Geometric deep manifold learning combined with natural language processing for protein movies

Valentin Lombard

SORBONNE UNIVERSITÉ

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

École doctorale d'informatique, télécommunications et électronique de Paris

Thèse de doctorat de Sorbonne Université

Geometric deep manifold learning combined with natural language processing for protein movies

Soutenue publiquement le 16/12/2024 par :

Valentin LOMBARD

Devant le jury composé de :

Slavica JONIC	Présidente du jury
Florence TAMA	Rapporteuse
Jean-Christophe GELLY	Rapporteur
Marco PASI	Examinateur
Sergei GRUDININ	Directeur de thèse
Élodie LAINE	Directrice de thèse

"Everyday above ground is a great day."

– ***Mr. Worldwide***

Remerciements

Cette thèse est le fruit de trois années passées au sein du LCQB, le Laboratoire de Biologie Computationnelle et Quantitative de l'Université de la Sorbonne. Je tiens à remercier toutes les personnes qui, directement ou indirectement, ont contribué à sa réalisation.

En premier lieu, je souhaite exprimer toute ma gratitude à Élodie Laine, ma directrice de thèse, pour m'avoir accueilli dans son équipe, éclairé sans relâche de ses conseils et de sa vision scientifique, tout en m'apportant son soutien et sa bonne humeur contagieuse. Je remercie également Sergei Grudinin, mon co-directeur de thèse, pour son aide précieuse, son soutien sans faille, les échanges enrichissants que nous avons eus, ainsi que pour son accueil chaleureux lors de ma visite à Grenoble.

Je tiens également à remercier Florence Tama et Jean-Christophe Gelly, les rapporteurs de cette thèse, pour leurs lectures attentives et leurs retours constructifs, ainsi que Slavica Jonic et Marco Pasi, membres du jury, pour avoir accepté d'évaluer mon travail.

Je souhaite également exprimer ma gratitude envers Juliana Bernardes et Pablo Chacón, qui ont accepté de participer aux comités de suivi de thèse et m'ont apporté leur soutien et leurs précieux conseils.

Maintenant, permettez-moi (s'il vous plaît !) de ne pas me lancer dans l'ambitieuse tâche d'établir une liste exhaustive de toutes les personnes avec qui j'ai partagé de bons moments au cours de ces années. Je garde des souvenirs précieux de tant d'échanges et de moments partagés, et quelques mots ne suffiraient pas à rendre justice à l'importance qu'ils ont pour moi.

J'ai toutefois une pensée spéciale pour les doctorants qui ont réussi à partager mon bureau sans devenir fous : Marina, Louis et Julien. Vous allez pouvoir travailler dans le calme maintenant.

Je pense bien sûr à mon père, ma mère, mon frère, mon oncle, mon grand-père, ma grand-mère, ainsi qu'à tous les autres membres de ma famille, qui m'ont

toujours soutenu et encouragé, chacun à sa manière.

Enfin, je tiens à remercier Naomi pour sa patience, son soutien, sa présence, et pour tout le reste.

Abstract

Proteins play a central role in biological processes, and understanding how they deform and move is essential to elucidating their functional mechanisms. Despite recent advances in high-throughput technologies, which have broadened our knowledge of protein structures, accurate prediction of their various conformational states and motions remains a major challenge. This thesis presents two complementary approaches to address the challenge of understanding and predicting the full range of protein conformational variability.

The first approach, Dimensionality Analysis for protein Conformational Exploration (DANCE) for a systematic and comprehensive description of protein families' conformational variability. DANCE accommodates both experimental and predicted structures. It is suitable for analyzing anything from single proteins to superfamilies. Employing it, we clustered all experimentally resolved protein structures available in the Protein Data Bank into conformational collections and characterized them as sets of linear motions. The resource facilitates access and exploitation of the multiple states adopted by a protein and its homologs. Beyond descriptive analysis, we assessed classical dimensionality reduction techniques to sample unseen states on a representative benchmark. This work improves our understanding of how proteins deform to perform their functions and opens ways for a standardized evaluation of methods designed to sample and generate protein conformations.

The second approach relies on deep learning to predict continuous representations of protein motion directly from sequences, without the need for structural data. This model, SeaMoon, uses protein language model (pLM) embeddings as inputs to a lightweight convolutional neural network with around 1 million trainable parameters. SeaMoon achieves a success rate of 40% when evaluated against around 1,000 collections of experimental conformations, capturing movements beyond the reach of traditional methods such as normal mode analysis, which relies solely on 3D geometry. In addition, SeaMoon generalizes to proteins that have no detectable

sequence similarity with its training set and can be easily retrained with updated pLMs.

These two approaches offer a unified framework for advancing our understanding of protein dynamics. DANCE provides a detailed exploration of protein movements based on structural data, while SeaMoon demonstrates the potential of sequence-based deep learning models to capture complex movements without relying on explicit structural information. Together, they pave the way for a more comprehensive understanding of protein conformational variability and its role in biological function.

Outline

This thesis manuscript addresses deep learning methods applied to the study of protein dynamics. The first two chapters of this manuscript are introductory. The first chapter aims to introduce biological concepts, while the second chapter introduces deep learning techniques useful for understanding the work discussed in the following chapters. Chapter 3 presents a method to extract a family-specific linear motion from a set of sparse observations. This method is applied to all known experimental structures to form a database of linear motions. Chapter 4 presents a deep learning method aimed at predicting the linear motion of a given protein based on its sequence embedding from a protein language model. This method is trained on a database created using the method described in Chapter 3. Chapter 5 concludes this manuscript and offers perspectives on possible improvements to the method described in Chapter 4.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to proteins, their 3D structures, and dynamics

The aim of this chapter is to introduce the essential concepts, outside of deep learning, that are necessary for understanding the content of this thesis. Section 1.1 aims to provide an introduction to proteins as biological entities. Since this thesis is entirely based on data derived from protein structures, it seemed important to describe how these data are obtained. Thus, section 1.2 offers an introduction to the main experimental techniques used to obtain the three-dimensional structure of proteins. Section 1.3 discusses the importance of the dynamic nature of proteins for their function, while section 1.4 explores the determination of these dynamics through two distinct methods based on physical principles.

1.1 Proteins

Proteins are macromolecules present in all living organisms. Essential to life, they participate in a wide variety of biological functions. A short and incomplete list of their biological roles might include the following: 1. A structural role in forming the support for the different structures of organisms, such as collagen in skin and bones, or keratin in hair and nails. 2. An enzymatic role in catalyzing essential chemical reactions, such as the degradation of starch by amylase. 3. A transport role, such as hemoglobin in the blood which brings oxygen to tissues. 4. A role in immunity, such as antibodies which serve to identify and neutralize pathogenic viruses and bacteria. 5. A role in movements, such as actin and myosin which participate in muscle contraction. 6. A storage role, such as ferritin which allow

the storage of iron.

Protein synthesis Proteins are formed in cells during a process called protein synthesis. This process can be divided into two phases called transcription and translation (Fig. 1.1).

During the transcription phase, some section of the deoxyribonucleic acid (DNA), known as a gene, is transcribed into RNA. The DNA and RNA are made of four nitrogenous bases, adenine, cytosine, thymine and guanine in DNA, and adenine, cytosine, guanine, uracil in RNA. The bases pair in a complementary manner, in DNA the adenine pairs with thymine and the cytosine pairs with guanine. This pairing gives DNA its double-stranded structure, forming a double helix of two antiparallel strands. The arrangement of these base pairs allows DNA to contain all the genetic information, called the genome, of the living being. The RNA, on the other hand, is formed of a single strand that is synthesized by the RNA polymerase, which moves along the opened DNA that serves as template. In eukaryotic organisms, this process takes place in the nucleus. Once synthesized, the initial RNA transcript, known as precursor messenger RNA (pre-mRNA) undergoes post-transcriptional modifications, to form a mature mRNA that is transported outside the nucleus for the translation phase of the protein synthesis.

The translation phase occurs in the cytoplasm. Here, the mRNA encounters a ribosome, a complex machinery formed with ribosomal RNA (rRNA) and proteins, that read its sequence to translate it into protein. The ribosome reads the mRNA three bases at a time. These sets of three nucleotides are called codons. Each one of the codons codes for one of the 20 standard amino acids that compose protein, or signal the end of the translation. The correspondence between the codons and the amino acid they encode is known as the genetic code (Fig. 1.2). It is noteworthy that this code contains redundancy, as there are $4^3 = 64$ possible codons, but they only result in 20 standard amino acids and the stop codon. Once a codon is read, the ribosome chooses the corresponding transfer RNA (tRNA) molecule that carries the specific amino acid. This amino acid is attached to the previous ones with peptide bond, forming what is called the peptide chain. This process occurs fast, at around 5 to 20 amino acids translated per second, and with high accuracy with an error rate estimated at one error every $\sim 10^4$ codon translated. The transcription occurs at similar speed, but with typically one order of magnitude lower error rate. The polypeptide chain then quickly folds itself into a specific shape that will determine its function within the organism.

Amino acids Amino acids are the building blocks of proteins and share a common structure known as the backbone, which forms the principal chain. This

backbone is composed of a central carbon atom (C_α) linked to an amine group (-NH_2), a carboxyl group (-COOH), and a hydrogen atom. In addition, amino acids have a variable side chain (-R) attached to the C_α. This variable chain, known as the side chain or the secondary chain, differentiates each amino acid and determines its unique chemical and physical properties, influencing the final protein structure and function. These side chains impart different chemical characteristics to amino acids, such as hydrophilicity or hydrophobicity. Hydrophobic amino acids are non-polar and uncharged, and tend to be found inside the protein to minimize their contact with water, whereas hydrophilic amino acids have polar or charged side chains that interact favorably with water, usually by forming hydrogen bonds, and tend to be found on the protein surface. Other type of amino acids, known as amphipathic, have both a polar and non-polar character, and therefore tend to be at the interface between hydrophobic and hydrophilic environnement (Fig. 1.2).

There are four levels of protein structure (Fig. 1.3).

Primary structure The primary structure corresponds to the linear succession of amino acids without spatial reference, it is often noted in the form of a sequence formed from an alphabet of 20 letters, each corresponding to one of the 20 standard amino acids. By convention, the primary structure has a direction from the amino acid with the free amine end, called the N-terminus, to the amino acid with the free carboxylate end, called the C-terminus.

Secondary structure The secondary structure of proteins corresponds to the local organization of the polypeptide chain into regular motifs, such as the α-helix and the β-sheet, stabilized by hydrogen bonds. The α-helix, first described by L. Pauling in 1951 [1], is the most common secondary structure, as well as the most predictable from analysis of the primary structure. It consists of a right-handed helix-shaped chain, where each N-H group of an amino acid forms a hydrogen bond with the C=O group of the main chain located four amino acids upstream. The β-sheet is another common secondary structure in proteins, characterized by the alignment of several parallel or antiparallel polypeptide chains, forming a sheet-like arrangement. It is also stabilized by hydrogen bonds between the N-H and C=O groups of different residues. The main chain of a protein contains three covalent bonds per amino acid. Since the peptide bond is highly constrained because of the partial double bond between carbon and nitrogen atom, this leaves two single bonds around which rotation is possible, making it possible to describe the conformation of an amino acid backbone from two dihedral angles, that are called ϕ and ψ. The dihedral angle ϕ is defined by four successive atoms of the backbone: CO-NH-C_α-CO, with the first carbonyl group belonging to the

preceding residue. The dihedral angle ψ, on the other hand, is defined by the four atoms: NH-C_α-CO-NH, with the second amide being that of the following residue (Fig. 1.4a). Not all values of the angles ϕ and ψ are achievable, as some lead to energetically unfavorable configuration because of electrostatic destabilization and steric hindrance. The Ramachandran diagram, introduced in 1963 by G.N. Ramachandran [2], graphically represents admissible combinations of angles ϕ and ψ, showing three main energetically favorable zones. Analyzing the structure of a protein, we see that the majority of amino acids have combinations of angles (ϕ, ψ) that fall within these zones, corresponding mainly to the two common secondary structure, α-helices and β-sheets (Fig. 1.4b).

Tertiary structure The secondary structure elements fold into a compact object stabilized by weak interactions involving polar and nonpolar groups, resulting in the tertiary structure. The tertiary structure is the protein's three-dimensional structure. Most of the time, the structure of a protein is entirely determined by its amino acid sequence, or primary structure. This property is known as Anfinsen's dogma [3]. Sometimes, proteins known as chaperones can assist protein folding, but in the majority of cases, a protein can be unfolded and will spontaneously refold into what is known as its native state, demonstrating the validity of Anfinsen's dogma.

Quaternary structure The quaternary structure is the highest level of organization of proteins, involving two or more polypeptides. They can be composed of several identical chains, forming what is called a homomer, or by different polypeptides chains, forming a heteromer. The primary stabilizing factor of quaternary structures is the hydrophobic interactions between the non-polar amino acids, the hydrophobic regions of the monomers come together to minimize the solvent exposure.

Representation In textbooks, scientific articles and in this manuscript, protein structures are represented in several ways. For the visualisation and production of these representations from atomic coordinates, the software Pymol will be used in this manuscript [4]. Different examples of representations are given in Figure 1.5.

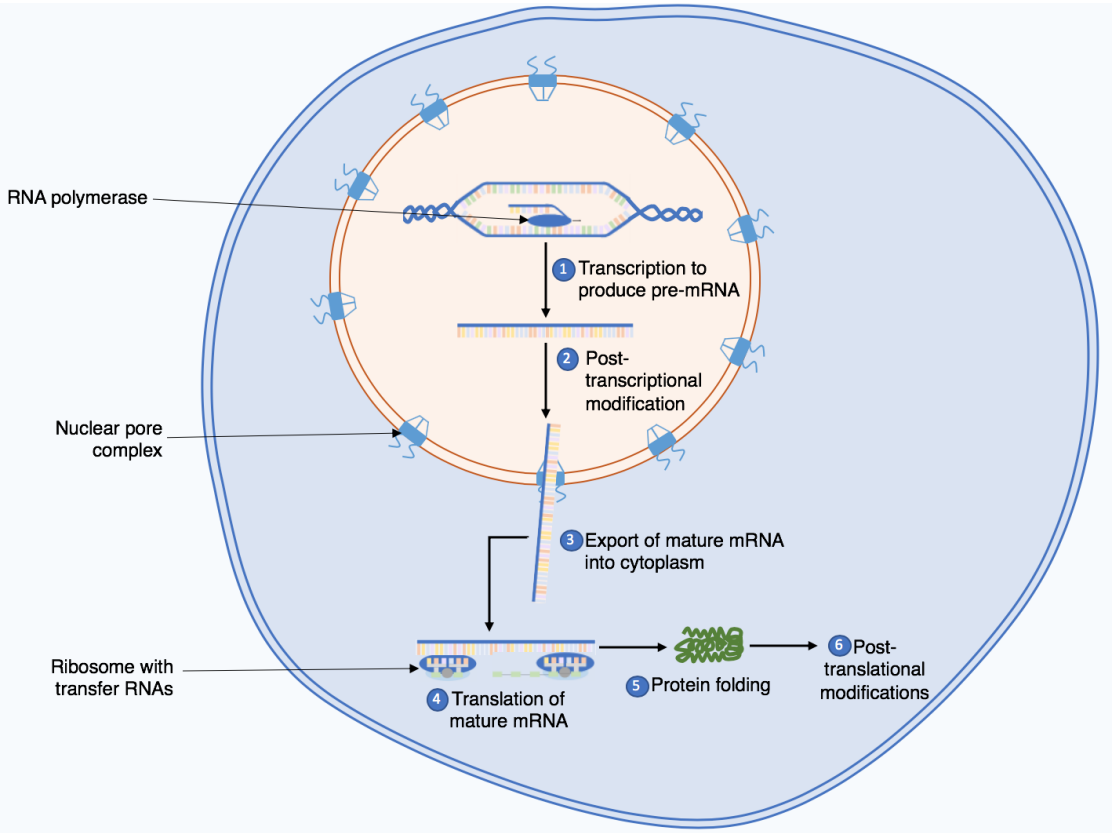


Figure 1.1: **Steps of the protein biosynthesis in eukaryote organism.** 1. RNA polymerase synthesizes pre-messenger RNA from the DNA in the nucleus. 2. The pre-mRNA undergoes post-transcriptional modifications to become mature mRNA. 3. The mRNA is exported from the nucleus to the cytoplasm through nuclear pores. 4. In the cytoplasm, ribosomes attach to the mRNA and use tRNAs to assemble amino acids according to the order of the codons, synthesizing a polypeptide chain. 5. The polypeptide chain folds into its three-dimensional structure. 6. The protein may undergo various additional modifications after its synthesis to become fully functional. *Kep17, CC BY-SA 4.0.*

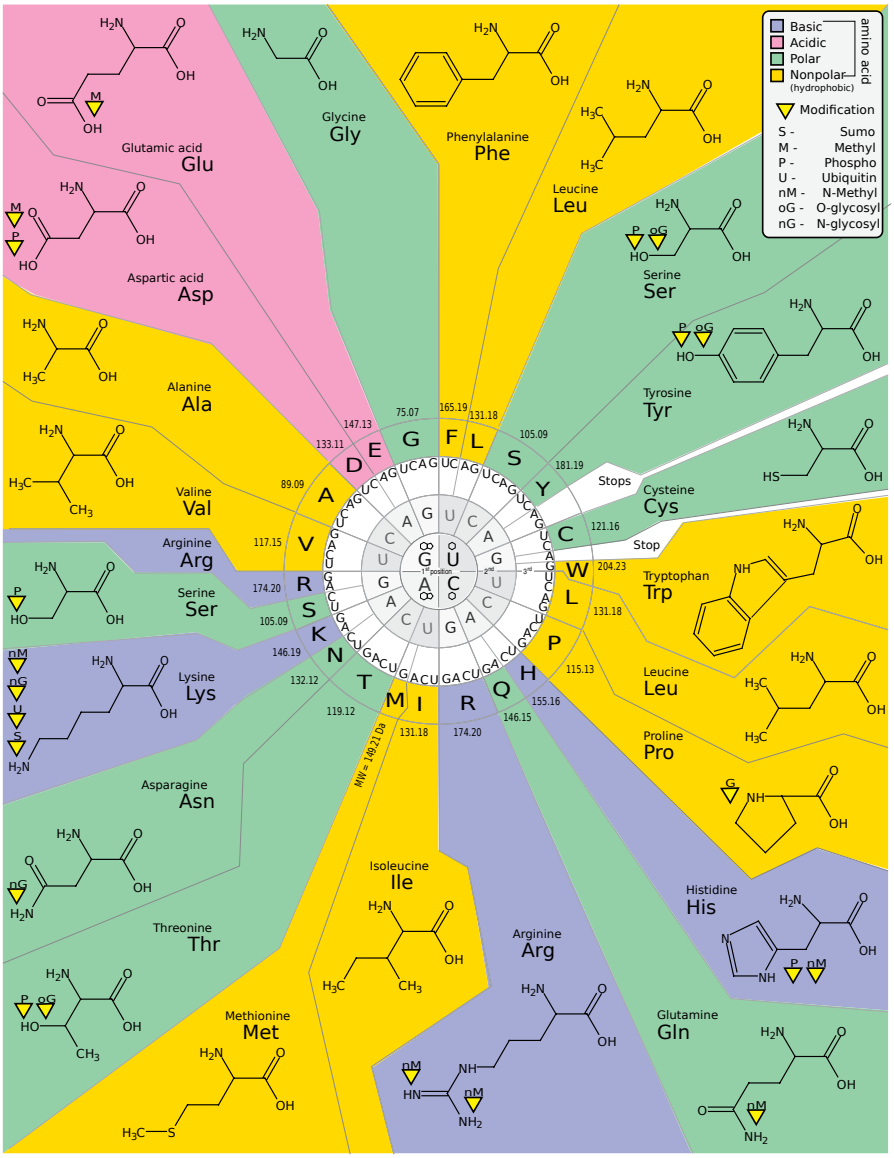


Figure 1.2: **The genetic code and properties of the 20 base amino acids.** The correspondence between the codon and the specific amino acid is given by reading the table from the center to the extremities. The molecular weight of each amino acid is given near the outer circle, in g/mol. The color of each amino acid's area corresponds to its physicochemical properties: blue for basic amino acids, pink for acidic amino acids, green for polar amino acids, and yellow for nonpolar amino acids. The yellow triangles represent potential modifications that the amino acids can undergo. *Edited by Seth Miller User:arapacana, Original file designed and produced by: Kosi Gramatikoff User:Kosigrim, courtesy of Abgent, also available in print (commercial offset one-page: original version of the image) by Abgent, Public domain*

