality as the Reference structure. This results agree with those obtain in a previous study (Bastolla and Dehouck, 2019). The main component of this improvement is the Rota_out, as can be seen in ***Figure 19***. The percentage of Rotamer. outliers has a radical drop after running SIDEpro on the structures. The change factor is 0.074, nearly 0, which means that the Rotamer outliers have dropped off. SIDEpro goes through the side-chains several times and compares them with a Rotamer library, re-setting their orientation so the geometrical and charge incompatibilities are minimized (***See Figure 2, A and B***).

However, the Rota_out component is even lower than in the Reference structures. This means that the side-chains in the models after SIDEpro are better than the experimental ones. We know that the Reference structures are high-quality well-contrasted proteins, and the possibility of the ensembles being even better than the experimental structures is unlikely and unrealistic. Therefore, it seems that the SIDEpro produced an over-fitting of the side-chains to its Rotamer library.

The Rama_iffy component is not affected by the SIDEpro optimization, as the backbone structure is not modified during the process. Quite surprisingly, the CS component is improved, by a 0.554 reduction factor. The CS has been reduced a little more than the half of the original CS. The Clashscore measures the backbone overlaps and the side-chains overlaps, so improving the last ones using a Rotamer library could affect the CS. Nevertheless, it is curious how a Rotamer library is able to reduce the CS, a space-energy related parameter. This is probably due to the fact that the high-quality library which SIDEpro uses has good conformations with low overlaps. Therefore, reconstructing the side-chain from these high-quality, low-overlap Rotamers produces the indirect decrease of the CS.

The SIDEpro approximation enables to keep the TNM backbone structure (so important when studying dynamics) while improving the quality of the model. Side-chain conformations can be also really important when dealing with protein dynamics, as some residues are vital in determined functions. TNM uses just the φ and ψ torsion angles of the backbone, leaving the side-chains aside. The SIDEpro software can deal with this unturned stone and be a great complement for the TNM.

Nevertheless, again, the improvement of the model quality does not mean necessarily that it is better for the protein dynamics study. The SIDEpro Rotamer library is composed by high-definition experimental structures, but they are *static*; as well as the judgment of the Rota_out component, which is focused on static structures Consequently, SIDEpro seems to over-fit the side-chains to the Rotamer library, and Molprobity gets this as positive, because its judgment is based also on a Rotamer library. During some protein conformation switches, it is possible that these Rotamer references do not match the real side-chain positions, as picks of energy, steric contacts or new bonding could take place. In summary, SIDEpro seems to be a nice general supporting feature for the TNM; but it must be taken with care when studying Protein dynamics where side-chains are known to be highly involved.

The Energy Minimization (EM) is usually performed in proteins before molecular dynamics simulations. It ensures that systems are at an energy minimum, preventing steric clashes or inappropriate geometry. The results show a very slight decrease of the Mpscore (by a 0.92 factor), but it is still significant. However, this low number tells us that the local changes have been subtler than in the other conditions. The most shocking result comes when looking at the local parameters. The Rota_out, which had been taken close to 0% average, has been partially reversed after the EM. The Rota_out is 7,22 times higher than the TNM ensemble refined just with SIDEpro. ***See Figure 24***. The reversion is not total, as the Rota_out median of the original TNM ensemble was about 2 and the Rota_out mean 2.4; and after the EM it has a median of 1.2 and a mean of 1.4. So, the effects of the SIDEpro refinement can still be seen after the EM.

This reversion is due to the re-position of some side-chains following a differing criteria than the SIDEpro program. While SIDEpro uses a Rotamer library, the EM process uses the OPLS force field. The discrepancies between the two methods result in a worse Rota_out. However, the final balance is an improvement of the quality of the model.

This partial reversion of the Rota_out when energy-minimizing supports the idea of over-fitting the side-chains to the Rotamer library by SIDEpro. The library is used to reconstruct the side-chains, which are fitted to the library, but poor in energy terms. As Molprobity uses its own Rotamer library to judge the Rota_out percentage, the score rises enormously. None the less, the EM shows us another reality. The side-chains are so well fitted to the SIDEpro library that are energetically unrealistic. So, after the EM, the side-chains are re-located following an energetic criteria. The result of the combination is improved Rotamers which combine the knowledge of a Rotamer library with the realism of a force field. Even if the Rota_out has become lower according to Molprobity, the side-chains after the SP-EM process are probably better and more realistic.

The Rama_iffy parameter is not affected after the EM. The TNM did affect the Rama_iffy, but no influence was seen when using SIDEpro. So, the changes produced in the backbone by the TNM agree with the EM criteria. The EM force field seems to leave the backbone angles intact, and no changes in the Rama_iffy are observed.

The clashscore is reduced by a 0.371 factor, and it explains the change in the Mpscore. The EM puts the conformation in a energy minimum, that is, Van Der Waals, ion interactions and other forces are reduced. As the overlaps mean an increment in the energy due to the repulsion of molecules, these are reduced too. Therefore, the CS is notably decreased, which explains the Mpscore reduction regardless the rise in Rota_out.

The conclusions obtained from this thesis leave the door open to ***future works.*** First, taking the TNM to its limits regarding to the target RMSD score. Both extremes, 0.25 and 2, can be lowered and increased respectively, to see if the system reaches a plateau: in the first case, having exactly the same scores and structures than the reference; and in the second, having the same result over and over even rising the target RMSD. However, this RMSDs would not be realist, and this experiment would be performed to test the behavior of the TNM program. Second, adjusting the target RMSD according to specific protein dynamics simulations. For example, having a chaperon protein in two different conformations (A, when free; B, when attached to another specific protein) experimentally deduced; and calculating from A the target RMSD that creates the most similar ensembles to B. In third place, applying the SIDEpro and EM refinements in Protein Dynamics simulations with the TNM, in order to get the most accurate ensembles. For example, using this pipeline to work with another set of proteins whose alternative conformations are known and see if the SIDEpro and EM procedures help the ensembles to be more similar to the real alternative structures.

The TNM gives an estimated value of the optimal target RMSD according to the temperature (B-factor) and the protein. The study of close ranges to the RMSD proposed would be interesting in order to know the best target RMSD for specific proteins. Also, bringing the temperature factor to the equation would be interesting in order to get more realistic and protein-dynamic results. A study with different temperature values, the RMSD variation and the Molprobity quality assessment would be remarkable. Recently, an option to run TNM positioning the side-chains has been developed. Even if no great consequences were observed when chosen, an approximation taking into account the side-chains since the TNM assembly could allow to improve even more the Rotamers.

7. Bibliography

Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D., & Gerstein, M. Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool. Protein science : a publication of the Protein Society, 2005, 14(3), 633–643.

Allinger, N. L. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. Journal American Chemical Society, 1977, vol. 99, no. 25,

Antunes DA., Devaurs D., Kavraki LE. Understanding the challenges of protein flexibility in drug design. Expert Opin Drug Discovery, 2015 Dec; 10(12):1301-13.

Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophysics, 2001, 80, 505.

Bastolla, Ugo and Dehouck, Yves. Can Conformational Changes of Proteins Be Represented in Torsion Angle Space? A Study with Rescaled Ridge Regression. Journal Chemical Information Modeling. 2019, 59, 11, 4929–4941

Battey,J.N. et al. Automated server predictions in CASP7. Proteins, 2007, 69 (Suppl. 8),

68–82.

Benkert,P. et al. QMEAN: a comprehensive scoring function for model quality

assessment. Proteins. 2008, 71, 261–277

Bhattacharya, D.; Cheng, J., 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. Proteins 2013, 81 (1), 119-31.

Bordoli,L. et al. Protein structure homology modeling using SWISS-MODEL

workspace. Nat. Protocols, 2009 4, 1–13.

Bryngelson J. D. and P. G. Wolynes Spin glasses and the statistical mechanics of protein folding Proc. Natl. Acad. Sci. U.S.A. 84, 7524 (1987).

Cheng,J. et al. SCRATCH: a protein structure and structural feature prediction

server. Nucleic Acids Res, 2005, 33, W72–W76.

Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica, 2010 Jan;66(Pt 1):12-21.

Christopher J. Williams, Jeffrey J. Headd, Nigel W. Moriarty, Michael G. Prisant, Lizbeth L. Videau, Lindsay N. Deis, Vishal Verma, Daniel A. Keedy, Bradley J. Hintze, Vincent B. Chen, Swati Jain, Steven M. Lewis, Bryan W. Arendall 3rd, Jack Snoeyink, Paul D. Adams, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. MolProbity: More and better reference data for improved all-atom structure validation. Protein Science, 2018, 27: 293-315.