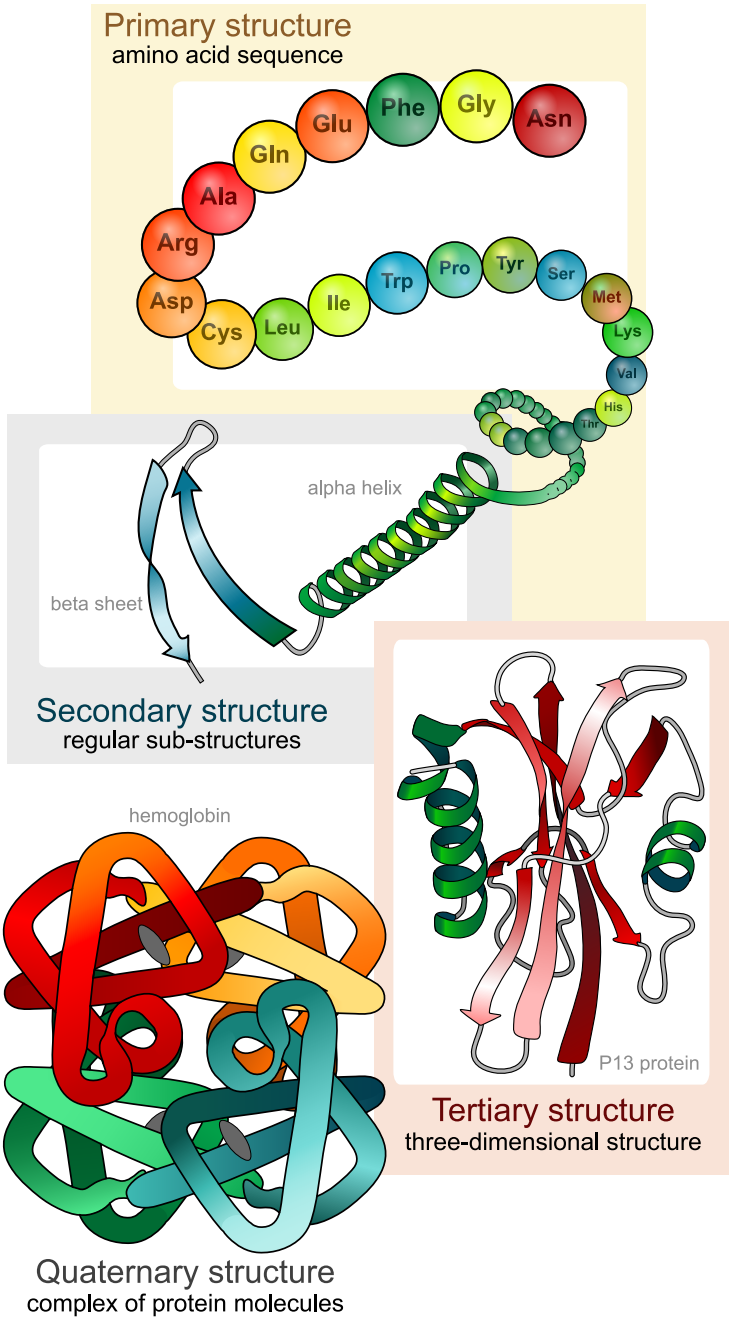
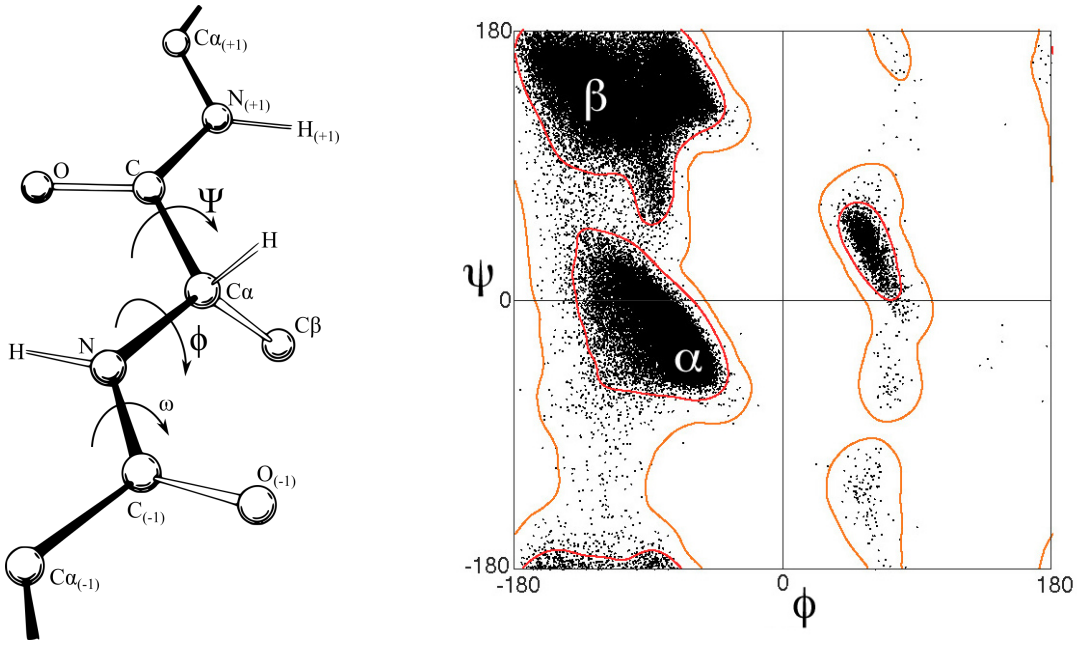


Figure 1.3: **The four levels of protein structure.** The one-dimensional amino acid sequence (primary structure) forms local motifs such as alpha helices and beta sheets (secondary structure). The entire chain adopts a global arrangement (tertiary structure), which can then associate with other chains to form complexes (quaternary structure) *LadyofHats, Public domain.*



(a) (b)

Figure 1.4: (a). **The dihedral angles illustrated on a protein backbone** The angles ϕ, ψ, and ω are shown, corresponding to the rotation around the N-C, C_α-C, and C-N bonds, respectively. *Dcrjsr, vectorised Adam Redzikowski, CC BY 3.0.*, (b). **Ramachandran plot of several proteins** The orange areas delineate the regions favorable for conformational stability. Two main regions are observed, corresponding to the secondary structures of alpha helices and beta sheets. The small region where $\phi > 0$ corresponds to a left-handed helical conformation. Glycines, which do not contain a side chain, are less restricted and can sometimes be found outside of the favorable regions. *Dcrjsr, CC BY 3.0.*

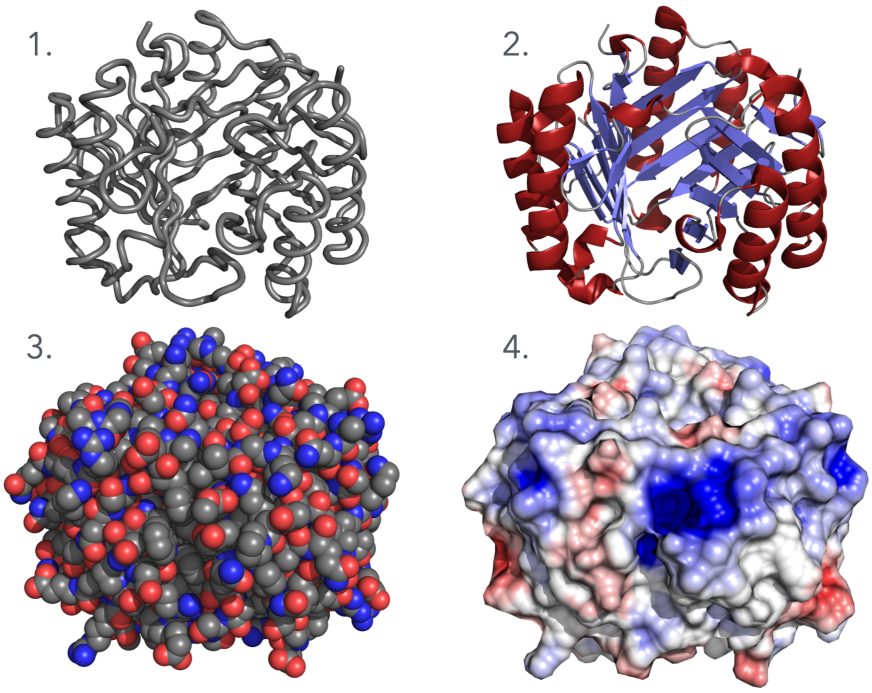


Figure 1.5: **Different ways of representing protein structures.** Four representations of the same macrophage migration inhibitory factor protein (Protein Data Bank code: 1ljt). 1. Wire representation showing the path of the protein backbone. 2. Ribbon representation highlighting the secondary structure elements. Alpha helices (red) are represented by a coiled ribbon, and beta strands (blue) are represented by arrows pointing toward the C-terminal extremity. 3. Sphere representation, where each non-hydrogen atom is shown as sphere the size of its van der Waals radius. The atoms are colored according to the element type. 4. Surface representation colored by electrostatic potential. The red areas are negatively charged, and the blue areas are positively charged.

1.2 Experimental determination of protein structures

1.2.1 The Protein Data Bank

History The Protein Data Bank [5] is a fundamental resource in structural biology. It is an international effort to regroup the three-dimensional structures of biological molecules such as proteins, nucleic acids, and complexes. It was first launched in 1971 at Brookhaven National Laboratory and originally contained 7 structures, that were shared on request on magnetic tape [6, 7]. Since then, it has seen a significant growth, crossing the 1,000 entries milestone in 1994, the 10,000 entries in 1999, the 50,000 entries in 2008 and the 100,000 entries in 2014 (Fig. 1.6). As of the time of writing, the PDB contains 223,790 entries. The PDB was based in Brookhaven until 1998. In 1999 the management of the PDB was transferred to a U.S. consortium named Research Collaboratory of Structural Bioinformatics (RCSB PSB). In 2003, an international collaboration agreement management of the PDB was established, known as the worldwide Protein Data Bank (wwPDB) [8, 9], gathering the U.S. pole (RCSB PDB), a European pole (PDBe), and a Japanese pole (PDBj).

Impact The PDB plays a considerable role in structural biology. In 2014, Nature journal published a list of the 100 most-cited research papers of all time, where the commonly cited PDB paper ranked 92nd [10]. A study made in 2017 placed it 5th in the list of the most cited papers since the year 2000 [11]. In 2023 alone, the wwPDB reported that over 3 billion structures were downloaded from them. In particular, open access to structures and active sites has contributed to the development of almost all 210 new drugs approved by the U.S. Food and Drug Administration (FDA) over the 2010-2016 period, and all new drugs approved over the 2019-2023 period [12, 13].

Entries The PDB entries come from three main methods for determining experimental structures, ranked by number of entries: the X-ray crystallography with 83.6% of the PDB entries, the Elecron Microscopy (EM) with 9.81%, and the Nuclear Magnetic Resonance (NMR) with 6.40%.

1.2.2 X-Ray crystallography

History X-ray crystallography is historically the most widely used method for determining the atomic structure of proteins. Its principle is based on the analysis

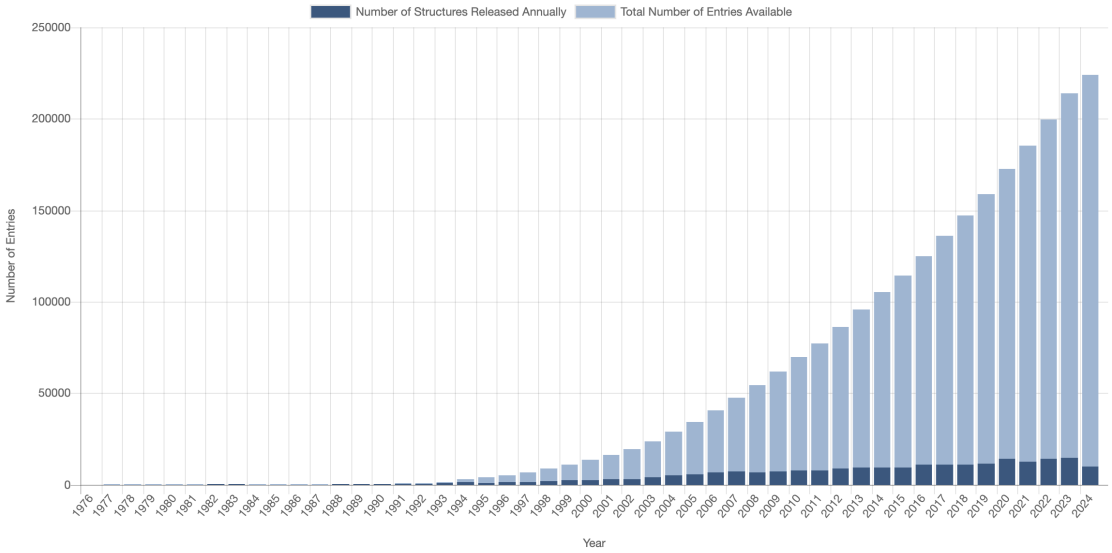


Figure 1.6: **Growth of the number of entries in the Protein Data Bank among the years.**
Source: `https://www.rcsb.org/stats/growth/growth-released-structures`.

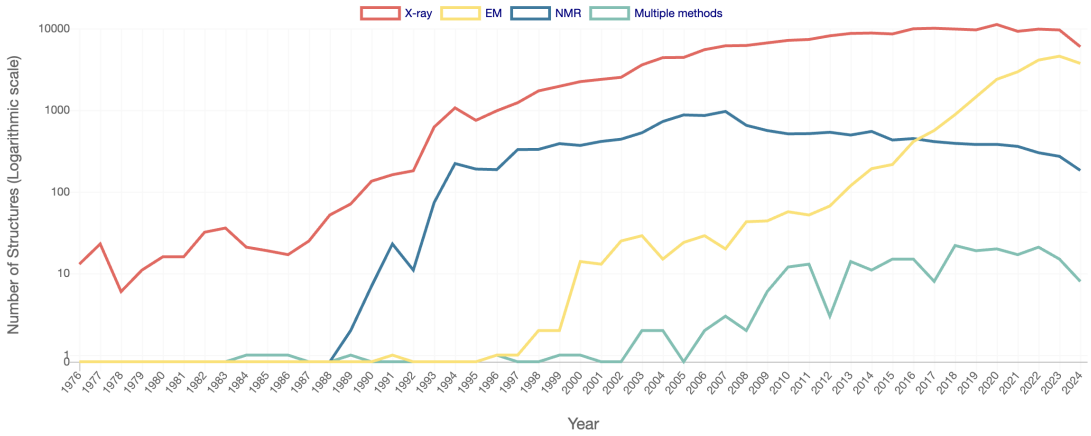


Figure 1.7: **Number of released PDB structures per year and per method.**
Source: `https://www.rcsb.org/stats/all-released-structures`

of diffraction patterns during the scattering of X-rays by a crystal. When a beam of X-rays passes through a crystal, which is a solid in which the atoms are regularly organized in all directions in space, it is diffracted at specific angles, generating a pattern that can be analyzed to reconstruct the crystal's structure. The origins of this method can be traced back to the discoveries of William Henry Bragg and his son William Lawrence Bragg, who in 1913 discovered the law linking the diffraction angle, the X-ray wavelength and the distance between the atomic planes of the crystal, thus enabling the position of atoms in the crystal to be calculated [14]. This discovery laid the foundations of modern crystallography and earned Braggs the Nobel Prize in Physics in 1915. To this day, William Lawrence Bragg remains the youngest winner of a scientific Nobel Prize, at the age of 25. The use of X-ray crystallography to study biomolecules took off in the mid-twentieth century. Major contributions to this field were made by Dorothy Crowfoot Hodgkin, who used X-ray crystallography to determine the structure of penicillin in 1945, and vitamin B12 in 1956 [15, 16]. For this latter contribution, Hodgkin was awarded the Nobel Prize in Chemistry in 1964, becoming the first British woman to win a Nobel Prize. It was only after this work that X-ray crystallography began to be applied to proteins. John Kendrew and Max Perutz, were the first to successfully determine the three-dimensional structure of a protein, the myoglobin, in 1958 [17]. For this work, they were awarded the Nobel Prize in Chemistry in 1962.

Experimental context The main difficulty with this method is that it is entirely dependent on the crystallization step. In order to get the protein structure with high resolution, the protein crystal needs to be large enough, typically more than 0.1 mm in its longest dimension - even though advances in microcrystallography and femtosecond crystallography and have enabled the analysis of much smaller crystals [18, 19] - , pure in its composition and with no internal imperfections [20]. During the crystallization process, proteins are dissolved in an aqueous environment until they reach a state of supersaturation. At this point, and if experimental conditions are favorable, the proteins assemble to form crystals. Favorable conditions for protein crystallization can be more or less complicated to find, and can vary greatly depending on the protein being studied. As an example, membrane proteins are often known to be difficult to crystallize due to the presence of hydrophobic segments crossing the lipid membrane, which tend to be denatured on exposure to water solvent. Proteins with high conformational flexibility are also known to be often hard to crystallize. After collecting diffraction data, crystallographers have to solve another challenge known as the phase problem. This is because sensors are only able to capture the resulting intensity of the

diffracted rays, and the phase information, which is the shift between the incident and diffracted wave and is essential for reconstructing the crystal structure, is lost. To solve this problem, crystallographers can use direct methods if the resolution is high enough, molecular replacement (MR) which exploits the already known phases of a similar molecule, and multiple isomorphic replacement (MIR) which consists of introducing heavy atoms into the crystal that modify the diffraction intensities and allow the missing phases to be calculated by comparison with the original diffraction patterns. After capturing the diffraction data and finding a phase, crystallographers construct an electron density map, which represents an envelope of electron presence in the crystal. The quality of this electron map is measured by its resolution. This resolution is expressed in terms of distance: if a structure has been determined at a resolution of 2 Å, then two atoms separated by more than 2 Å will appear as separate maxima on the electron density map. Resolution depends directly on the fineness of the diffraction pattern, which in turn depends directly on the quality of the diffraction experiment and the protein crystal. At a resolution of between 0 and 1 Å, i.e. a distance less than the typical length of a covalent bond, it is easy to distinguish and position each atom in the structure. At a resolution of 3 Å or more, the envelope only describes the basic contours of the protein chain, and atom positions can be inferred with limited accuracy (Fig. 1.9). The structure obtained in an X-ray crystallography experiment is considered to be a static structure averaged over time and space. It is an average of time, because it takes time to capture the data, and it is an average of space because the position obtained is an average of the atomic positions over the different meshes of the crystal. However, the X-ray experiment allows for a certain degree of measurement flexibility.

B-factor The B-factor, or Atomic Displacement Parameter (ADP), measures the attenuation of X-ray diffraction caused by thermal agitation of the atoms. It is interpreted as describing the amplitude of fluctuation of an atom around its mean position. The higher an atom's B-factor, the greater its fluctuation amplitude. They have been used to study protein dynamics for a long time [21, 22, 23]. Nevertheless, the B-factor is known to have certain limitations [24]. As explained above, the B-factor is supposed to represent the amplitude of fluctuation around an equilibrium position. But due to potential conformational variation, this equilibrium position may not be unique and some atoms may have two or more stable positions. When the resolution of the crystallography experiment is not high enough, it is likely that the inferred position of the atom is an average of the stable positions and that the fluctuations around this position are therefore

overestimated. Thus, the B-factor does not clearly distinguish between disorder arising from thermal motion and disorder arising from structural variability, which limits its interpretation. However, it may be possible to distinguish between the two by cooling the crystal, which supposedly only reduces thermal fluctuations, but this approach also has its limitations [21]. In addition, like structural resolution, they are also affected by potential crystal defects and by potential damage to the crystal caused by X-rays during data acquisition [25, 26]. Furthermore, B-factors are extremely dependent on the resolution of the experiment. For all these reasons, B-factors can be particularly difficult to compare between two proteins. Another limitation if the resolution of the X-ray experiment is not high enough, the B-factor will be isotropic, i.e. it assumes that the amplitude of fluctuation is independent of the direction of space. This is an important restriction because, in reality, it is likely that there are preferential directions for the movement of atoms.