Clore,G.M. and Garrett,D.S. R factor, free R, and completecross-validation for dipolar coupling refinement of NMR structures. J. Am. Chem. Soc. 1999, 121, 9008–9012.

Cole, Jason C. and Korb, Oliver and McCabe, Patrick and Read, Murray G. and Taylor, Robin. Knowledge-Based Conformer Generation Using the Cambridge Structural Database. Journal of Chemical Information and Modeling. 2018. 58, 3; 615-629

Dehouck, Y. Mikhailov, A. S.; Effective Harmonic Potentials: Insights into the Internal

Cooperativity and Sequence-Specificity of Protein Dynamics. PLoS Comput. Biol. 2013, 9, No. e1003209

Dehouck, Y. Bastolla, U.; The Maximum Penalty Criterion for Ridge Regression:

Application to the Calibration of the Force Constant in Elastic Network Models. Integr.

Biol. 2017, 9, 627−641.

DePristo MA, de Bakker PI, Blundell TL. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 2004;12:831–838.

Dunbrack RL, Jr. Rotamer libraries in the 21st century. Curr Opin Struct Biol 2002;12:431–440

Eyal E., Yang L.-W., Bahar I. Anisotropic network model: Systematic evaluation and a new web interface. Bioinformatics. 2006;22:2619–2627.

Eramian,D. et al. How well can the accuracy of comparative protein structure

models be predicted? Protein Sci. 2008, 17, 1881–1893.

Eran Eyal, Gengkon Lum, Ivet Bahar, The anisotropic network model web server at 2015 (ANM 2.0), *Bioinformatics*, 2015, 31, 9, 1487–1489

FADLAN, Arif; NUSANTORO, Yesaya Reformyada. The Effect of Energy Minimization on The Molecular Docking of Acetone-Based Oxindole Derivatives. JKPK , [S.l.], v. 6, n. 1, p. 69-77, apr. 2021.

Feig, M.; Mirjalili, V., Protein structure refinement via molecular-dynamics simulations: What works and what does not? Proteins 2016, 84 Suppl 1, 282-92.

Gal Masrati, Meytal Landau, Nir Ben-Tal, Andrei Lupas, Mickey Kosloff, Jan Kosinski, Integrative Structural Biology in the Era of Accurate Structure Prediction, Journal of Molecular Biology, 2021, 167127, ISSN 0022-2836,

Georgii G. Krivov, Maxim V. Shapovalov, Roland L. Dunbrack Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins Volume77, Issue4 December 2009 778-795

Go N, Noguti T, Nishikawa T Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci U S A. 1983 Jun; 80(12):3696-700.

Lim Heo, Collin F. Arbour, and Michael Feig. Improved Sampling Strategies for Protein Model Refinement based on Molecular Dynamics Simulation. *J. Chem. Theory Comput.* 2021, 17, 3, 1931–1943

Heo, L.; Feig, M., High-accuracy protein structures by combining machine-learning with physicsbased refinement. Proteins 2020, 88 (5), 637-642

Hintze, B. J., Lewis, S. M., Richardson, J. S., & Richardson, D. C. Molprobity's ultimate rotamer-library distributions for model validation. *Proteins*, 2016.*84*(9), 1177–1189.

Howe, PW Journal: Principal components analysis of protein structure ensembles calculated using NMR data. Biomol NMR. 2001 May; 20(1):61-70.

Huang,Y.J., Powers,R. and Montelione,G.T. Protein NMRrecall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J. Am. Chem. Soc. 2005, 127, 1665–1674.

Ian W. Davis, Andrew Leaver-Fay, Vincent B. Chen, Jeremy N. Block, Gary J. Kapral, Xueyi Wang, Laura W. Murray, W. Bryan Arendall III, Jack Snoeyink, Jane S. Richardson and David C. Richardson. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research, 2007, 35:, W375-W383.

Jones,T.A., Zou,J.-Y., Cowan,S.W. and Kjeldgaard,M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. 1991, A, 47, 110–119.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 1–7,

Ihaka R. and Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 1996, 5:3, 299-314,

Jorgensen, W. L. * Maxwell D. S., and Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids J. Am. Chem. Soc. 1996, 118,

Karplus M, Brooks B Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci U S A. 1983 Nov; 80(21):6571-5.

Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 2002, 646–652

Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A Chem Coarse-Grained Protein Models and Their Applications. Rev. 2016 Jul 27; 116(14):7898-936.