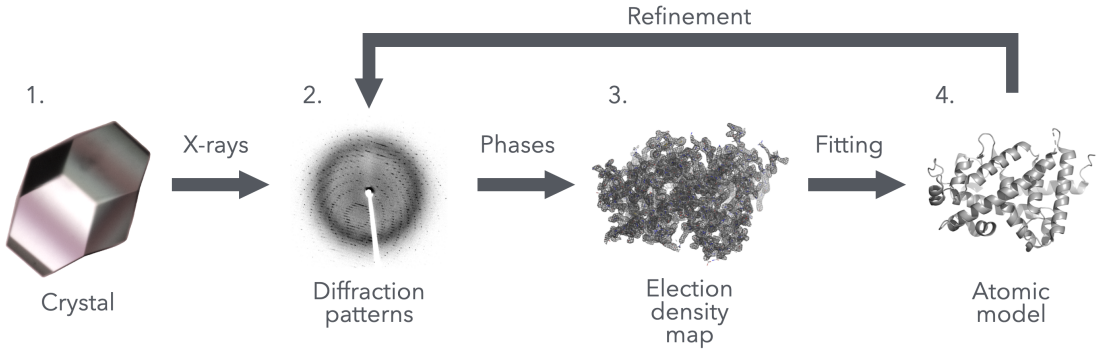


Figure 1.8: **Workflow of structure determination by X-ray crystallography.** 1. The protein is purified and crystallized. These crystals are irradiated with X-rays, often generated by a synchrotron. 2. The X-rays scatter on the crystal lattice planes, producing a diffraction pattern that is captured by a detector. The phases are calculated. 3. Using the diffraction pattern and the phases, electron density maps are generated. 4. These maps enable crystallographers to construct an initial model of the protein's structure. The model is refined by comparing the calculated diffraction pattern of the model with the actual pattern observed in the crystal. Through iterative adjustments, the model is optimized until the calculated and observed patterns match as closely as possible. The quality of the final structure is assessed by measuring the percentage difference between the calculated and actual diffraction patterns, often known as the R-factor.

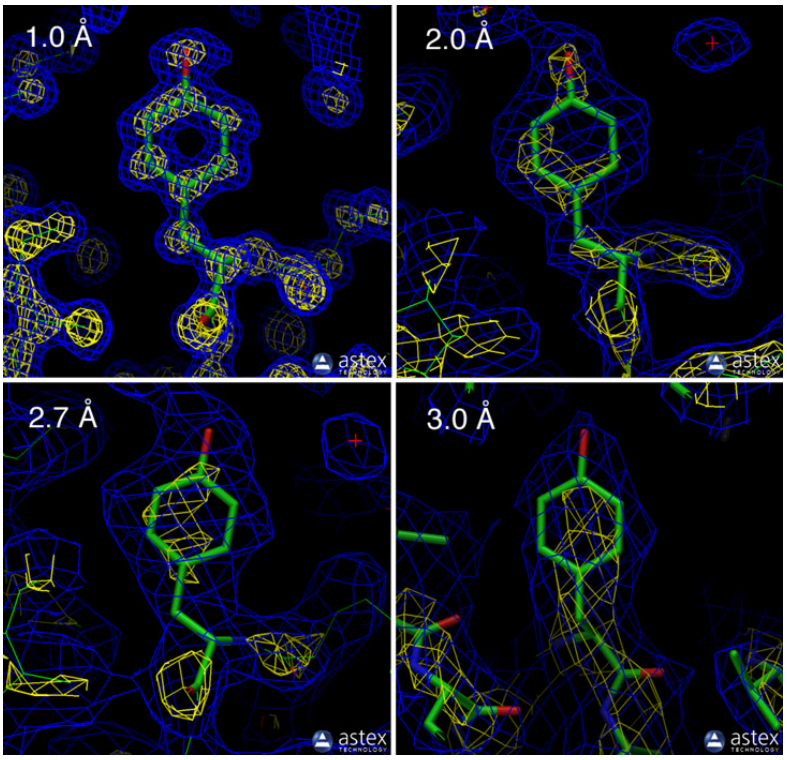


Figure 1.9: **Electron density maps for structures with different resolutions.** At 1 Å resolution, the atoms are visible and resolved. The interatomic distances can be measured to a few hundredths of an Ångstrom. At 3 Å the interatomic distances can only be measured to about ± 0.5 Å. Blue and yellow represents region with high and higher electron density.

Source:
`https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/resolution`.

1.2.3 NMR

History Nuclear magnetic resonance (NMR) is another popular technique for determining the atomic structure of proteins. This method uses the interaction of the spins of atomic nuclei with a magnetic field to obtain information on interatomic distances and the local environment of atoms. Otto Stern's first measurement of the magnetic moment of the proton in 1933 earned him the Nobel Prize in Physics in 1946 [27, 28]. In 1946, Felix Bloch and Edward Purcell independently discovered new measurement methods using nuclear magnetic resonance, for which they shared the Nobel Prize in Physics in 1952 [29, 30]. These works founded the field of NMR spectroscopy, which led to the first applications of NMR in molecular biology in the 1960s [31]. But it was not until the 1980s that the technique was mature enough for determining the three-dimensional structures of proteins in solution, with the work of Richard Robert Ernst and Kurt Wüthrich. Both made decisive contributions to improving the accuracy of NMR techniques for studying protein structures. Ernst introduced methods for better resolving NMR signals by developing multidimensional techniques [32]. Unlike 1D NMR, which produces a single spectrum showing the resonance of nuclei as a function of a single parameter, usually frequency, multidimensional NMR records the interactions between different nuclei. This makes it possible to distinguish signals that would be superimposed in a one-dimensional spectrum, and to obtain correlations between different atoms, making it possible to analyze more complex structures. Wüthrich then used these new tools to map atomic interactions in proteins in solution, facilitating the precise reconstruction of their three-dimensional structure [33]. For their respective contributions, Ernst was awarded the Nobel Prize in Chemistry in 1991, followed by Wüthrich in 2002.

Experimental context Unlike X-ray crystallography, NMR structure determination is generally carried out in solution. This is an advantage because proteins are in an environment closer to their physiological environment than in a crystal. It is also very useful for studying proteins that are difficult to crystallise. However, the protein studied by NMR must be soluble at very high concentrations without forming aggregates. NMR is generally considered to be limited to the study of small proteins, with a small molecular mass, because the complexity and noise of NMR spectra increases significantly with protein size, making data assignment more difficult. In the 1970s and up to the early 1980s, 2D NMR experiments enabled detailed structures to be resolved down to around 10kDa [34]. Improvements in experimental techniques enabled larger and larger proteins [35] to be studied. One

of the first of these methods was isotopic labelling of C and N, which consists of replacing these atoms with their isotopes ^{13}C and ^{15}N, which contribute to the NMR signals and allow their interactions with other atoms to be detected. Using this technique, proteins of the order of 20 kDa have been studied [36]. Deuteration, which is the random replacement of hydrogen atoms by the ^{2}H isotope, pushed the limit of precise resolution towards 30 kDa in the 1990s [37]. Specific pulse sequences for studying large proteins, such as transverse relaxation optimized spectroscopy (TROSY), have also been developed [38]. By combining all these advances in an appropriate way, NMR was able to be applied to the study of the dynamic properties of large protein complexes such as the 670 kDa 20S proteasome [39]. In contrast to X-ray diffraction, which provides a static image representing a spatial and temporal average of a protein's structure, NMR has the advantage of being able to analyse dynamic properties on a large range of time scales [40, 41, 42, 43].

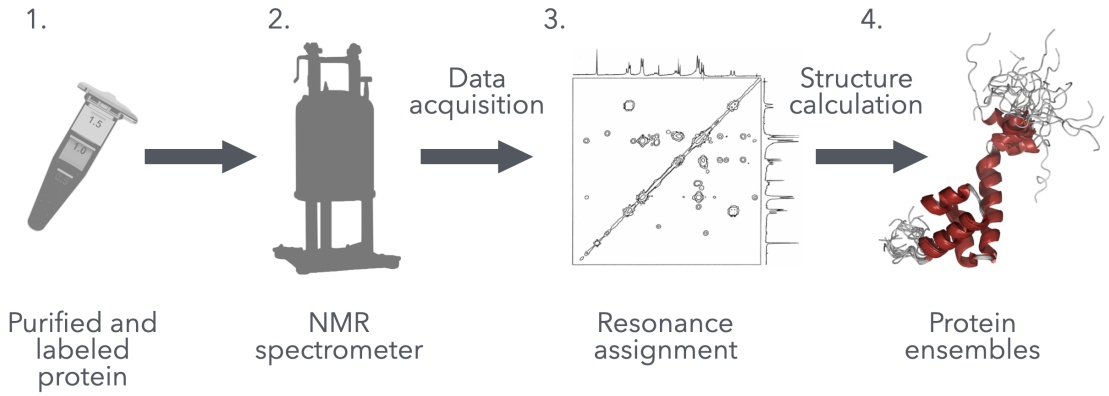


Figure 1.10: **Workflow of structure determination by NMR.** 1. Proteins are purified and isotopically labeled, then dissolved at very high concentration. 2. Radio frequency pulses are applied to the sample, temporarily exciting the nuclear spins of specific atoms. When these nuclei relax back to their original state, they emit radiation that depends on their chemical environment in the protein. 3. The emitted radiation is recorded for different pulse sequences, and the resulting spectra are analyzed to measure chemical shifts and other interactions. 4. The relative positions of atoms are calculated based on the NMR data, and a series of possible protein structures are generated.

1.2.4 Cryo-EM

History Optical microscopy does not allow the observation of objects smaller than the wavelength of visible light, in the order of a hundred nanometers. In his thesis in 1924, the French physicist Louis de Broglie presented his theory on the existence of a wave-like nature for electrons [44]. This hypothesis was confirmed in 1927 by the electron diffraction experiments of Clinton Davisson and Lester Germer and by George Paget Thomson, independently [45, 46, 47]. For the validation of his hypothesis, Louis de Broglie was awarded the Nobel Prize in Physics in 1929, and for the experimental discovery Davisson and Thomson became co-recipients of the Nobel Prize in 1937. Following these discoveries, the first prototype electron microscope was constructed by Ernst Ruska and Max Knoll in 1931 [48]. The idea is to use the much shorter wavelength of high-speed electrons, giving far better theoretical resolution. This first prototype didn't allow magnification beyond that of optical microscopy, but two years later Ruska built the first electron microscope to exceed the capabilities of optical microscopy [49, 50]. For his work in electron microscopy (EM), Ruska was awarded half of the 1986 Nobel Prize for Physics. Immediately after these discoveries, this new technique was applied to the study of biological objects [51]. It led to a better understanding of the cell, with the progressive discovery of organelles. Notably, George Emil Palade discovered the existence of the ribosome in the cell using an electron microscope in 1955, which earned him a share of the 1974 Nobel Prize for Medicine [52]. Although it has enabled all these advances, traditional electron microscopy suffers from limitations when it comes to observing biological samples, particularly when preparing them for observation. In order to prevent any disturbance to the trajectory of the incident electron beam, the sample is observed in a vacuum. However, biological samples contain water, which evaporates in a vacuum. To solve this problem, the samples were dehydrated and fixed by being covered with a layer of metal, but this treatment fundamentally denatures the biological sample. In 1981, Jacques Dubochet and his team invented a new sample preparation technique, which led to the development of cryo-electron microscopy and earned him a share of the Nobel Prize in Chemistry in 2017 [53, 54, 55]. The principle lies in the extremely rapid cooling of the sample. The method involves rapidly immersing the sample contained in a thin layer of water in a solution of liquid ethane. The water in the sample cools at a rate of around 10,000°C per second and has no time to crystallise, turning into vitreous ice. The latter has the same density as water, so the sample is embedded in it without being destroyed. Gradually, improvements in electron

detectors and advances in image processing algorithms have made cryoEM a key technique in structural biology [56]. As of 2017, the number of PDB entries from electron microscopy experiments has exceeded the number of entries from NMR experiments (Fig. 1.7). Today, the most precise cryoEM experiments report atomic resolution [57, 58].

Experimental context CryoEM has many advantages over X-ray crystallography and magnetic resonance methods. Indeed, it does not require crystallization of the macromolecule studied, which simplifies sample preparation and poses fewer restrictions on its purity. Moreover, this method does not require a large quantity of samples. In addition, rapid cooling of the sample preserves it in a state close to its native state, which is important for understanding the object in its biological context. High-resolution images and classification algorithms now make it possible to distinguish between the different conformations present in the sample, which allows a better understanding of the functional aspect of the protein or complex studied. These practical aspects of CryoEM also stimulated the development of integrative structural biology approaches toward characterising protein continuous conformational heterogeneity [59], as we will describe in more details in Section 2.4.2.

1.3 Protein dynamics

Proteins move The function of most proteins, whether in catalysis, transport, or signaling activities, often relies on their ability to bind to other molecules, known as ligands. This binding capability is intricately linked to the inherent flexibility of proteins. Indeed, proteins are not rigid and static objects, but dynamic entities whose movements play an essential role in their biological function. Protein structures are stabilised by weak interactions that are easily broken and reformed, giving them their plasticity. This flexibility allows them to adapt to environmental changes and the presence of other molecules by adopting different three-dimensional conformations. In this way, a protein is better represented by a set of conformations than by a single static structure. Proteins therefore inhabit different states, and these transitions between different states can occur over a wide range of time and distance scales (Table 1.1), ranging from small adjustments on a side chain to global conformational rearrangements. These changes have long been known to enable the biological function associated with the protein to be realised [60, 61]. One way of describing the populations of different conformations is through the energy landscape. The most populated and stable states correspond to energy minima,

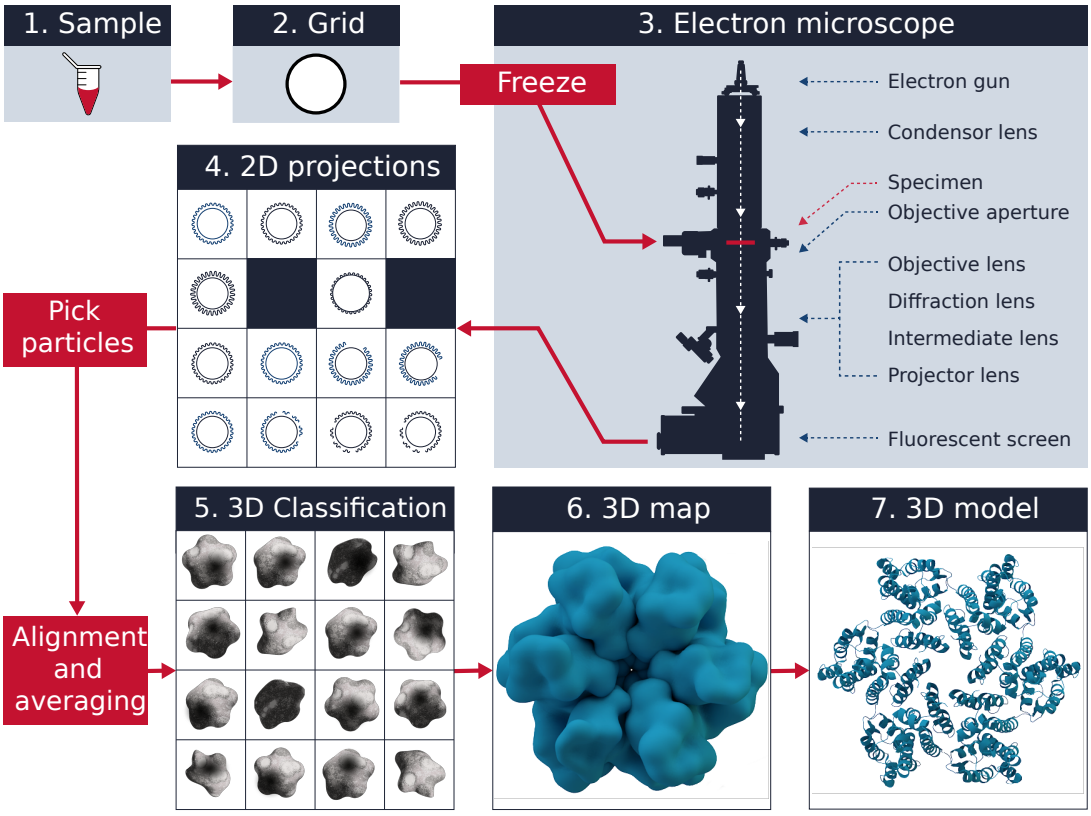


Figure 1.11: **Workflow of structure determination by cryoEM.** 1. The sample, present in the aqueous solution and purified. 2. A few microliters of solution are placed on a grid containing holes of the order of a micrometer. The excess water is removed and the grid is plunged in a solution of liquid ethane, which causes the formation of a thin layer of vitreous ice containing the samples within the holes. The samples all have a random orientation of their own. 3. The frozen grids are then loaded into the electron microscope. In its upper part, the latter creates and condenses a beam of coherent electrons on the sample. Its lower part is dedicated to the enlargement and acquisition of the electrons having passed through the sample. 4. Two-dimensional images of the samples, called poses, are collected. They correspond to the 2D projection of the 3D objects under their own orientation, which is unknown at this stage. Typically, tens to hundreds thousands poses are captured. Depending on the method used, a pose classification algorithm can be used to classify conformational heterogeneities and give rise to several three-dimensional reconstructions. 5. The poses are processed by algorithms that allow them to be assigned a specific orientation. 6. A three-dimensional map of the sample is created. 7. A model of the protein is built. The last steps are repeated iteratively in order to refine the model.

Source: Leeds University Library, CC BY-SA 4.0

represented by wells, and the transitions between the different states correspond to the crossing of an energy barrier. This energy landscape is often very complex, with a large number of minima. At the temperature of the organism, the energy supplied by thermal agitation is often sufficient for a given protein to explore several minima [62, 63].

Intrinsic and interacting motions Two types of motions can be distinguished in proteins: intrinsic motions and motions induced by interactions with other molecules. Intrinsic motions refer to spontaneous motions of the protein in the absence of interactions. They are induced by thermal agitation and can refer to small atomic fluctuations as well as large collective motions (Table 1.1). Interaction-induced motions, on the other hand, are induced by the binding of the protein to other molecules, such as small ligands, nucleic acids or other proteins. This concept, known as induced fit, was first suggested by Daniel E. Koshland in 1958 [64]. The idea is that the ligand causes a conformational change in the protein that increases its affinity for the ligand, stabilizing the interaction and optimizing the fit between the protein and its binding partner. This concept was originally widely used to describe conformational changes in proteins. However, while this model can explain changes that are plausible on a local scale, it struggles to explain the collective movements that can lead to the rearrangement of entire domains [65]. This is where the concept of conformational selection comes in, a concept first proposed by Gregorio Weber in 1972 [66]. The idea is that proteins are constantly exploring a set of conformations, known as a pre-existing equilibrium, and that only a fraction of these states are predisposed to binding to the ligand. The ligand then preferentially binds to the conformations that are favourable to it, known as conformational selection. In this view, structural rearrangements occur because the intrinsic dynamics of proteins allow them to do so [67]. These two views are not contradictory but complementary [68], and are illustrated in the Fig. 1.12.

Motions in enzymes Enzymes, which are proteins that catalyse chemical reactions, are among the most widely studied proteins. They bind to their substrate and reduce the activation energy required for the chemical reaction. They can also bind to other molecules called cofactors, which can regulate their activity by increasing it (activation) or decreasing it (inhibition). The places on the enzyme that bind with ligands and catalyse the reaction are called active sites. They generally exist in the form of cavities in the protein's three-dimensional structure where a microenvironment favourable to binding with a specific ligand exists. Interactions between ligands and proteins are stabilised by non-covalent bonds of the same type as those stabilising the protein structure. To illustrate the diversity

of protein movements and their purpose, we propose 3 examples of movements that exist in different enzymes. One very local movement is the rotation of side chains. The different conformations obtained by side-chain rotation are called side-chain rotamer. These changes play an essential role in the adaptation of the protein to the ligand by precisely modulating the environment of the active site, and are present on 90% of active sites [69].

Another important type of movement during interaction is loop closure around the active site, as observed in triosephosphate isomerase (TPI). This enzyme catalyses the conversion of dihydroxyacetone phosphate (DHAP) to glyceraldehyde-3-phosphate (G3P) [70]. This is a step in the metabolism of glucose by the organism. When the loop closes over the active site, it protects it from the external influence of solvent. The distance between the top of the loop and the open and closed conformations is around 7 Å. In the absence of ligand, both states are accessible, and the characteristic time between the two states is about 10^{-4} s [71].

Another notable example of movements within enzymes is whole-domain movements. This type of movements is seen, for example, in aspartate aminostransferase (AST), which is a key enzyme in amino acid metabolism. AST catalyses the transamination reaction between aspartate and alpha-ketoglutarate, leading to the formation of oxaloacetate and glutamate. Upon binding to the substrate, a pivot movement of the small domain (approximately 130 amino acids) is observed. This movement has the effect of bringing the two parts of the enzyme together, closing the active site, protecting the substrate from the external environment and positioning it for catalysis [72]. Figure 1.13 shows some examples of movements at different scales.

Intrinsically Disordered Proteins Intrinsically Disordered Proteins (IDPs) and Intrinsically Disordered Regions (IDRs) refer to proteins or protein segments that do not exhibit a stable three-dimensional structure under physiological conditions [73]. Unlike classical globular proteins, which adopt a defined conformation, IDPs exist as flexible and dynamic chains. The amino acid sequence of a protein dictates its folding, and in the same way, it determines the presence or absence of disorder [74]. IDPs and IDRs play a central biological role by facilitating interactions with a wide range of molecular partners, such as other proteins or nucleic acids, thereby endowing these proteins with high functional versatility, sometimes referred to as functional promiscuity [75, 76, 77]. This ability to interact with multiple partners allows IDPs/IDRs to participate in essential cellular processes, such as cell signaling and transcriptional regulation, and to assume complex regulatory functions [78, 79].

IDPs/IDRs exist at various levels of potential foldability and can acquire a more rigid structure upon specific interactions, a phenomenon known as induced folding [80, 81]. In eukaryotes, more than 30% of the proteome consists of proteins containing disordered regions with more than 50 consecutive residues [74]. These disordered regions are often rich in charged amino acids and are devoid of bulky hydrophobic residues, which prevents the formation of a compact hydrophobic core, characteristic of globular proteins [82].

Motion	**Spatial displacement (Å)**	**Characteristic time (s)**	**Energy source**
Fluctuations (e.g., atomic vibrations)	0.01 to 1	10^{-15} to 10^{-11}	k_bT
Collective motions (A) fast, infrequent (e.g., Tyr, Phe ring flips) (B) slow (e.g., domain movement; hinge-bending)	0.01 to > 5	10^{-12} to 10^{-3}	k_bT
Triggered conformational changes	0.5 to > 10	10^{-9} to 10^{3}	Binding interactions

Table 1.1: **Typical characteristics of protein motions.**
Values extracted from [83]. k_bT is the thermal energy.

1.4 Physics based approaches to protein dynamics

1.4.1 Normal mode analysis

Context Normal mode analysis (NMA) is a technique used to study the collective motions of proteins by modelling their vibrations around a steady state. This technique has been applied to proteins since the early 1980s [86, 87, 88]. It is a computationally inexpensive technique, especially when compared with molecular dynamics (MD) simulations. NMA makes it possible to study low-frequency and large-amplitude deformation. These motions are valuable from a structural biology point of view because numerous examples have shown that functionally important transitions follow the trajectory of one or more low-frequency normal modes [89, 90, 91, 92, 93, 94, 95, 96, 97, 98]. Moreover, low-frequency normal modes are highly

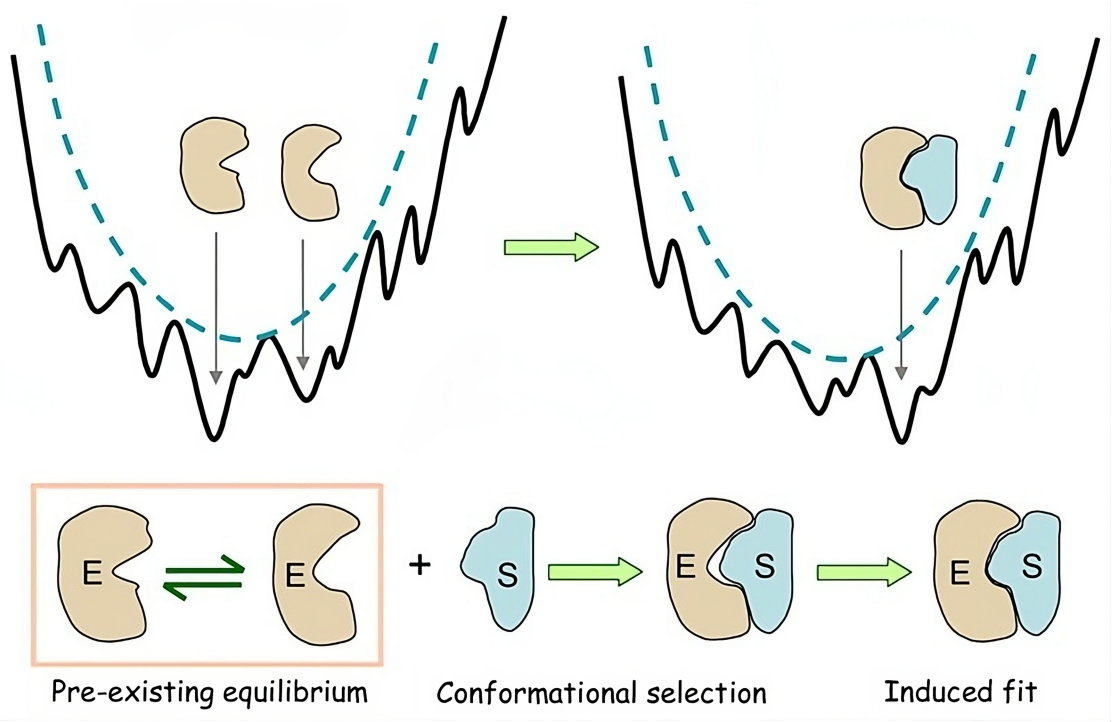


Figure 1.12: **Representation of the mechanisms of conformational selection and adjustment induced in the enzyme-substrate interaction**. Left: The diagram represents the pre-existing equilibrium between two conformations of the enzyme (E) in the absence of the substrate (S). The dashed blue curve corresponds to the harmonic potential of the protein approximated by NMA. Right: When the substrate binds to the enzyme with the most favourable conformation (conformational selection), its energetic landscape is modified in favour of the stability of the attached form (induced fit).

Figure extracted from [84].

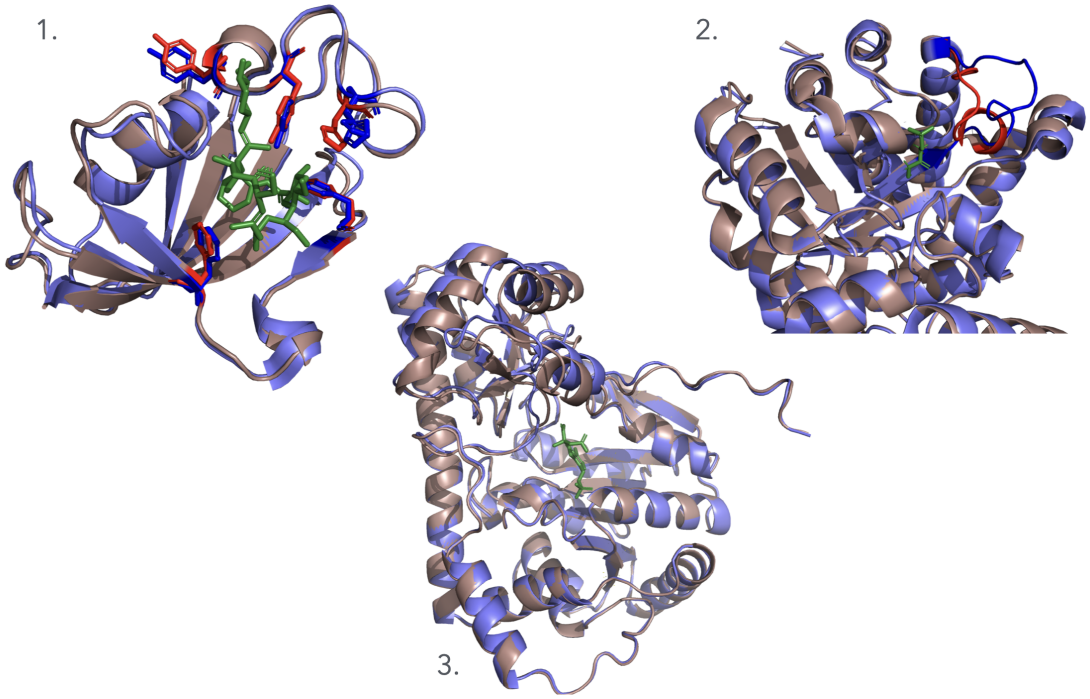


Figure 1.13: **Different scales of motions in enzymes**. Rearrangement of the aromatic chains of the FKBP12 enzyme on binding to the substrate. The bound form (1fkj) is shown in red and the unbound form (2ppn) is shown in blue. Example taken from [85]. 2. Loop opening and closing at the triosephosphate isomerase active site. The closed form (1ney) is shown in red and the open form (1ypi) is shown in blue. 3. Movement of the small domain of the enzyme L-aspartate aminostransferase during binding with its substrate. The closed form (1art) is shown in brown and the open form (1ars) is shown in blue.

conserved between homologous structures, which tends to confirm their functional validity [99, 100]. Tiwari et al. states that low-frequency modes are even more conserved than the structure itself [101]. This has significant implications, such as the possibility of detecting distant homologs by using their similarities in flexibility [102].

Principle Normal mode analysis is a method used to describe the flexible states available to a protein around its equilibrium position, i.e. when it is in a conformation corresponding to a minimum of energy. When an oscillating system, such as a protein, is slightly perturbed, a restoring force acts to bring the system back to its equilibrium state. This approach makes it possible to model small oscillations around this stable conformation. At equilibrium, the potential energy of the system $V(q)$ can be written by a Taylor expansion:

$$V(\mathbf{q}) = V(\mathbf{q}^0) + \left(\frac{\partial V}{\partial q_i}\right)^0 \eta_i + \frac{1}{2}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 \eta_i \eta_j + \ldots, \tag{1.1}$$

where q_i and q_j represent the position of components i and j and the deviation of component i from its equilibrium configuration is given by $\eta_i = q_i - q_i^0$. The exponents 0 indicate that the development is carried out in the equilibrium state. The first term represents the minimum of the potential, which can be set to 0, and the second term is zero because the system is in a state of equilibrium. The expression for the second-order potential therefore becomes

$$V(\mathbf{q}) = \frac{1}{2}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 \eta_i \eta_j = \frac{1}{2}\eta_i H_{ij} \eta_j, \tag{1.2}$$

where H_{ij} is the Hessian matrix containing the second derivatives of the potential with respect to the components of the system.
The kinetic energy T of the system is expressed in terms of the velocities of the particles:

$$T(\mathbf{q}) = \frac{1}{2} m_i \dot{\eta}_i \dot{\eta}_i = \frac{1}{2} \dot{\eta}_i m_i \dot{\eta}_i, \tag{1.3}$$

where $\dot{\eta}_i = \frac{d\eta_i}{dt}$ is the velocity of component i, and m_i is the mass of component i. In matrix notation, this becomes:

$$T(\mathbf{q}) = \frac{1}{2} \dot{\boldsymbol{\eta}}^{\mathrm{T}} \mathbf{M} \dot{\boldsymbol{\eta}}, \tag{1.4}$$

where $\mathbf{M}$ is the diagonal mass matrix.

The Lagrangian $\mathcal{L}$ of the system is given by:

$$\mathcal{L} = T - V = \frac{1}{2}\dot{\eta}_i m_i \dot{\eta}_i - \frac{1}{2}\eta_i H_{ij} \eta_j. \tag{1.5}$$

Using the Euler-Lagrange equations:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\eta}_i}\right) - \frac{\partial \mathcal{L}}{\partial \eta_i} = 0, \tag{1.6}$$

we compute:

- The partial derivative of the Lagrangian with respect to the velocity:

$$\frac{\partial \mathcal{L}}{\partial \dot{\eta}_i} = m_i \dot{\eta}_i, \tag{1.7}$$

  and its time derivative:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\eta}_i}\right) = m_i \ddot{\eta}_i, \tag{1.8}$$

  where $\ddot{\eta}_i = \frac{d^2\eta_i}{dt^2}$ is the acceleration.

- The partial derivative of the Lagrangian with respect to the displacement:

$$\frac{\partial \mathcal{L}}{\partial \eta_i} = -H_{ij}\eta_j. \tag{1.9}$$

  Substituting these into the Euler-Lagrange equations gives the equations of motion:

$$m_i \ddot{\eta}_i + H_{ij}\eta_j = 0, \tag{1.10}$$

  or, in matrix form:

$$\mathbf{M}\ddot{\boldsymbol{\eta}} + \mathbf{H}\boldsymbol{\eta} = 0. \tag{1.11}$$

Assuming a harmonic solution for the displacements:

$$\boldsymbol{\eta}(t) = \mathbf{a}_k \cos(\omega_k t + \delta_k), \tag{1.12}$$

where $\mathbf{a}_k$ is the amplitude vector for mode k, ω_k is the angular frequency of mode k, and δ_k is the phase constant.
Calculating the second derivative of $\boldsymbol{\eta}(t)$:

$$\ddot{\boldsymbol{\eta}}(t) = -\omega_k^2 \mathbf{a}_k \cos(\omega_k t + \delta_k). \tag{1.13}$$

Substituting $\boldsymbol{\eta}(t)$ and $\ddot{\boldsymbol{\eta}}(t)$ back into the equation of motion:

$$\mathbf{M}\left(-\omega_k^2 \mathbf{a}_k \cos(\omega_k t + \delta_k)\right) + \mathbf{H}\left(\mathbf{a}_k \cos(\omega_k t + \delta_k)\right) = 0. \tag{1.14}$$

Simplifying, we get:

$$\left(\mathbf{H} - \omega_k^2 \mathbf{M}\right) \mathbf{a}_k \cos(\omega_k t + \delta_k) = 0. \tag{1.15}$$

Leading to the generalized eigenvalue problem:

$$\mathbf{H}\mathbf{a}_k = \omega_k^2 \mathbf{M}\mathbf{a}_k. \tag{1.16}$$

The eigenvectors $\mathbf{a}_k$ contain the directions and relative amplitudes of motion for each atom in the system, and the eigenvalues ω_k^2 correspond to the squared angular frequencies of the normal modes.

The first 6 normal modes correspond to the translation and rotation of the rigid body, without deformation, and are therefore trivial. Apart from these zero-frequency modes, it is generally the lowest-frequency modes that are analysed to describe the motion of the protein.

RTB and Elastic Network Models Early studies of normal modes at the atomic level used empirical potentials similar to those used in molecular dynamics. However, this implies certain limitations. Indeed, the structure must be relaxed to its minimum energy in the force field used, which deforms the structure, frequently leading the NMA to be performed on a different structure from the initial one. In addition, for classical NMA is normally performed with all atoms, including hydrogen atoms, so diagonalization of the Hessian matrix of size $3N \times 3N$ where N is the number of atoms can quickly be intractable for large proteins. To solve this problem, low-resolution approaches such as Rotation-Translation of Blocks (RTB) have been developed [103, 104]. This approach involves dividing the protein into rigid blocks containing one or more residues, in order to reduce the number of

elements to be considered.

Another approach, introduced by Tirion, showed that these complicated potentials could be replaced by much simpler potentials while largely preserving the low-frequency movements [105]. These potentials are of the form

$$V = \sum_{r_{ij}<R_c} C \left(\vec{r_{ij}} - \vec{r_{ij}}^{\,0}\right)^2 \tag{1.17}$$

where $\vec{r_{ij}}$ is the displacement vector between the atoms i and j, and $\vec{r_{ij}}^{\,0}$ is the initial vector between these two atoms, R_c is a distance threshold beyond which pairs are no longer considered and C is a rigidity constant associated with the bond between two atoms. In practice, this means that we model the protein as a network of springs where a pair of atoms is linked if it is below the distance threshold, hence the name Elastic Network Model. This aproximation takes into account the relative directions of the atoms, and is often referred to as the Anisotropic Network Model (ANM) [106]. Generally the spring constant C is chosen to be uniform for all pairs, but some studies have tried to assign higher constants to rigid domains [107, 108].

An even simpler elastic network called the Gaussian Network Model (GNM) consists of an isotropic version of the previous model, i.e. only the amplitudes of the fluctuations of the atoms are considered, and the Hessian matrix of dimension $3N \times 3N$ is replaced by a matrix of size $N \times N$, called the Kirchoff matrix [109, 110]. However, since the deformations considered here are isotropic, this model does not allow alternative conformations to be generated in the vicinity of the starting conformation by following collective deformation directions, as is possible with the anisotropic model. Nevertheless, the fluctuations obtained by the GNM tend to correlate well with the experimental B-factors [111, 112]. Kundu et al. found an average correlation of 0.59 on a set of proteins, which is slightly better than the correlation obtained when using ANM to predict isotropic B-factors [112]. Normal mode analysis produce linear motions, which deforms the structure unrealisticly at large amplitude. To tackle this issue Hoffmann and Grudinin introduced a RTB-based method with non-linear extrapolation of the motion, using instantaneous linear and angular velocities of the rigid block [113]. Additionally, HOPMA is a method that proposes breaking certain connections within the Elastic Network Model (ENM), specifically those corresponding to isolated residues that are distant in the sequence. This allows, for example, the facilitation of protein opening by removing non-covalent interactions that would otherwise keep the protein in a closed state [114].

1.4.2 Molecular Dynamics (MD) simulations

Principle Molecular dynamics is a computer simulation technique used to study the movement of atoms and molecules over time. It involves solving Newton's equations of motion for each particle in the system. The forces exerted on atoms are described by a collection of interatomic potentials called force fields. The most common force fields used in MD include interatomic potentials describing the energy between covalent atoms as a function of their bond distance, bond angles and bond torsion. These force fields also include terms that apply to atoms not bound by a covalent bond, modelling van der Waals forces via the Lennard-Jones potential and electrostatic forces via the Coulomb potential [115]. The forces exerted on the particles determine their acceleration, and numerical integration algorithms are used to calculate the trajectories of each atom over very short time steps, generally of the order of femtoseconds. Molecular dynamics is the most accurate calculation method for studying the dynamics of a protein. However, because of the small integration step and the number of operations required to calculate the potential of each atom, molecular dynamics simulations are very demanding in terms of computing resources, which is their main limitation.

Context The first molecular dynamics simulation applied to a protein dates back to 1977, when the dynamics of a small protein of 58 amino acids, the bovine pancreatic trypsin inhibitor, was simulated over a period of 9.2 picoseconds [116]. This work, among others, is one of the achievements recognised in the Nobel Prize for Chemistry awarded in 2013. Since then, the increase in computing power and the evolution of MD techniques have made it possible to apply these techniques to larger systems and over much longer simulation times. In the 2000s, the longest molecular dynamics simulations were typically of the order of microseconds, and in the 2010s, of the order of milliseconds [117].

Principal component analysis of MD trajectory PCA is a linear dimensionality reduction technique used to project high-dimensional coordinates into a low-dimensional space. This space is formed with the axes that maximise the variance of the data, thereby preserving as much information as possible [118]. In molecular dynamics, it is used to extract the main deformation directions from conformations from different simulation timesteps. The first step in this analysis is to choose a reference structure and align the other structures with it, minimising the sum of the squared deviations of the pairs of atoms. Next, we calculate the

covariance matrix of the C positions, whose elements are given by

$$C_{ij} = \frac{1}{N} \sum_{k=1}^{N} \left(x_i^{(k)} - \overline{x_i} \right) \left(x_j^{(k)} - \overline{x_j} \right), \tag{1.18}$$

where $x_i^{(k)}$ and $x_i^{(k)}$ denote the coordinates of structure k at position i and j respectively, and $\overline{x_i}$ and $\overline{x_j}$ the mean coordinates at these positions. This matrix is then diagonalized, i.e. we find a Λ matrix for which

$$C = U \Lambda U^T \tag{1.19}$$

where the matrix U is the matrix containing the eigenvectors corresponding to the principal directions of variance and Λ is the diagonal matrix containing the eigenvalues associated with the eigenvectors. The eigenvectors are interpreted as the principal directions of deformation and the eigenvalues provide information about the amount of positional variance of the structures explained by the associated eigenvector. PCA was first used to describe protein dynamics in 1992, by Garcia on a 240 picosecond simulation of the dynamics of Crabin [119]. He observed that the motions of distant parts could be highly correlated, but that the motion in his example was essentially non-linear because the main direction of deformation contributed only 36% of the total positional variance observed in the simulation. Amadei et al. applied the same method and highlighted the fact that it is generally possible to reduce the dimensionality of the motions to 1% of the size of the original Cartesian space, showing that the internal motions of proteins are highly constrained [120]. They named the space formed with the principal eigenvectors essential space (ES). Since then, the use of PCA to analyse the movements produced by molecular dynamics has become widespread. In particular, this method has been used to test the validity of normal modes experimentally, by comparing the subspaces produced by non-trivial normal modes and the ES [121, 122]. While PCA was first applied to MD trajectories, it has also proven valuable for the analysis of structural ensembles derived from experimental techniques. Teodoro et al. showed that PCA can efficiently analyze experimentally determined X-ray structures to capture dominant motions and reduce the complexity of modeling large-scale protein flexibility, as shown by their study of HIV-1 protease using over 130 crystal structures [123]. Similarly, Yang et al. applied PCA to ensembles of NMR structures and X-ray structures, demonstrating the extracted collective motions can be similar to those predicted by elastic network models and can

identify motions that are functionally relevant [124, 125]. One measure of subspace similarity used in this case is the root mean square inner product (RMSIP),

$$\mathrm{RMSIP}(I, J) = \left(\frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} (u_i \cdot v_j)^2 \right)^{\frac{1}{2}}, \tag{1.20}$$

where I and J denote the dimensions of the subspaces formed by the eigenvectors from each method, u_i is the i^{th} vector in the first subspace, and v_j is the j^{th} vector in the second subspace [126]. The RMSIP is between 0 and 1. An RMSIP of zero corresponds to mutually exclusive spaces, and a score of 1 corresponds to identical spaces. David and Jacobs. found a RMSIP between the 20 first non-trivial normal modes and the ES of around 0.5 on several examples, which is considered to be a good correspondence [122].