Kmiecik, S., Kouza, M., Badaczewska-Dawid, A. E., Kloczkowski, A., & Kolinski, A. Modeling of Protein Structural Flexibility and Large-Scale Dynamics: Coarse-Grained Simulations and Elastic Network Models. International journal of molecular sciences, 2018, 19(11), 3496.

Köfinger, J., Stelzl, L. S., Reuter, K., Allande, C., Reichel, K., & Hummer, G. Efficient ensemble refinement by reweighting. *Journal of Chemical Theory and Computation*, 2019, 15(5), 3390– 3401

Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J., Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins, 2019, 87 (12), 1011-1020.

Kuriyan J, Osapay K, Burley SK, Brunger AT, Hendrickson WA, Karplus M. Exploration of disorder in protein structures by X-ray restrained molecular dynamics. Proteins 1991;10:340–358.

Laskowski,R.A., Macarthur,M.W., Moss,D.S. and Thornton, J.M. ProCheck - A program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 1993, 26, 283–291.

J. Levin 1,2, Dmitry A. Kondrashov 1, Gary E. Wesenberg 2, and George N. Phillips Jr. Ensemble refinement of protein crystal structures. Structure. 2007 September ; 15(9): 1040–1052.

Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M Simultaneous

determination of protein structure and dynamics. Nature 2005, 433: 128–132

López-Blanco JR, Chacón P New generation of elastic network models. Curr Opin Struct Biol. 2016 Apr; 37():46-53.

Lovell SC, Davis IW, Arendall WB, III, Bakker PIWD,Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Ca geometry and Cb deviation. Proteins 2003, 50:437–450.

Lu M, B. Poon, and J. Ma, J. Chem. A new method for coarse-grained elastic normal-mode analysis. Theory Comput. 2003, 2, 464

Ma J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure. 2005;13:373–380

Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C., Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011, 6 (12), e28766.

Matsumoto, S., Ishida, S., Araki, M. Extraction of protein dynamics information from cryo-EM maps using deep learning. Nat Mach Intel. 2021, 3, 153–160

McCammon, J.A., Gelin, B.R. and Karplus, M. Dynamics of folded proteins. Nature, 1977, 267.

McGuffin,L.J. The ModFOLD server for the quality assessment of protein

structural models. Bioinformatics, 2008, 24, 586–587.

Meiler J, Baker D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins 2006;65:538–548.

Meńdez,R.; Bastolla,U. Torsional Network Model: Normal Modes in Torsion Angle

Space Better Correlate with Conformation Changes in Proteins. Phys. Rev. Lett. 2010,

104, No. 228103.

Mirjalili, V.; Feig, M., Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. J Chem Theory Comput 2013, 9 (2), 1294-1303.

Mishra S.K., Jernigan R.L. Protein dynamic communities from elastic network models align closely to the communities defined by molecular dynamics. PLoS ONE. 2018;13

Mitchell D. Miller, George N. Phillips, Moving beyond static snapshots: Protein dynamics and the Protein Data Bank. Journal of Biological Chemistry, 2021, 296,

Nagata, K., Randall, A., & Baldi, P. SIDEpro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations. 2012 *Proteins*, *80*(1), 142–153.

S. Omori, S. Fuchigami, M. Ikeguchi, and A. Kidera, J. Linear response theory in dihedral angle space for protein structural change upon ligand binding Comput. Chem. 2009, 30, 2602

Orellana L., Yoluk O., Carrillo O., Orozco M., Lindahl E. Prediction and validation of protein intermediate states from structurally rich ensembles and coarse-grained simulations. Nat. Commun. 2016;7:12575.

Orozco M A theoretical view of protein dynamics. Chem Soc Rev. 2014 Jul 21; 43(14):5051-66.

Pan AC, Weinreich TM, Piana S, Shaw DEJ Demonstrating an Order-of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems. Chem Theory Comput. 2016 Mar 8; 12(3):1360-7.

P**anwar A., Ashok K. In-silico Analysis and Molecular Dynamics Simulations of Lysozyme by GROMACS 2020.2.** *Annals of the Romanian Society for Cell Biology*, 2021, *25*(6), 9679–9685.

Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci 2004;13:735–751.

Poma A.B., Li M.S., Theodorakis P.E. Generalization of the elastic network model for the study of large conformational changes in biomolecules. Phys. Chem. Chem. Phys. 2018;20:17020–17028.