Chapter 2

Deep Learning background in bioinformatics

This chapter introduces deep learning techniques relevant to understanding the manuscript, as well as methods commonly used in the field of structural biology and protein dynamics. Section 2.1 introduces the principles of artificial neural networks and presents two types of architectures that will be used later in the manuscript. Section 2.2 covers concepts from natural language processing and connects their application to natural language with their application to protein sequences. Section 2.3 discusses protein structure prediction from amino acid sequences, while section 2.4 focuses on various methods used to predict protein flexibility.

2.1 Introduction to Artificial Neural Networks

Artificial neural networks (ANNs) have a wide range of applications in bioinformatics. This section aims to give a short introduction to the essential concepts of this field.

Machine learning Machine learning is a sub-field of artificial intelligence whose aim is to create statistical algorithms that can learn from a set of numerical data and generalise to new data without specific instruction. One of the most important areas of machine learning is artificial neural networks.

Artificial neural networks Artificial neural networks are mathematical models inspired by the function of biological neurons. They consist of a set of interconnected neurons where the strength of the connection is represented by a numerical parameter called a weight. They are used to approximate complex functions by

learning non-linear relationships between input and output data. We introduce the important concepts in this field using the classic example of multilayer perceptrons (MLP), first introduced in the 1960s [127, 128]. Nowadays, these basic networks provide a building block for more complex architectures.

Neural networks are generally composed of several layers, with each layer containing a certain number of neurons. A neuron is a parameter whose value is chosen arbitrarily when it belongs to an input layer of the network, or calculated using the values of the neurons in the previous layer when it does not belong to the first layer.

In MLPs, two neurons in two adjacent layers are linked by a parameter called the weight. By analogy with biology, this parameter can be seen as the strength of the connection between two neurons. The value taken by a neuron is called activation. In the general case of fully connected networks, this activation is calculated as the sum of the activations of the neurons in the previous layer, each multiplied by its corresponding connection weight. To this sum is added a parameter specific to each neuron, called the bias, which can be seen as the neuron's activation threshold, and then a non-linear function called the activation function is applied to the whole. Mathematically, the activation $a_k^{(j)}$ of neuron k in layer j is equal to

$$a_k^{(j)} = \sigma\left(\sum_{i=0}^{N} w_{ik}^{(j)} a_i^{(j-1)} + b_k^{(j)}\right), \tag{2.1}$$

where σ is the activation function, $w_{ik}^{(j)}$ is the weight linking neuron i in layer $j-1$ to neuron k in layer j, and $b_k^{(j)}$ is the bias of neuron k in layer j.

Generally speaking, the first layer of the network will take parameters from a dataset as its activation. The last layer, or output layer, contains the prediction of the neural network.

Training In order for the network's prediction to make sense, it must go through a training phase. Learning consists of iteratively modifying the parameters of the network in order to minimise a quantitative metric called the loss function. In the case of supervised learning, the prediction targets associated with the training data are known, so the error function is commonly a metric measuring the difference between the prediction and the target. The lower the loss function, the more similar the network prediction and the target.

The classic learning method is gradient backpropagation, first introduced in 1970 and popularized for the training of neural networks in 1980 [129, 130]. This involves calculating the partial derivative of the error function for each network

parameter, then updating the parameters by subtracting this derivative multiplied by a certain coefficient. Mathematically, the parameters are updated as follows

$$\omega_i \to \omega_i - \gamma \frac{\partial L}{\partial \omega_i}, \tag{2.2}$$

where ω_i is a network parameter, L is the loss function, and γ is the learning rate. This procedure for updating the network parameters corresponds to classic gradient descent. By repeating the operation many times and on many examples, we hope that the network will reduce its error function. In general, the network parameters are not updated for each example, but for a batch of examples. An average correction is then made by averaging the gradients for each parameter. There are variants to these training methods. The most widely used is a stochastic method based on estimating the mean and variance of the gradients called Adam (for Adaptive Moment Estimation), which we will use in the following work [131]. These approaches aim to make network training faster and more reliable.

Convolutional Neural Networks (CNNs) are a class of deep neural networks that are particularly effective in processing data with a grid structure, such as sequences and time series [132, 133]. Although CNNs are often associated with two-dimensional data such as images, they are also applicable to one-dimensional data, making them suitable for the analysis of sequential data such as protein sequences. In one-dimensional CNNs (1D CNNs), convolutional layers apply a set of learnable filters that slide along the time dimension of the input sequence. These filters capture local patterns by calculating the dot product between filter weights and input sequence segments. A first key property of CNN is the translation equivariance, i.e. a shift in the input results in a corresponding shift in the output feature maps. This property enables CNNs to recognize motifs independently of their position in the sequence, which is particularly useful in the analysis of protein sequences where functional motifs may appear at different locations. Another key property is the ability to process sequence data of arbitrary size, meaning that the same network can be used on sequences of different sizes, which is useful in the case of protein analysis since proteins exist in different sizes.

Mathematically, the $a_k^{(j)}(t)$ activation of the k-th feature map at position t in the j layer is given by:

$$a_k^{(j)}(t) = \sigma\left(\left(w^{(j,k)} * a^{(j-1)}\right)(t) + b_k^{(j)}\right), \tag{2.3}$$

where $*$ is the cross-correlation operation, $w^{(j,k)}$ are the weights of the filter for

feature map k, $a^{(j-1)}(t)$ is the activation from the previous layer, $b_k^{(j)}$ is the bias term, and σ is the activation function. We will use this type of architecture in section 4.

Message Passing Graph Neural Networks (MPGNNs) are a family of neural networks designed to operate on graph-structured data [134]. In MP-GNNs, each node in the graph updates its representation by aggregating information from its neighbors through iterative message-passing steps. This process enables the network to learn representations that utilize both the features of the nodes and the topology of the graph. At each layer k, the hidden state $h_v^{(k)}$ of node v is updated based on its previous state and the messages received from its neighboring nodes $\mathcal{N}(v)$:

$$h_v^{(k)} = \sigma\left(W^{(k)} h_v^{(k-1)} + \text{AGGREGATE}_{u \in \mathcal{N}(v)} M^{(k)}\left(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}\right)\right), \quad (2.4)$$

where $W^{(k)}$ is the weight matrix at layer k, $M^{(k)}$ is the message function that computes the message from node u to node v using their node features and possibly the edge features e_{uv}, AGGREGATE is an aggregation function invariant to node permutation (such as sum, mean, or max) applied over the neighboring nodes, and σ is the activation function. Proteins can be represented in an invariant manner using a graph, where, for example, the nodes represent the types of amino acids and the edges represent the distances between them. This is very useful because the orientation of a single protein in 3D space is arbitrary, and having an invariant representation ensures that the learned model focuses on the intrinsic properties of the protein structure rather than its spatial orientation.

2.2 Protein Language Models (PLMs)

2.2.1 The Transformer

Context Technical advances in the field of natural language processing (NLP) are of great interest in bioinformatics. This is understandable given that DNA, RNA, and amino acid sequences are expressed in their own language. Indeed, many analogies can be drawn between natural language and biological sequences. To begin with, they share a limited number of units of meaning, arguably the words in the dictionary for natural language and the amino acid alphabet for amino acid sequences. The amino acids are arranged in a linear order in the same way as the words in a sentence, and this order is crucial to the meaning of the text and the

function of the protein. They also have the importance of context in common: in natural language, the meaning given to a word may depend on the other words in the sentence. The same applies to the functional interpretation of an amino acid or group of amino acids, which may vary according to the context of the sequence. What they also have in common is the presence of recurring patterns. In natural language, certain groups of words are often used together, and similarly, certain amino acids are often found together, forming easily identifiable patterns. In natural language, syntax and grammar impose strong constraints that can prevent the use of a word in a certain context, just as the replacement of one amino acid by another can be deleterious to the function of the protein. Finally, just as in natural language, several sentences can have the same meaning, and several amino acid sequences can code for proteins with similar functions. It is these similarities that have motivated the application of language processing techniques to proteins.

The field of NLP took a big leap forward in 2017 with the introduction of the transformer architecture by Google engineers [135], for its ability to process information in parallel, for its scalability and for its ability to capture long-distance dependencies in text. I will introduce a few important concepts from this architecture because it is the basis on which all large language models (LLMs) and protein language models (PLMs) are now built. These models are trained on large quantities of data, with learning tasks forcing them to acquire an internal representation of the data

Tokenization Tokenisation is a basic step that consists of dividing the sequence into small units called tokens. The set of tokens constitutes the network's vocabulary, i.e. the set of units it can process. This stage is often fairly trivial for PLMs compared with LLMs. For PLMs, it is generally carried out at the level of amino acids, of which there are a limited number (20 standard amino acids). PLMs generally have a vocabulary of 20 to 35 tokens [136, 137, 138, 139, 140, 141]. In addition to the 20 standard amino acids which are always included, there are often tokens assigned to non standard amino acids (for example, U for Selenocysteine) as well as a common token for all unknown amino acids (X). In addition, there are tokens specific to the operation of the model, such as the markers for the beginning (BOS) and end (EOS) of the sequence, the token for masking residues, and a padding token for adjusting the sequences to give them a uniform size.

Some PLMs [142, 143] adopt an approach used in language models, Byte Pair Encoding (BPE) [144], which involves dividing the sequence into individual characters, then iteratively merging the most frequent character pairs to form longer tokens. ProtGPT2 therefore works with a vocabulary of 50,256 tokens,

corresponding on average to 4 amino acids per token [143].

Vector representation In language models, tokens are translated into high-dimensional vectors called embeddings. The aim is to associate each token with a representation rich in meaning, capturing not only its own meaning, but also its relationships with other words. In this way, two vectors with similar meanings will also have a close direction in the embedding space. For example, playing with *glove-wiki-gigaword-50* [145], a simple model that encodes words in 50 dimensions - a relatively small size but sufficient to illustrate the point - we can see that the 5 vectors most similar to the word *proteins*, according to cosine similarity, are:

protein	0.91
molecules	0.85
genes	0.83
enzymes	0.83
bind	0.81

All these words share a meaning linked to *proteins*. What is more, by performing the operation

$$\vec{\text{proteins}} + \vec{\text{computer}} - \vec{\text{protein}},$$

we obtain a vector most similar to $\vec{\text{computers}}$ (with a cosine similarity of 0.88), suggesting the existence of a specific direction that encodes the notion of plural. Similarly, by performing

$$\vec{\text{tokyo}} + \vec{\text{france}} - \vec{\text{japan}},$$

we obtain a vector most similar to $\vec{\text{paris}}$ (cosine similarity of 0.92), indicating a direction corresponding to the concept of *capital*. This is a simple example, in practice in LLMs the embedding associated with a token depends on its context. In a similar way, PLMs learn representations for amino acids which we hope will contain as much biological information as possible.

Self-Attention The self-attention mechanism is an important component of Transformer. It allows the embeddings associated with each token to be updated using the embeddings of other tokens, independently of their positions, enabling them to enrich their representation by the context in which they are.

The first step is to calculate a linear transformation of the embeddings of each vector, by calculating $Q = XW^Q$, $K = XW^K$ and $V = XW^V$, where Q, K, and V are the matrices of the *queries*, *keys*, and *values*, and W^Q, W^K and W^V are the matrices of weights associated with this transformation, which are learned as the transformer is trained.

Attention is calculated using the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.5}$$

where d_k is the size of the vector K. Multiplying QK^T produces attention scores that indicate the relative importance of each token compared with all the others. By applying the softmax function, these scores are normalised to form a probability distribution. This probability distribution is then used to calculate a weighted sum of the *values* V, which will update the original embedding X by adding to it. The transformers' original paper proposed the idea of *heads* consisting of performing the attention operation in parallel with several matrices of weights W^Q, W^K and W^V, then concatenating them and linearly transforming them to calculate the update of each of the embeddings. The idea is to allow each *head* to focus on different ways in which context can change the meaning of an embedding.

Linear projection into higher dimensional space An important step, which should not be overlooked as it contains 2/3 of the Transformer's parameters [146], is the MLP part applied to the embeddings after the self-attention step. This projects each of the embeddings into a high-dimensional space (4x the dimension of the embedding in the original paper [135]) and then a non-linear function is applied, before projecting the result back into the original dimensional space of the embeddings. The aim of this operation is to increase further the expressiveness of each of the embeddings.

The attention and the MLP are the two fundamental parts forming a block. These blocks of attention are then stacked in the Transformer's architecture. The original implementation uses 6 of these stacks for the encoder, and 6 for the decoder.

Transformers as GNNs I want to make the connection here with the previous section on MPGNN by noting that Transformers can be seen as a special case of the general GNN framework. In this special case, Transformers operate on fully connected graphs. Viewed this way, with one head, the update of the node (or the token embedding) i of the layer $\ell + 1$ follows this logic:

$$h_i^{\ell+1} = h_i^\ell + \sum_{j \in \mathcal{S}} w_{ij}(W^{V(\ell)} h_j^\ell), \tag{2.6}$$

In this equation, the presence of h_i^ℓ represents the residual connection, meaning the preservation of the previous information of node i. The term $w_{ij}(W^{V(\ell)} h_j^\ell)$ corresponds to the message that i receives from j, where the coefficient w_{ij} is

calculated through the attention mechanism:

$$w_{ij} = \text{softmax}_j \left(\frac{(W^{Q(\ell)} h_i^\ell) \cdot (W^{K(\ell)} h_j^\ell)}{\sqrt{d_k}} \right), \tag{2.7}$$

The aggregation function consists of summing these messages weighted by the attention coefficients w_{ij}, allowing i to update its representation by taking into account all the other nodes in the graph.

2.2.2 Overview of PLMs

Encoder only Most PLMs allow a representation of a protein sequence to be calculated in the form of an amino acid embedding of a fixed size, specific to the model. The models that calculate these representations essentially use the BERT [147] architecture, or variations of it. This architecture is based on the transformer encoder with an unsupervised learning task, the mask language modelling. Mask language modelling involves corrupting input sequences by randomly masking some of their residues, then forcing the PLM to find the type of the masked amino acid. By repeating this operation on billions of sequences, the model becomes capable of predicting the probability of finding an amino acid at a given position in the sequence, showing that it is acquiring a contextual representation of the protein language. These representations can then be used for various prediction tasks, such as predicting function, contacts, structure, mutation effects, etc. These prediction tasks based on the representation learned by the model are called downstream tasks. These models are generally trained on large collections of sequences, such as UniRef [148, 149]. The main PLMs for representation learning include the Meta ESM series: ESM-1b [136], ESM-1v [150], ESM2 [138], which are based on the RoBERTa architecture [151], which is an evolution of BERT without the auxiliary task of predicting the next sequence. MSA-Transformer [152] extended the concept of self-attention to Multiple Sequence Alignment (MSA) columns in order to produce an embedding for MSAs. This component will be key in AlphaFold2 [153]. Rostlab, in ProtTrans paper [140], has trained a series of encoders used in language models on protein sequence databases, protBERT, ProtAlBert, protElectra based on BERT [147], Albert [154], Electra [155]. They also trained ProtT5, based on T5 architecture [156], which is based on the full transformer architecture (encoder and decoder), but they remove the decoder part during the inference of the embeddings.

Decoder only Some models are auto-regressive, i.e. they follow the principle of the Transformer decoder architecture. They use masking which forces them

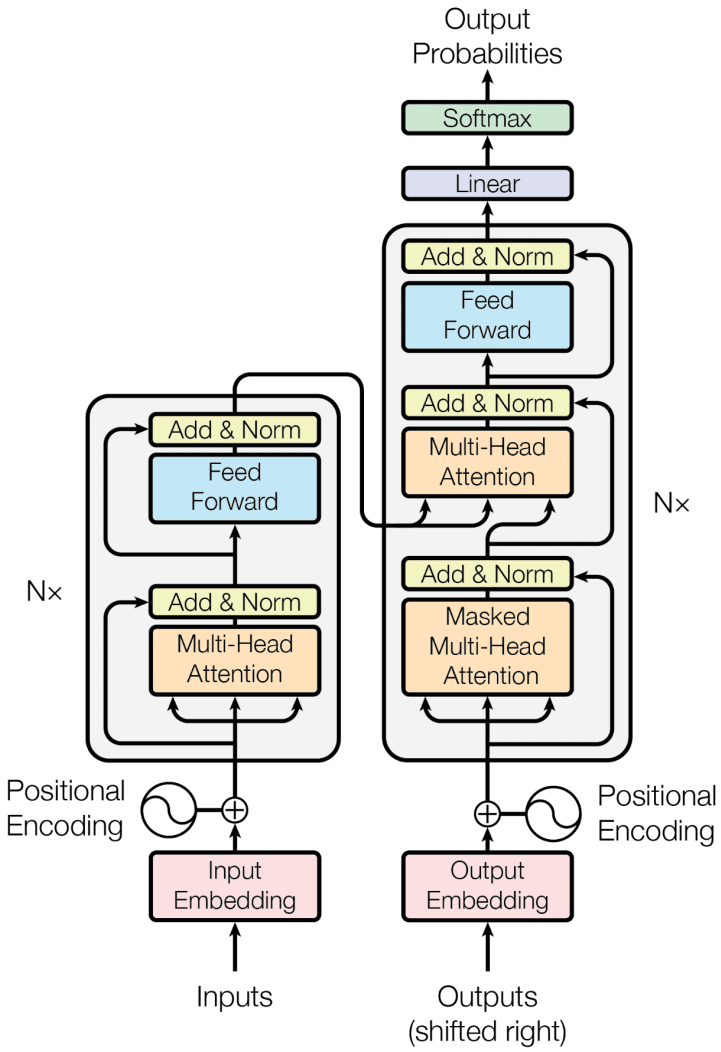


Figure 2.1: **Transformer architecture** The transformer architecture consists of two parts: the encoder on the left, and the decoder on the right. Both parts are formed with stacks of attention and position-wise linear layers. The decoder adopts a similar architecture to the encoder, but is distinguished by the masking task preventing the network from informing itself with the subsequent context, and by the presence of cross-attention enabling a correspondence to be established between the encoder's embeddings and those of the decoder. The figure comes from the original transformer paper [135]

to predict the next tokens from the beginning of the sequence, in the same way as the GPT model [157]. These include Progen [158] and its higher-level version Progen2 [159] and ProtGTP2 [143]. ProtTrans also contains auto-regressive models, ProtTXL and ProtXLNet, based on the Transformer-XL [160] and XLNet [161] architectures. These models are more designed for a generative approach rather than for representation learning. In fact, in ProtTrans [140], the extracted embeddings from these models underperform for the downstream tasks studied compared to the encoder-only models.

Multimodality Some language models take the approach of augmenting the information contained in the embedding with other forms of input, notably the three-dimensional structure of the protein associated with the sequence. For example ProstT5 [162] is a model trained to translate amino acid sequences into 3Di sequences, which are a sequential representation of the 3D structure, introduced in Foldseek [163]. Similarly, SaProt [164] is a language model trained on amino acid sequences enriched with the structural tokens produced by Foldseek. ESM-Gearnet [165] uses the Gearnet structural encoder [166] in order to enrich the ESM2 embeddings with structural information. ESM3 [167] is the latest generation of multimodal models succeeding ESM2, including encoding and decoding of protein sequence, structure, and function from a shared embedding.

2.3 Protein structure prediction with deep learning

2.3.1 CASP

Context The problem of predicting the three-dimensional conformation of a protein from its amino acid sequence has been around since the first protein structures were solved over 50 years ago. In addition to the importance of structure for characterizing protein function, the growing gap between the number of sequences and the number of experimentally known structures has helped to reinforce the importance of this problem. For example, in 2022, less than 0.03% of all known proteins were experimentally solved [168]. It was in this context that the Critical Assessment of Structure Prediction (CASP) was introduced in 1994 [169], an experiment aimed at assessing the performance of algorithms for predicting the three-dimensional structure of proteins from their sequence, in order to provide a clear framework for evaluating these methods, with the aim of accelerating

research in this field. This event takes place every two years, and evaluates the methods proposed by the different groups of participants on common targets, whose structure has not yet been made public. These targets are ranked according to their difficulty, which is assessed by their similarity to already known structures. Recent editions of CASP have been enriched by new categories, such as complex prediction, or, as from edition 15 in 2022, conformational ensemble prediction [170], reflecting the fact that today's major challenges in structural biology are protein dynamics and interactions. CASP12 saw the emergence of deep learning methods using coevolution signals to predict protein structures, before becoming a dominant approach since CASP13 [171, 172].

AlphaFold At CASP editions 13 and 14, the models developed by Google Deepmind, AlphaFold and AlphaFold2 respectively, distinguished themselves [173, 153]. On two occasions, they were the best performers in the high accuracy and topology modelling categories. AlphaFold2's performance was such a breakthrough that it is often considered to have solved the problem of folding single chain proteins [174]. Indeed, at CASP 14, Alphafold 2 achieved a median Global Distance Test (GDT) of 92.4 on all targets, far surpassing the second best method with a median GDT of 72.8. This score measures the similarity between the predicted protein and the target, ranging from 0 to 100, where 100 denotes maximum similarity. A score above 90 is considered competitive with experimental methods, and therefore a valid solution to the problem. The evolution of the performance of the best performing method of each edition is given in the Fig. 2.2. Since AlphaFold2 has had an extremely important impact on structural biology, having modeled structures for more than 200 million UniProt sequences [175, 176] and earned its creators half of the 2024 Nobel Prize in Chemistry, we will briefly present a few key elements of its architecture.

2.3.2 AlphaFold2

Preprocessing AlphaFold2 first starts with a preprocessing step. This consists of taking the amino acid sequence, finding its homologous sequences and constructing a multiple sequence alignment (MSA). The intuition behind this process is to exploit the co-evolution signals available in the MSA, which are informative for determining structure. Indeed, two residues close together in three-dimensional space will tend to coevolve [177]. AlphaFold2 also looks for the existence of resolved structures for homologs, and if available extracts a pairwise representation of the distance, used as template.

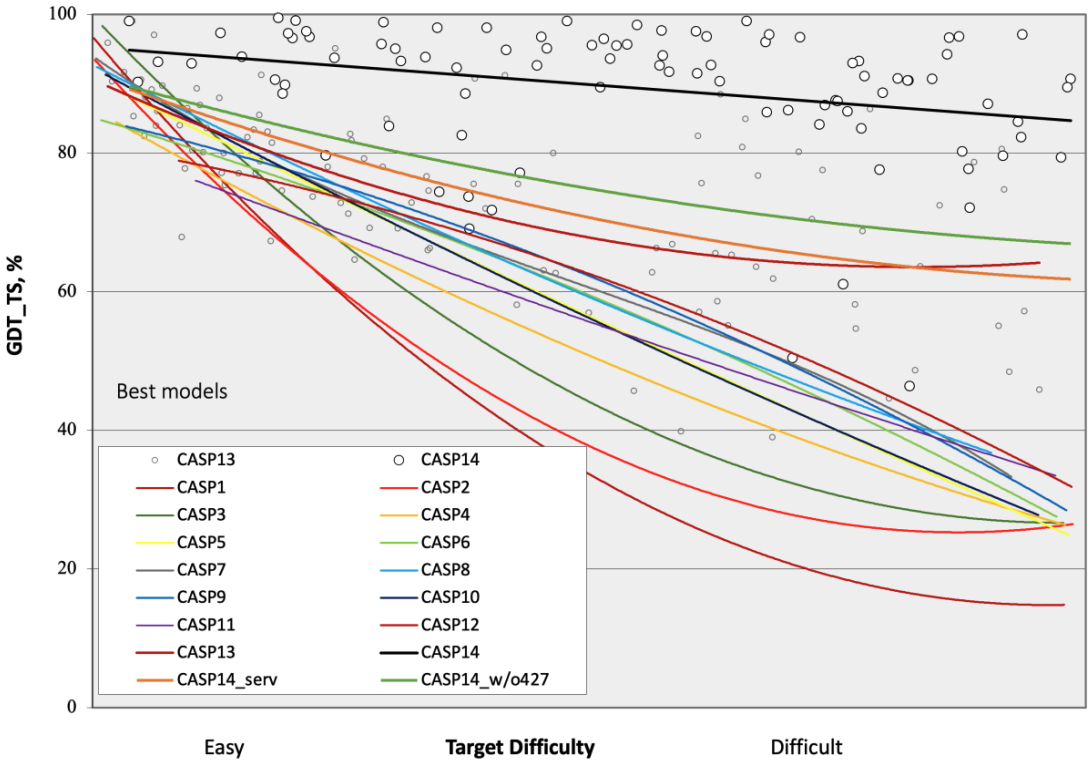


Figure 2.2: **28 years of progress between the first edition of CASP and CASP14** Evolution of the performance of the best method in each edition as a function of the difficulty of the predicted target. AlphaFold2's performance (CASP14) showed a big gap with the second best method ($CASP14\_w/o427$).

Source: John Moult's presentation at CASP14 `https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf`

Evoformer The second part is called Evoformer, inspired by the Transformer architecture described above. This part is made up of 48 blocks, each with its own weights. Each block has two inputs, the MSA representation and the pairwise representation. The output of each Evoformer block is an updated version of these two representations. In each Evoformer block, the MSA representation is updated by applying axial attention to the rows and columns of the MSA representation, enabling it to find the most informative relationships between the different sequences (rows) and amino acids (columns) of the MSA. This is followed by the MSA transition, which is analogous to the standard MLP of the transformer architecture. Pair representations are updated via triangle updates, which consist in learning updates that do not break the geometrical constraint of triangle equalities, and then via a self-attention mechanism. These two representations communicate within the Evoformer, the pair representation informs the row-wise attention of the MSA representation with bias, and the MSA representation updates the pair representation by adding its outer product. The Evoformer is the main part of the model, and contains 91M of the model's 93M parameters. The rest of the parameters belong to the structure module.

Structure module The structure module can be interpreted as a decoder that transforms Evoformer representations into three-dimensional structures. It contains 8 blocks, but unlike the Evoformer, its weights are shared. The structure module takes as input a linear projection of the first line of the MSA, the pair representation, and "backbone frames", which are local landmarks centered on alpha carbons and formed using the position of neighboring N and C. These backbone frames are initially all placed at the origin, without consideration of covalent bonds, and are then iteratively updated to form the protein. The key feature of the structure module is Invariant Point Attention, which makes the attention mechanism invariant by rotation and translation, which is important because these choices are arbitrary for a protein.

Flexibility, pLDDT and PAE In addition to the structure, AlphaFold2 outputs a confidence measure called the predicted Local Distance Difference Test (pLDDT). This is a score ranging from 0 to 100 assigned to each residue of the predicted structure. A score above 90 is considered very reliable, between 90 and 70 as reliable, between 70 and 50 as less reliable, and below 50 as very unreliable. It has been repeatedly established that regions with low pLDDT scores are often associated with flexible regions [178, 179, 180, 181]. Al Masri et al. [179] report a strong anti-correlation (Spearman coefficient > -0.70) between the pLDDT and B-factors for 10 out of the 15 kinases they studied. Saldaño et al. [178] found

a Pearson correlation of -0.44 between the pLDDT and the RMSF (Root Mean Squared Fluctuations) across a set of apo (non-binded) and holo (binded) proteins conformations. The RMSF is defined as the average fluctuation of an atom around its mean for a set of conformations:

$$\mathrm{RMSF}_i = \sqrt{\langle (r_i - \langle r_i \rangle)^2 \rangle}, \tag{2.8}$$

where r_i is the position of atom i and $\langle r_i \rangle$ is its average position over the different conformations observed. In order to study the dynamics via pLDDT, Guo et al. [180] defined the $\mathrm{AF2}_{score}$,

$$\mathrm{AF2}_{\mathrm{score}} = \frac{(\mathrm{pLDDT}_{\mathrm{max}} - \mathrm{pLDDT})}{(\mathrm{pLDDT}_{\mathrm{max}} - \mathrm{pLDDT}_{\mathrm{min}})}, \tag{2.9}$$

and found an average Pearson correlation of 0.85 between the $AF2$score and the RMSF derived from molecular dynamics simulations of four distinct proteins. However, other studies found only a very weak correlation (Pearson coefficient of -0.069) between the pLDDT and the associated experimental B-factors of 330 X-ray crystal structures of proteins [182]. This contradictory result could perhaps be partially explained by the author's use of the Pearson correlation, which yields large absolute values for linear correlation, whereas, based on previous works, the linear correlation of the pLDDT is more aligned with the RMSF, which is theoretically proportional to the square root of the B-factors via the following relationship:

$$B = \left(\frac{8\pi^2}{3}\right) \mathrm{RMSF}^2. \tag{2.10}$$

In addition to pLDDT, AlphaFold2 also outputs the Predicted Aligned Error (PAE), which estimates the error in the relative position of each pair of residues in the protein. Guo et al. [180] report that, generally, the PAEs associated with a pair of residues within the same domain are lower than the inter-domain PAEs, suggesting a link between PAE and the dynamics of the protein. Furthermore, they report a high correlation (PCC > 0.7) between the PAEs and the Cα distance variation maps derived from molecular dynamics.

Other approaches inspired by AlphaFold2, although often slightly less precise than the latter, have emerged, such as RoseTTAFold, which is also based on a similar principle of information exchange between a sequence representation and a pair representation [183]. Conceptually different approaches have also emerged, based on the use of a single sequence instead of MSAs. These methods rely on PLMs trained

in a self-supervised manner, i.e. without the objective of structure prediction. The hope is that homology information is learned independently by the PLM, rather than being explicitly constructed via the MSA. These include OmegaFold [184], HelixFold [185], MonoFold [186] and ESMFold [138]. In May 2024, AlphaFold3 was released [187]. It brings improvements in the prediction of protein-protein, protein-RNA and protein-ligand interactions. It includes architectural modifications of AlphaFold2. The Evoformer is simplified, with the MSA representation replaced by a single representation (analogous to ESMFold's Folding Block [138]) and renamed Pairformer. The structure module is replaced by a generative diffusion module trained to denoise noisy coordinates. These different methods tend towards the prediction of a single state. However, since proteins are dynamic objects, taking account of deformations and alternative conformations in prediction is an important future step, as emphasized by the AlphaFold3 authors: "*A key limitation of protein structure prediction models is that they typically predict static structures as seen in the PDB, not the dynamical behaviour of biomolecular systems in solution. This limitation persists for AF3, in which multiple random seeds for either the diffusion head or the overall network do not produce an approximation of the solution ensemble.*".

2.4 Deep learning methods for protein dynamics prediction

2.4.1 Prediction of isotropic flexibility

From sequence The prediction of RMSF, defined in Eq. (2.8), or B-factors is an initial approach to predicting a protein's dynamics. Since B-values vary greatly from protein to protein depending on the crystallographic experiment performed, this prediction generally involves predicting B_{norm}, the normalized B-factor,

$$B_{\text{norm}} = \frac{B - \langle B \rangle}{\sigma_B}, \tag{2.11}$$

where $\langle B \rangle$ is the average of the B-factors and σ_B is the standard deviation of the B-factors. The prediction of B_{norm} is an old problem; the first approach to predicting B_{norm} from the amino acid sequence dates back to 1985, when P.A. Karplus proposed a method based on a sliding window along the amino acid sequence to weight empirical B-factor values from the neighbors of each amino

acid [188]. Since this preliminary work, other approaches based on artificial neural networks have emerged to predict protein flexibility. Schlessinger and Rost [189, 190] developed PROFbval, which also uses a sliding window processed by two feed-forward networks, one of which is specific to buried residues. The predictions from these networks are made based on the amino acid sequence, evolutionary profiles, predicted secondary structure, and solvent accessibility.

In 2012, de Brevern et al. presented PredyFlexy [191], with an updated version in 2019 [192], a model based on Support Vector Machines (SVM) predicting structural flexibility from the sequence. This SVM was trained on experimental B-factors from 169 X-ray structures and the RMSFs from molecular dynamics simulations of these structures. This approach makes predictions per position, providing residue classification (flexible, intermediate, rigid), normalized B-factors, normalized RMSFs, and a confidence measure.

Yaseen et al., in 2016, proposed FLEXc [193], a feed-forward neural network consisting of 250 hidden nodes, applied to a window of 15 residues. The network predicts 3 classes (flexible, intermediate, rigid) for the central residue. The network was trained on a set of 5,547 proteins to predict B-factors from Position-Specific Scoring Matrices (PSSM), which represent the evolutionary conservation of amino acid positions in a protein sequence, along with various contextual scores and physical properties of the amino acids considered. They report performance improvements compared to PredyFlexy on the task of classifying different levels of flexibility.

In 2021, Vander Meersche et al. introduced MEDUSA [194], a deep convolutional neural network trained on B-factors from 9,880 high-resolution proteins from the PDB. The input features include PSSM, physical and chemical properties of amino acids, and one-hot encoding of the sequence, while the output is a flexibility class for each residue, with a selectable number of classes ranging from 2 to 5. The authors report improved performance in flexibility class classification compared to PROFbval. Other sequence-based approaches using bi-directional Long Short-Term Memory (biLSTM) networks [195, 196, 197] report very good correlations between predicted and experimental B-factors.

From structure To my knowledge, the only deep learning method for the prediction of B-factor that includes structural information is OPUS-BFactor, a method based on the use of ESM2 sequence embeddings and inter-residue backbone information [198]. We can also mention other non-deep learning methods that rely solely on representing protein structures as graphs, similar to the Gaussian Network Model described in section Section 1.4.1. The first is Multiscale Weighted

Colored Graphs (MWCGs) [199], which represent the protein as a graph, where each interaction is weighted by distance. The graph is "colored" to distinguish different types of interactions depending on the atom types involved in the bond. The flexibility of each atom is then inferred from a centrality measure based on the sum of weighted distances between the atom and other atoms. Scaramozzino et al. [200] also developed a purely structural approach based on ENM. They define the pairwise structural compliance, a measure of node displacement in the elastic network model in response to an applied force, divided by the magnitude of this force, and observe better correlations between this measure and B-factors compared to simple fluctuations from the ANM.

2.4.2 Prediction of conformational landscape

Conformational landscape from CryoEM CryoEM allows proteins to be captured in states close to their native environment. Although information about different conformations and their relative distribution is embedded in the poses, how this information is exploited depends on the methods used. Traditional approaches reconstruct a 3D volume by estimating the poses and handle conformational heterogeneity by performing discrete classification into a limited number of conformations [201, 202, 203]. These methods enable high-resolution reconstructions for multiple discrete states but may struggle when continuous heterogeneity is present in the sample. To address this, several methods have been proposed. Traditional approaches represent heterogeneity using a low-dimensional linear manifold in the space of volumes, relying on techniques like PCA [204, 205, 206] and normal mode analysis (NMA) [207, 208], or even combining NMA with atomic displacements from MD simulations [209, 210].

Deep learning approaches have aslo been proposed, such the variational autoencoder (VAE) CryoDRGN [211, 212], that learn a nonlinear manifold from the 2D images specific to each dataset. This low-dimensional manifold allows arbitrary sampling of the latent conformations, generating a continuous spectrum of conformations and creating a dynamic trajectory through 3D space. Other methods using VAEs to capture conformational heterogeneity have also been proposed [213, 214, 215, 216]. Similarly, 3DFlex [217], implemented in the cryoSPARC software, employs an autodecoder to predict a deformation field for the cryoEM map, further enabling the modeling of continuous conformational variation.

Conformational sampling with AF2 As previously mentioned, AlphaFold2 was trained to predict a single conformation. In fact, it has been observed that

in a set of 98 fold-switching proteins, AlphaFold2 captures only one of the two conformations in 94% of cases [218]. Many researchers have attempted to modify AlphaFold2 to enable it to predict alternative conformations of proteins. Most of the methods works by altering the input MSA. Alamo et al. [219], showed that reducing the number of sequences in the MSA used as input for AlphaFold2, coupled with reducing the number of recycles in the model from three to one, allowed AlphaFold2 to sample a wider variety of structures. The idea is that using a shallower MSA introduces uncertainty into the co-evolutionary signals, which increases conformational diversity. The authors demonstrated this approach by predicting alternative conformations of protein transporters and G Protein Coupled Receptors. Another approach, also involving modifications to the input MSAs, was explored by Stein and Mchaourab [220]. The authors call their approach in silico mutagenesis, which consists of artificially mutating entire columns of the MSA at contact zones predicted by AlphaFold2, which are believed to stabilize the structure. The authors illustrated their approach by predicting various conformations of adenylate kinase. Wayment-Steele et al. and Monteiro da Silva et al. [221, 222] propose two similar methods to predict different conformations, both based on subsampling the initial MSA, which involves creating subgroups within the MSA. Wayment-Steele et al. chose to form clusters based on sequence similarity, while Monteiro da Silva et al. proceeded with repeated random subsampling, allowing them to estimate the frequency distribution of different conformations. Another approach, which acts not on the MSA but on the network parameters, involves using dropout, a technique that randomly disables connections within the network. This method is typically employed during training to regularize learning but can also be applied during inference to introduce diversity into the predictions [223, 224]. Cfold [225], also proposes using sampling and clustering of the input MSA, but they retrain the network specifically with a split of alternative conformation to make sure that alternative conformations are not seen during training, making sure that the network is really able to predict alternative conformations by itself.

Generative models Denoising Diffusion Probabilistic Models (DDPMs) are generative models trained to denoise (inverse diffusion) random noise added (diffusion) to data [226]. This allows them to generate new data from noise alone. These models have been successfully applied to protein design, with approaches such as Chroma [227], RFDiffusion [228], Genie and Genie 2 [229, 230]. Building on DDPMs, Komorowska et al. proposed conditioning the generation of new conformations through normal mode analysis [231, 232]. EigenFold [233] is also a generative diffusion model that allows sampling a distribution of protein structures

with normal modes. The authors of EigenFold define harmonic diffusion, which, unlike adding Gaussian noise, incorporates the structural constraints of the protein by modeling it as a system of harmonic oscillators. This approach ensures biologically plausible structures by maintaining the constraints between adjacent residues, and it generates new conformations by progressively refining the structure along the eigenmodes of the system during the reverse diffusion process. Similarly, the Distributional Graphormer (DiG) framework introduced by Zheng et al. [234] also leverages diffusion models to sample equilibrium distributions of molecular systems. Flow matching, a generalization of diffusion models, learns continuous transformations between distributions without stepwise noise addition, and has been combined with Alphafold and ESMFold to capture conformational diversity [235]. Diffusion models offer great flexibility in generating new structures; however, they are prone to hallucination, and models trained on different protein families still fail to generate accurate ensembles [187].

Chapter 3

DANCE

This chapter is based on the scientific article *Explaining Conformational Diversity in Protein Families through Molecular Motions*, published on July 10, 2024, in Scientific Data and available open access at the following link: `https://www.nature.com/articles/s41597-024-03524-5`.

Explaining Conformational Diversity in Protein Families through Molecular Motions

Valentin Lombard[1], Sergei Grudinin[2*], Elodie Laine[1,3*]

[1]Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France
[2]Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[3]Institut Universitaire de France (IUF)

Corresponding author(s): Sergei Grudinin (sergei.grudinin@univ-grenoble-alpes.fr), Elodie Laine (elodie.laine@sorbonne-universite.fr)

Abstract

Proteins play a central role in biological processes, and understanding their conformational variability is crucial for unraveling their functional mechanisms. Recent advancements in high-throughput technologies have enhanced our knowledge of protein structures, yet predicting their multiple conformational states and motions remains challenging. This study introduces Dimensionality Analysis for protein Conformational Exploration (DANCE) for a systematic and comprehensive description of protein families conformational variability. DANCE accommodates both experimental and predicted structures. It is suitable for analysing anything from single proteins to superfamilies. Employing it, we clustered all experimentally resolved protein structures available in the Protein Data Bank into conformational collections and characterized them as sets of linear motions. The resource facilitates access and exploitation of the multiple states adopted by a protein and its homologs. Beyond descriptive analysis, we assessed classical dimensionality reduction techniques for sampling unseen states on a representative benchmark. This work improves our understanding of how proteins deform to perform their functions and opens ways to a standardised evaluation of methods designed to sample and generate protein conformations.