Porter, J. R., Zimmerman, M. I. & Bowman, G. R. Enspara: modeling molecular ensembles with scalable data structures and parallel computing. *J. Chem. Phys.* 2019, 150, 044108

Rahman, A., Correlations in the motion of atoms in liquid argon. Physical Review, 1964 136(2A), p.A405

RAMACHANDRAN GN, RAMAKRISHNAN C, SASISEKHARAN V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963 Jul;7:95-9.

Ravikumar, A., de Brevern, A. G., & Srinivasan, N. Conformational Strain Indicated by Ramachandran Angles for the Protein Backbone Is Only Weakly Related to the Flexibility. *The Journal of Physical Chemistry B*, 2021, *125*(10), 2597-2606.

Read, R. J.; Sammito, M. D.; Kryshtafovych, A.; Croll, T. I., Evaluation of model refinement in CASP13. Proteins 2019, 87 (12), 1249-1262. **Rice L. M. , Axel T. BrüNger. Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement.** Proteins:1994, 19, 277-290

Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004;55:656–677.

K. Roy, S. Kar, & R. N. Das,"Computationalchemistry," in Understanding The Basics of QSAR for Applications in Pharmaceutical Sciences Risk Assessment, K. Roy, S. Kar and R. N. Das, Elsevier, SanDiego, 2015b, pp. 357-425.

Schwede,T. et al. Outcome of a workshop on applications of protein models in

biomedical research. Structure, 2009, 17, 151–159.

Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D., Improved protein structure prediction using potentials from deep learning. Nature 2020, 577 (7792), 706-710

Sippl,M.J. Recognition of errors in three-dimensional structures of proteins.

Proteins, 1993, 17, 355–362

Stillinger, F.H. and Rahman, A. Improved simulation of liquid water by molecular dynamics. The Journal of Chemical Physics, 1974, 60(4), pp.1545-1557

Stumpff-Kane AW, Maksimiak K, Lee MS, Feig M Sampling of near-native protein conformations during protein structure refinement using a coarse-grained model, normal modes, and molecular dynamics simulations. Proteins. 2008 Mar; 70(4):1345-56.

Tirion M.M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. 1996;77:1905–1908.

Vakser IA. Curr Opin. Challenges in protein docking. Struct Biol. 2020, 64, 160-165.

Veenstra DL, Kollman PA. Modeling protein stability: a theoretical analysis of the stability of T4 lysozyme mutants. Protein Eng 1997; 10:789–807.

Vincent B. Chen, W. Bryan Arendall III, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson and David C. Richardson. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica, 2010, D66: 12-21.

Vriend,G. WHAT IF: A molecular modeling and drug design program. J. Mol. Graph. 1990, 8, 52–56.

Wang,Z. et al. Evaluating the absolute quality of a single protein model using

structural features and support vector machines. Proteins, 2009, 75, 638–647.

Yang L., Song G., Carriquiry A., Jernigan R.L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. Structure. 2008;16:321–330.

Xiang, Zhexin; Steinbach, Peter J.; Jacobson, Matthew P.; Friesner, Richard A.; Honig, Barry. Prediction of side-chain conformations on protein surfaces. Proteins: Structure, Function, and Bioinformatics. 2007; 66(4):814–823.

Zhang, Y., Protein structure prediction: when is it useful? Curr. Opin Struct Biol 2009, 19 (2), 145

Annex

GROMACS script:

```
#!/bin/bash
### 11/07/2021 ###

## Created by David Roncero ##

### This script allows to perform the Energy Minimization of a Protein, when the input pdb
### is given for the first argument, and the output pdb file is given as second argument

source /usr/local/gromacs/bin/GMXRC

grep -v HOH $1 > structure_clean.pdb

gmx pdb2gmx -f structure_clean.pdb -o structure_processed.pdb -water spce

gmx editconf -f structure_processed.pdb -o structure_newbox.pdb -c -d 1.0 -bt cubic

gmx solvate -cp structure_newbox.pdb -cs spc216.gro -o structure_solv.pdb -p topol.top

gmx grompp -f ions.mdp -c structure_solv.pdb -p topol.top -o ions.tpr

gmx genion -s ions.tpr -o structure_solv_ions.pdb -p topol.top -pname NA -nname CL -neutral

gmx grompp -f minim.mdp -c structure_solv_ions.pdb -p topol.top -o em.tpr

gmx mdrun -v -deffnm em -c $2
```