3.1 Introduction

Proteins orchestrate all biological processes, and their malfunctions often result in disease. In recent years, high-throughput technologies have greatly improved our knowledge of their amino acid sequences and 3D shapes [236, 175, 237, 5]. While reaching the single-structure frontier [174], these advances have also highlighted the complexities of how proteins move and deform to carry out their biological functions [238, 63]. They have stimulated a renewed interest in the modeling of protein and protein complex multiple conformational states [170]. In particular, the success of the protein structure prediction neural network AlphaFold2 [153] has inspired innovative strategies for modifying or repurposing it toward exploring protein conformational space. These approaches involve forced sampling [239], modulation of input multiple sequence alignment content and depth [221, 219], or guidance with state-annotated templates [240, 241]. Although they have achieved promising results for specific protein families, systematic assessments have revealed limitations [218, 242]. In addition, studies sampling from low-dimensional representations or manifolds learned from observed or simulated conformations [233, 234, 243] have underscored the difficulty in predicting new, completely unseen states and the importance of high-quality data for training or benchmarking.

Experimental techniques like X-ray crystallography, cryogenic-electron microscopy (cryo-EM), and nuclear magnetic resonance spectroscopy (NMR) are essential for capturing protein functional states [244, 238]. The Protein Data Bank (PDB) [5] offers access to multiple structural states for various proteins, solved independently in different conditions, oligomeric states, and with diverse cofactors and molecular partners. Researchers have actively engaged in efforts to collect, cluster, curate, represent, visualise, and functionally annotate these states [245, 244, 246, 247]. These endeavours have provided valuable insights into the biologically meaningful conformational space for specific protein families such as protein kinases [248], RAS isoforms [249], ABC (ATP Binding Cassette) transporters [250], and G-protein coupled receptors (GPCRs) [251]. However, producing or validating functional annotations for structural states involves a substantial amount of manual intervention. Despite the wealth of experimentally resolved protein conformational variability, its full exploitation remains an ongoing challenge.

Ideally, one would like to comprehensively describe protein conformational variability with low-dimensional representations or manifolds amenable to visualisation and interpretation. Principal Component Analysis (PCA) serves as a convenient and robust means to reduce the dimensionality of a dataset, capturing maximum

variability [252, 118]. The principal components extracted from a conformational ensemble define 3D directions for every atom, and motions along them allow navigating the conformational space [120]. PCA has proven useful for extracting structural transitions from sparse disconnected low-energy structural states [253, 254, 65, 124, 255, 256]. Unlike more complex non-linear dimensionality reduction techniques, it offers the advantage of not depending on numerous adjustable parameters and provides a straightforward geometrical interpretation.

Here, we describe a PDB-wide analysis of protein conformational variability across various levels of sequence homology. Our fully-automated computational pipeline, named Dimensionality Analysis for protein Conformational Exploration (DANCE), systematically compiles collections of aligned protein conformations and extracts their principal components. We interpret the representation space defined by the main principal components as the *linear motion manifold* underlying the observed conformations. We provide estimates of the intrinsic dimensionality of these motion manifolds. To assess generative methods, we introduce a benchmark set comprising ten conformational collections representing therapeutic targets with substantial functional transitions. Additionally, we provide baseline performances from classical linear and non-linear manifold learning techniques.

DANCE is versatile, handling both experimental and predicted structures with varying amino acid sequences. It adopts an unbiased approach, avoiding predetermined protein or domain definitions when building the conformational collections. Considering the complete context of input protein chains enables a thorough examination of inter-domain motions. Furthermore, DANCE accommodates uncertainty from unresolved protein regions without assuming potential conformations. It introduces a weighting scheme to mitigate the imbalanced coverage of variables.

We provide several databases of conformational collections representing the whole PDB as well as detailed information about the benchmark on Figshare [257]. In addition, DANCE's source code is available at: `https://github.com/PhyloSofS-Team/DANCE`.

3.2 Methods

3.2.1 Overview of DANCE

DANCE takes as input a set of protein 3D structures (in Crystallographic Information File or CIF format) and outputs a set of protein- or protein family-specific conformational collections or ensembles (in CIF of PDB format). It first clusters

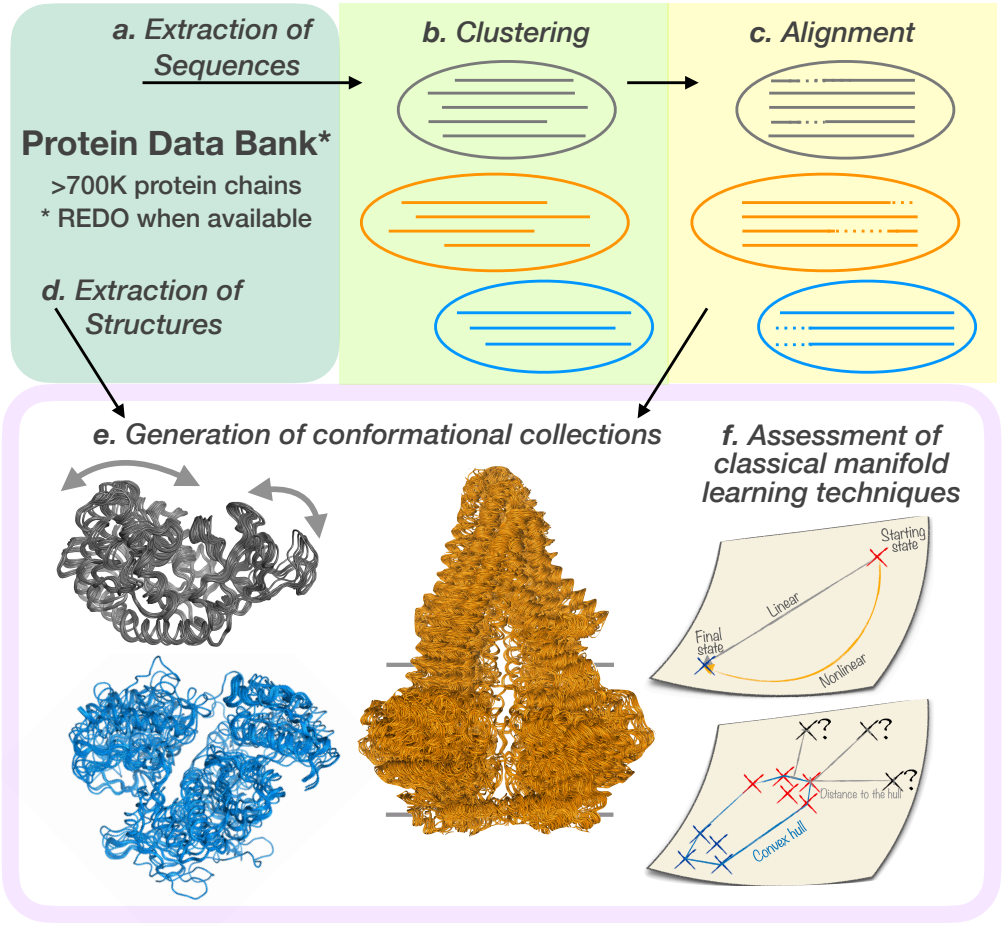


Figure 3.1: **Outline of the study.** Our approach, DANCE, exploits both amino acid sequences and 3D coordinates. We applied it to all experimentally determined protein-containing 3D structures from the PDB. Alternatively, users can provide a custom set of experimental structures or predicted models. DANCE first concentrates on sequences. It extracts them from the input structures (A) and clusters them with MMseqs2 based on user-defined similarity and coverage thresholds (B). For each cluster, It generates a multiple sequence alignment using MAFFT (C). It then extracts all 3D coordinates (D), groups the conformations according to the clusters identified in B and superimposes them to generate conformational ensembles (E). The superimposition aims at minimizing the Root Mean Square Deviation to a chosen reference, using the alignments produced by C for mapping the residues. The examples of the bacterial enzymes adenylate kinase (in grey, reference PDB code: 1AKEA) and MurD (in blue, 1E0DA), and the murine ABC transporter P-glycoprotein (5KOYB) are depicted. The arrows indicate adenylate kinase's main motion. The horizontal lines behind the P-glycoprotein indicate the boundaries for the membrane bilayer. Finally, DANCE summarises conformational diversity through Principal Component Analysis (F). We further assessed the ability of classical manifold learning techniques to reconstruct and extrapolate conformations.

and superimposes the input structures based on the similarities found in their corresponding amino acid sequences. The users can choose to analysis all input structures or only those representing monomeric biological units. DANCE then determines the set of principal components sufficient to explain the variability observed within each conformational ensemble. The algorithm unfolds in six main steps depicted in **Fig. 4.1**.

- **a- Extraction of sequences.** The first step extracts the one-letter amino acid sequences of all polypeptidic chains contained in the input CIF files. In case of multiple models, DANCE retains only the first one. The names of the residues with resolved 3D coordinates are taken from the *_atom_site.label_comp_id* column. Residues missing from the protein structure are included as lowercase letters in the sequence if they are defined in the *_entity_poly_seq* category. This information will help in clustering and aligning the sequences (see below). Otherwise, they are replaced by the "X" symbol. The "X" symbol is also used for unknown amino acid types and for modified amino acids without a close natural neighbour. Sequences comprising less than 5 non-"X" residues are then filtered out.

- **b- Clustering of the sequences.** DANCE clusters sequences using MMseqs2 [258]. The users can choose the desired levels of sequence similarity and coverage, both set to 80% by default. The coverage is bidirectional by default. This step outputs a TSV file specifying the clusters.

- **c- Multiple sequence alignments.** DANCE then aligns the sequences within each cluster using MAFFT [259] with default parameters and the BLOSUM62 substitution matrix [260]. It further removes all the columns containing only Xs or gaps, and reorders the sequences according to their PDB codes.

- **d- Extraction of structures.** DANCE extracts 3D coordinates of the backbone atoms N, C, Cα, and the O atom, of all polypeptidic chains contained in the input CIF files. It reconstructs missing O atoms based on the other atom's coordinates. It disregards residues with missing backbone atoms and chains shorter than 5 residues.

- **e- Generation of the conformational collections.** DANCE then uses the sequence clusters defined in (b) to group conformations and the residue matching provided by (c) to superimpose them. The superimposition puts

their centers of mass to zero and then aims at determining the optimal least-squares rotation matrix minimizing the Root Mean Square Deviation (RMSD) between any conformation and a reference conformation (see below). This is achieved through the ultrafast Quaternion Characteristic Polynomial method [261, 262]. The users can choose to account for all the atoms in the superimposition, or only the Cα atoms. Optionally, the users can filter out the conformations with too few (less than 5 by default) residues aligning to the reference. As a post-processing step, DANCE reduces structural redundancy. Namely, it removes any conformation A deviating by less than rms_{cut} Å from another one B, provided that the sequence of A is identical to or included in that of B. The value of rms_{cut} is 0.1 Å by default and is customizable by the users. Finally, DANCE saves the conformational ensemble as a multi-model file in PDB or CIF format. Notice that the models can display different amino acid sequences. DANCE also outputs the corresponding multiple sequence alignments (MSA) in FASTA format, and the matrix of all-to-all pairwise RMSDs.

- **f- Extraction of linear motions.** DANCE performs PCA on the 3D coordinates from each collection. This dimensionality reduction technique identifies orthogonal linear combinations of the variables, namely the Cartesian coordinates, maximally explaining their variance (see below). These linear combinations, which we refer to as principal components or PCA modes, represent directions in the 3D space for every atom. Deforming the protein structure using these components produce motions that connect the conformations observed in the collection. For the sake of simplicity, we directly refer to the principal components as to *linear motions*, although they may not represent actual physical motions undergone by the protein. Furthermore, we estimate the *intrinsic dimensionality* of the linear motion manifold underlying an ensemble's conformational variability as the number of principal component explaining essentially all its positional variance. The higher the dimensionality – the more complex the linear motions.

Choosing a reference

We choose the reference conformation for the superimposition as the one with the amino acid sequence most representative of the MSA. For this, we first determine the consensus sequence s^* by identifying the most frequent symbol at each position. We consider "X" symbols as equivalent to gaps. Hence, each position is described

by a 21-dimensional vector giving the frequencies of occurrence of the 20 amino acid types and of the gaps. In case of ambiguity, we prefer an amino acid over a gap, hence longer sequences over shorter ones, and an amino acid with a higher BLOSUM62 score over a lower-scored one. Then, we compute a score for each sequence s in the MSA reflecting its similarity to s^* and expressed as,

$$\text{score}(s) = \sum_{i=1}^{P} \sigma(s_i, s_i^*), \tag{3.1}$$

where P is the number of positions in the MSA and $\sigma(s_i, s_i^*)$ is the BLOSUM62 substitution score between the amino acid s_i at position i in sequence s and the consensus symbol s_i^* at position i. We set the gap score to $\min_{a,b}(\sigma(a,b)) - 1 = -5$.

Judging the quality of the MSA

We compute the identity level of an MSA as the average percentage of sequence pairs sharing the same amino acid in a column, and the coverage as the percentage of positions having less than 20% of gaps. In addition, we evaluate the global quality of the MSA with a sum-of-pairs score, with $\sigma_{match} = 1$ and $\sigma_{mismatch} = \sigma_{gap} = -0.5$. We normalise the raw sum-of-pairs scores by dividing them by the maximum expected values. The final score for an MSA is thus expressed as,

$$\text{score}_{rel}(MSA) = \frac{\text{score}(MSA)}{\binom{n}{2} L_{eff}}, \tag{3.2}$$

where is the raw MSA score, n is the number of chains or sequences, and L_{eff} is the effective length of the MSA, computed as,

$$L_{eff} = \max_{s \in \mathcal{S}} \sum_{i=1}^{L(s)} \mathbb{I}\{s_i \in \mathcal{A}\}, \tag{3.3}$$

where $\mathcal{S}$ is the set of sequences comprised in the MSA, $L(s)$ is the length of the aligned sequence s, and $\mathcal{A}$ is the 20-letter amino acid alphabet (*e.g.*, excluding gap characters).

Extracting linear motions

The Cartesian coordinates of each conformational ensemble can be stored in a matrix R of dimension $3m \times n$, where m is the number of positions in the associated

MSA and n is the number of conformations. Each position is represented by a C-α atom. We compute the covariance matrix as,

$$C = \frac{1}{n-1} R^c (R^c)^T = \frac{1}{n-1} (R - \bar{R})(R - \bar{R})^T, \tag{3.4}$$

where $\bar{R}$ is obtained by averaging the coordinates over the conformations. Alternatively, the users can choose to center the data on the reference conformation. The covariance matrix is a $3m \times 3m$ square matrix, symmetric and real.

The PCA consists in decomposing C as $C = VDV^T$ where V is a $3m \times 3m$ matrix where each column defines an eigenvector or a PCA mode that we interpret as a linear motion. D is a diagonal matrix containing the eigenvalues. The sum of the eigenvalues $\sum_{k=1}^{3m} \lambda_k$ amounts to the total positional variance of the ensemble. The portion of the total variance explained by the kth eigenvector or linear motion is estimated as $\frac{\lambda_k}{\sum_{k=1}^{3m} \lambda_k}$.

In addition, we estimate the collectivity [263, 264] of the kth eigenvector as,

$$\mathrm{coll}(\mathbf{v}_k) = \frac{1}{m} \exp\left(- \sum_{i=1}^{3m} v_{ki}^2 \log v_{ki}^2 \right). \tag{3.5}$$

If $\mathrm{coll}(\mathbf{v}_k) = 1$, then the corresponding motion is maximally collective and has all the atomic displacements identical. In case of an extremely localised motion, where only one single atom is affected, the collectivity is minimal and equals to $1/m$.

We also apply PCA to the correlation matrix computed by normalising the covariance matrix as,

$$\mathrm{Cor}_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i}}\sqrt{C_{j,j}}}. \tag{3.6}$$

In that case, the sum of the eigenvalues $\sum_{k=1}^{3m} \lambda_k$ amounts to 1.

Handling missing data

As stated above, the conformations in a collection may have different lengths reflected by the introduction of gaps in the associated MSA. We fill these gaps with the coordinates of the conformation used to center the data (average conformation, by default). In doing so, we avoid introducing biases through reconstruction of the missing coordinates. Moreover, this operation results in low variance for highly gapped positions, thus limiting their contribution to the extracted motions. To go further and explicitly account for data uncertainty, we implemented a weighting

scheme. Specifically, DANCE assigns confidence scores to the residues and include them in the structural alignment step and the PCA. The confidence score of a position i reflects its coverage in the MSA, $w_i = \frac{1}{n}\sum_S \mathbb{1}_{a_i^S \neq \text{"X"}}$, where "X" is the symbol used for gaps. The structural alignment of the jth conformation onto the reference conformation amounts to determining the optimal rotation that minimises the following function [265],

$$E = \frac{1}{\sum_i w_i} \sum_i w_i (r_{ij}^c - r_{i0}^c)^2, \tag{3.7}$$

where r_{ij}^c is the ith centred coordinate of the jth conformation and r_{i0}^c is the ith centred coordinate of the reference conformation. The resulting aligned coordinates are then multiplied by the confidence scores prior to the PCA.

Implementation details

We implemented DANCE in C/C++ and Python. It relies on the C++ GEMMI library [266] to parse the CIF files and manipulate the structures. It runs MMseqs2 through the following command: *cluster DB clusterDB tmp –cov-mode 0 -c $cov –min-seq-id $id*. It launches MAFFT with the options *auto*, *amino* and *preservecase*. The multiple sequence alignment and structure superimposition steps are parallelized. For the PCA, we use the singular value decomposition (SVD) implemented in NumPy [267] on the R matrix directly. SVD is computationally more advantageous when $3m \gg n$, which is typically the case of our data, since we only compute the required number of n components. We created structure visualisations in Pymol v2.5.0 [4].

3.2.2 Application and extension of DANCE

DANCE is applicable to experimental 3D structures as well as predicted 3D models, as long as they comply with the CIF standards.

Describing conformational variability over the whole PDB

We applied DANCE to all 748 297 protein chains with experimentally resolved 3D structures available in the PDB, as of June 2023. We downloaded all the PDB entries in CIF format from the RCSB [268]. We replaced the raw CIF files with their updated and optimised versions from PDB-REDO whenever possible [269]. It took about 2.25 hours to run DANCE on the whole PDB on a desktop

computer with Intel Xeon W-2245 @ 3.90GHz and 32Go of RAM (**Table A.1**). The most time consuming steps are the extraction and superimposition of the 3D structures to create the conformational ensembles. We ran DANCE at eight different levels of sequence similarity, designated as l_{cov}^{id}, where *id* and *cov* are the sequence identity and coverage thresholds, correspondingly, and range from 50 to 80%. For investigating how the ensembles transformed across levels, we focused on the 18 616 conformational ensembles detected in the most relaxed set up, namely at 30% identity and 50% coverage (l_{50}^{30}). For each ensemble, we extracted its reference protein chain and we traced back the conformational ensembles to which it belonged upon progressively applying stricter thresholds.

Focusing on the ABC superfamily

We extended DANCE usage beyond the single-chain and sequence-similarity paradigms to describe the conformational variability of ABC (ATP Binding Cassette) transporters. We retrieved a set of 354 ABC protein experimental 3D structures from `https://abc3d.hegelab.org` [250]. They correspond to functionally relevant states annotated as biological units in the PDB. In most of these structures, several polypeptidic chains, typically 2 or 4, encode the two nucleotide-binding domains (NBDs) and two transmembrane domains (TMDs) of the ABC architecture. In addition, some structures contain several ABC protein copies or some ABC protein cellular partners (small molecules, substrate peptides, interacting proteins). We chose the murine ABC transporter P-glycoprotein (5KOYA) as reference for the subsequent analysis. Its 1182-residue long single polypeptidic chain the full-length transporter architecture.

To cope with the high sequence divergence of the ABC superfamily, we relied on structural similarity for grouping and matching the ABC conformations. Specifically, we used the method Foldseek [163] to identity structures sharing significant similarity with the reference and align them. We performed a first screen by querying the reference against all individual chains (1 244 in total) and defined significant hits as those with an e-value lower than 10.0. Then, for each structure, we estimated an upper bound on its coverage of the reference by summing up the reference residue ranges appearing in the alignments associated with its significant hits. We filtered out the structures with coverage upper bounds lower than 90%. We performed a second screen by querying the reference against the 209 remaining structures defined as monomers by concatenating their chains. We identified two structures (5NIK, 5NIL) spanning less than 90% of the reference. Permuting their chains

did not increase their coverage and thus we removed them. To further detect potentially suboptimal chain orderings, we computed reference to target residue span ratios. We identified one structure, namely 7AHD, with a highly imbalanced ratio of 1.6. Such a high value is indicative of large parts of the reference that could not be aligned to the target structure. Permuting the four chains (A,B,C,D) of 7AHD into (A,D,B,C) led to a more balanced ratio of 0.86. We did not observe discrepancies for other structures and thus we retained their original chain ordering. Finally, we removed the structures with low-quality alignments, *i.e.*, with more than 200 gaps or with a continuous gapped region of more than 60 positions.

Among the 195 structures finally selected, 4F4C, 7SHN and 7AHD contained unknown or unrecognized amino acids which we removed. We ran Foldseek one more time to generate a structure similarity-based multiple sequence alignment centred on the reference 5KOYA. We trimmed the alignment and the 3D structures by removing the residues inserted with respect to the reference. We gave the trimmed alignment and 3D coordinate files as input to DANCE, starting directly from step *d* (see the overview of DANCE algorithm above). For consistency and comparison purposes, we asked DANCE to center the data on the reference. To mitigate the impact of potential alignment errors, we applied weights reflecting position-specific confidence scores (see above, *Handling missing data*). DANCE structural redundancy reduction step removed 7 conformations, resulting in an ensemble of 188 conformations.

We compared this ensemble with those generated by DANCE default sequence similarity-based end-to-end procedure applied to the whole PDB. More specifically, we took the ensembles generated at l_{80}^{80} and l_{50}^{30} and containing 5KOYA and we rebuilt them with DANCE, applying the 5KOYA centering and the uncertainty weighting scheme. We estimated the similarity between the ensembles' motion subspaces as the Root Mean Square Inner Product (RMSIP) [270, 126]. The latter measures the overlap between all pairs of the l first PCA modes and is defined as,

$$\mathrm{RMSIP} = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{l}(\mathbf{v}_i^{\mathcal{S}_\mathcal{A}}.\mathbf{v}_j^{\mathcal{S}_\mathcal{B}})^2}, \tag{3.8}$$

where $\mathbf{v}_i^{\mathcal{S}_\mathcal{A}}$ and $\mathbf{v}_j^{\mathcal{S}_\mathcal{B}}$ are the ith and jth PCA modes extracted from the conformational ensembles $\mathcal{S}_\mathcal{A}$ and $\mathcal{S}_\mathcal{B}$, and l is the number of modes considered for the comparison. Moreover, we monitored the distance between the geometric centres of the two NBDs defined by the C-α atoms of residues numbered 346-596 and 929-1182, respectively, in the reference 5KOYA.

3.2.3 Benchmarking for the generation of unseen conformations

We further investigated whether the extracted linear principal components could be useful to predict unseen conformations. Moreover, since the manifold underlying our data is *a priori* non-linear, we tested whether non-linear methods could achieve better reconstructions than linear PCA. We focused on the widely used kernel Principal Component Analysis (kPCA) [271, 272] and the uniform manifold approximation and projection (UMAP) [273].

Dimension reduction with non-linear kernel PCA

The intuition behind kPCA is to map the input data points to a higher dimensional space where they will be linearly separable by a classical PCA. The mapping function $\phi : \mathbb{R}^{3m} \to \mathbb{R}^M$ is not known. Instead of explicitly calculating it, we use a kernel function $k(\mathbf{r}_i, \mathbf{r}_j) = \phi(\mathbf{r}_i)^T \phi(\mathbf{r}_j)$, where $\mathbf{r}_i$ and $\mathbf{r}_j$ are two conformations. We considered three commonly used kernels,

- the polynomial kernel $k(\mathbf{r}_i, \mathbf{r}_j) = \left(\frac{1}{2\sigma^2}\mathbf{r}_i\mathbf{r}_j^T + c\right)^d$, where $c = 1$ and $d = 3$ by default,
- the sigmoid kernel $k(\mathbf{r}_i, \mathbf{r}_j) = \tanh\left(\frac{1}{2\sigma^2}\mathbf{r}_i\mathbf{r}_j^T + c\right)$, where $c = 1$ by default,
- and the radial basis function (RBF) or Gaussian kernel $k(\mathbf{r}_i, \mathbf{r}_j) = \exp\left(-\frac{d(\mathbf{r}_i,\mathbf{r}_j)^2}{2\sigma^2}\right)$, where $d(\mathbf{r}_i, \mathbf{r}_j)$ is the Euclidean distance between the two conformations $\mathbf{r}_i$ and $\mathbf{r}_j$.

We explored different values of the hyperparameter σ. For sufficiently large values, *i.e.*, $\frac{1}{2\sigma^2}\mathbf{r}_i\mathbf{r}_j^T \ll 1$ or $\frac{1}{2\sigma^2}d(\mathbf{r}_i, \mathbf{r}_j)^2 \ll 1$, the kernel becomes effectively linear.

Thus, given the input coordinates R representing n conformations, we computed the corresponding kernel matrix K of dimension $n \times n$ and decomposed it using the classical PCA. The resulting principal components $\{\boldsymbol{\nu_1}, \boldsymbol{\nu_2}, ..., \boldsymbol{\nu_n}\}$ can then be expressed as,

$$\boldsymbol{\nu_j} = \sum_{i=1}^{n} a_{ji}\phi(\mathbf{r}_i), \text{where } a_{ji} = \frac{1}{\lambda_j(n-1)}\phi(\mathbf{r}_i)^T\boldsymbol{\nu_j}. \tag{3.9}$$

Uniform manifold approximation and projection

The UMAP algorithm first builds a graph representing the data in the ambient space, and then determines the most similar graph in a lower dimension. It relies on the assumptions that there exists a low-dimensional manifold on which the original data would be uniformly distributed and that this manifold is locally connected. Under such assumptions, any ball of fixed volume on the low-dimensional manifold should contain approximately the same number of points. Thus, to build the graph, UMAP defines balls in the ambient space centred at each point and encompassing its n_{neigh} nearest neighbours. The balls have variable sizes that reflect the topology of the dataset in the ambient space. UMAP then connects points whose corresponding balls overlap and computes the edge weights by combining the balls' radii. The resulting graphical representation is projected into a lower-dimensional space by minimising the cross entropy between the high- and low-dimensional graphs, which can be viewed as a force-directed graph layout algorithm. We explored two hyperparameters, namely the number of neighbours n_{neigh} controlling the balls' radii and the minimum distance d_{min} apart that points are allowed to be in the low dimensional representation. Low values of n_{neigh} will make UMAP focus on local details of the dataset topology while high values will account for more global properties. Increasing d_{min} will push points far from each other in the representation space.

Generating conformations

For linear PCA, generating 3D conformations by combining the principal components is straightforward. More specifically, given a set of l PCA modes computed from the coordinates R, we generate a new conformation $\mathbf{r}^*_{\mathbf{pred}}$ as,

$$\mathbf{r}^*_{\mathbf{pred}} = \mathbf{p}^* V_l^T + \bar{\mathbf{r}}, \tag{3.10}$$

where the matrix $V_k \in \mathbb{R}^{3m \times l}$ contains the modes, $\bar{\mathbf{r}} \in \mathbb{R}^{3m}$ is the average conformation, and $\mathbf{p}^* \in \mathbb{R}^l$ is a point in the l-dimensional representation space defined by the modes. The coordinates of $\mathbf{p}^*$ specify the amplitudes of the modes.

For kPCA and UMAP, we need to learn an inverse transform function that maps points in the l-dimensional representation space defined by the components back to the input space. This problem is known as the *pre-image problem*. To solve it for kPCA, we used kernel ridge regression of the input coordinates R on their low-dimensional projections in the representation space as described in [274, 275] and implemented in the scikit-learn Python library [276]. The contribution of the L2-norm regularisation is controlled through the hyperparameter α. More

technically, α connects the squared L2-norm between a point in the representation space and its reconstruction with the squared L2-norm of the kernel weights used for the reconstruction. In the case of UMAP, we used the built-in inverse_transform function [273]. It relies on stochastic gradient descent to minimise the cross entropy between the low-dimensional graph and its high-dimensional pre-image graph.

Leave-one-cluster-out cross-validation procedure

We assessed the predictive performance of PCA and kPCA with a *leave-one-out* cross-validation procedure. Since the conformations are not evenly distributed within an ensemble, we grouped them into clusters prior to the evaluation. We performed the clustering in the l-dimensional PCA representation space, where l is the minimal number of linear components sufficient to explain 90% of the ensemble's total positional variance. We used the k-means clustering [277] with $k = l + 2$.

Given a clustered ensemble, we systematically tested the ability of the principal modes inferred from $l + 1$ clusters to predict the conformations belonging to the held-out cluster. We reconstructed each test conformation $\mathbf{r}^*$ from its projection $\mathbf{p}^*$ in the l-dimensional representation space. For the classical PCA, we computed the projection as,

$$\mathbf{p}^* = (\mathbf{r}^* - \bar{\mathbf{r}})V_l. \tag{3.11}$$

For the kPCA, the projection onto the principal component $\nu_{\mathbf{j}}$ is expressed as,

$$\phi(\mathbf{r}^*)\nu_{\mathbf{j}} = \sum_{i=1}^{n} a_{ji}\phi(R)^T\phi(\mathbf{r}^*) = \sum_{i=1}^{n} a_{ji}K(R, \mathbf{r}^*). \tag{3.12}$$

We evaluated the reconstruction error as the RMSD between the predicted conformation $\mathbf{r}^*_{\mathbf{pred}}$ and the original conformation $\mathbf{r}^*$.

Distance to the training set

We estimated the difficulty of reconstructing a given conformation by computing its distance to the convex hull defined by the conformations used for training in the l-dimensional representation space. Setting the number of clusters in the training set to $l + 1$ ensures that the convex hull will be a polytope of dimension at least l. For instance, in 1 dimension, we need at least 2 affine-independent points to define a 1-polytope. The explicit computation of the convex hull of n points in l dimensions is an operation whose complexity is of the order of $O(n^{l/2})$

[278] and rapidly becomes computationally infeasible as the value of l increases. Nevertheless, the calculation of the distance of a given point to the hull does not require computing the convex hull explicitly and is a much simpler computational problem. It can be solved in quasilinear time with quadratic programming (QP). Here, we used the efficient and exact QP simplex solver proposed in [279] and implemented in the Computational Geometry Algorithms Library (CGAL) [280]. It takes advantage of the low dimensionality of the representation space by observing that the closest features of two l-polytopes are always determined by at most $l + 2$ points.

In order to compare distances across systems of different sizes, we scale them by the number of positions m,

$$d^{norm} = \frac{d}{\sqrt{m}}. \tag{3.13}$$

This normalisation also allows relating distances in the representation space with RMS deviations in the 3D Cartesian space. Indeed, let us consider an ensemble of conformations exhibiting a purely one-dimensional motion. Any two conformations distant by an RMSD of 1 Å in the original 3D space will be separated by a normalised distance of 1 Å in the one-dimensional representation space.

Interpolating between states

We generated interpolation trajectories between ATPase states with PCA and kPCA. We started from the conformational clusters defined in the leave-one-out procedure and identified clusters 0 and 4 as the most extreme ones along the first PCA component. Secondly, we used these two clusters only to learn PCA and kPCA low-dimensional representation spaces. We computed the coordinates of the clusters' centres in these spaces and defined interpolation trajectories between them with 50 regularly spaced intermediate points. We then generated 50 conformations from the 50 intermediate points. We finally determined the minimal RMS deviation between each generated conformation and the known conformations from clusters 1, 2 and 3. We qualitatively compared these trajectories with physics-based non-linear trajectories computed with NOLB [113]. NOLB extracts normal modes from a starting conformation and models the transition to a target conformation as a series of twists extrapolated from these modes with optimal amplitudes, as described in [281]. We chose 1KJUA from cluster 0 as the starting conformation and 1T5SA from cluster 4 as the target conformation.

3.3 Results

We used DANCE to chart the experimentally resolved conformational diversity of protein families (**Fig. 4.1**). We explored eight levels of sequence similarity (*sim*) and coverage (*cov*), denoted as l_{cov}^{sim}, to group the ∼750K chains included in the PDB as of June 2023 (**Fig. A.1A** and **Table A.2**). In the most conservative set up, namely l_{80}^{80}, less than 3% of the conformations remain isolated (**Fig. A.1A**, *singletons*). Most of the conformational collections (or ensembles) are associated with multiple sequence alignments of high quality across all levels (**Fig. A.1B**). Sequence identity and coverage are more widely distributed in more relaxed conditions, but the median values always remain very high, above 0.95 (**Fig. A.1C-D**).

3.3.1 Experimentally resolved conformations lie on low-dimensional manifolds

Only one or two linear principal components suffice to explain almost half of the ensembles' conformational diversity (**Fig. 3.2a**). We interpret these components as directions of motion, and by simplification, we will denote them as linear motions in the following (see *Methods*). In the overwhelming majority of cases, less than eight linear motions explain more than 90% of the total positional variance. These observations hold true across all sequence identity and coverage levels. They indicate that the conformational states captured by experimental techniques for a protein or a protein family lie on a low-dimensional manifold. This low dimensionality is only partially determined by the cardinality of the ensembles (**Fig. A.2A-B**). Almost 30% of the most highly populated ensembles (>50 conformations) detected at l_{80}^{80} can be comprehensively described with less than three linear motions (**Fig. A.2C**). This proportion increases up to 46% in the most relaxed conditions, namely at l_{50}^{30} (**Fig. A.2D**).

The bacterial adenylate kinase gives an example of a one-dimensional motion underlying its 42 conformations (**Fig. 4.1e**, in grey). One can easily classify the conformations by visual inspection into two main states, open and closed, deviating by about 7 Å. The bacterial enzyme MurD (**Fig. 4.1e**, in blue) and the murine ABC transporter P-glycoprotein (**Fig. 4.1e**, in orange) also exhibit low-dimensional opening-closing motions. In particular, the P-glycoprotein's collection reveals a rich spectrum of intermediate conformations between the open and closed forms (**Fig. 4.1e**, in orange). The main motion involves about 70% of the protein and modulates the volume of the transporter's internal cavity within the lipid bilayer

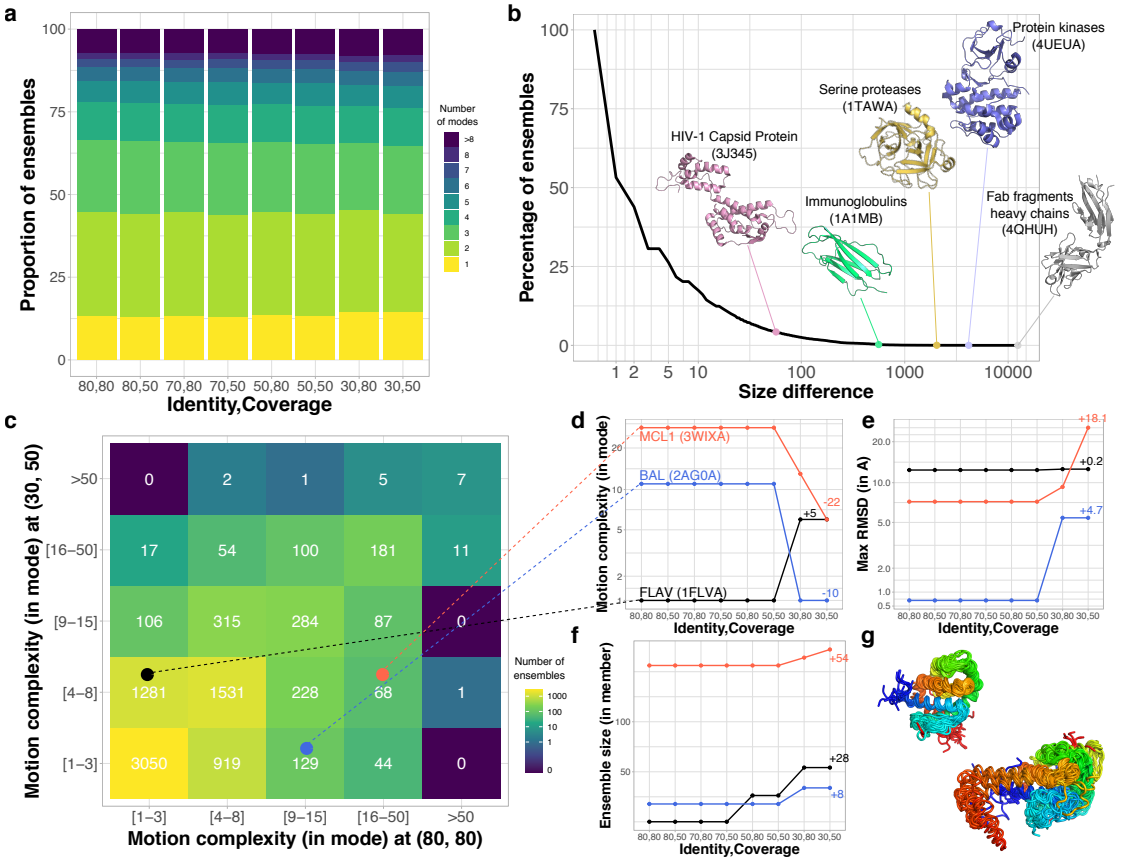


Figure 3.2: **Evolution of protein conformational diversity across sequence similarity levels. a.** Proportion of conformational ensembles requiring n linear PCA modes to explain 90% of their total positional variance, with n varying from 1 to 8. The number of modes n is an indicator of motion complexity. Singletons and pairs are excluded. **b.** Cumulative distribution of the number of conformations gained from the most stringent level, namely l_{80}^{80}, with 80% sequence similarity and coverage, to the most permissive one, l_{50}^{30}, with 30% similarity and 50% coverage. The 3D structures of the reference protein chains are depicted for a few ensembles. **c.** Comparison of motion complexity between the most stringent and most relaxed set ups. We considered only the cases where the ensemble at l_{50}^{30} is bigger than the corresponding one at l_{80}^{80}. Singletons and pairs are excluded. **D-G.** Detailed evolution of three ensembles marked by colored dots in panel C. **d.** Motion complexity expressed as a number of modes. The names and PDB codes of the reference chains are indicated. **e.** Motion amplitude, measured as the maximum RMSD between any two conformations (in Å). **f.** Conformational collection size. **G.** Conformational diversity observed for the Bcl-2 family. On the top left, the 54 conformations comprised in the MCL1 ensemble at l_{80}^{80}. At the bottom right, the 218 additional conformations at l_{50}^{30}. The color code indicates the position in the sequence, from the N-terminus in blue to the C-terminus in red.

up to over 6,000 Å^3 [282]. It explains about 80% of the total positional variance on its own. The remaining variability is mostly due to rotations of the nucleotide binding domains with respect to the transmembrane helical bundles and to loop deformations.

3.3.2 A few protein families display huge conformational expansion upon relaxing the sequence selection criteria

To investigate how the conformational ensembles transformed with sequence similarity, we systematically backtracked the 18 616 representative protein chains identified at l_{50}^{30} across more stringent levels (see *Methods*). The fragment antigen-binding regions display the largest growth between the most stringent and most relaxed sequence selection criteria (**Fig. 3.2**). For instance, while the Fab6785 light chain's ensemble at l_{80}^{80} comprises a bit less than 300 conformations, it expands up to over 12 500 conformations at l_{50}^{30} (**Fig. 3.2b**, PDB id: 4QHUH). With the largest number of conformations at l_{80}^{80}, the HIV-1 capsid protein's ensemble however displays a relatively limited expansion across the different levels, from 3 334 to 3 391 (**Fig. 3.2b**, 3J345). Bovine trypsin and its close homologs give an example of an extensively characterized subfamily, with 470 different conformations detected at l_{80}^{80}. This ensemble expands by more than 5 folds, aggregating different serine proteases, upon relaxing the criteria to l_{50}^{30} (**Fig. 3.2b**, PDB id: 1TAWA). Likewise, the Beta-2-microglobulin and its close homologs have a large body of 1 465 conformations at l_{80}^{80}, growing further up to 2 025 conformations at l_{50}^{30} by including other immunoglobulins (**Fig. 3.2b**, 7MX4B). By contrast, the reconstructed ancestral tyrosine kinase AS, a common ancestor of Src and Abl, has only 2 conformations available in the PDB and no close homologs. At l_{50}^{30}, it serves as representative for a huge ensemble of over 4 000 protein kinase conformations (**Fig. 3.2b**, 4UEUA). Apart from these over-represented protein families or superfamilies, the ensembles generally gain only a few conformations, with a median value of 4.

3.3.3 Family expansion may lead to an apparent motion simplification

As an ensemble grows, the gained conformations may lie on the same motion manifold, defined by the subset of principal components explaining the variance, or give rise to new motions represented by new components (**Fig. 3.2c**). The

bacterial long-chain flavodoxin exemplifies the second scenario (**Fig. 3.2d-f**, in black). At l_{80}^{80}, it undergoes a one-dimensional motion describing the transition between a compact state and a partially unfolded conformation (**Fig. A.3**). Upon relaxing sequence similarity to l_{50}^{30}, the ensemble roughly doubles in size (**Fig. 3.2f**) and the newly added conformations exhibit complex deformations of the FMN binding pocket. As a result, five more linear motions are required to explain the positional variance (**Fig. 3.2d**). Hence, in this case, the motions get more complex when considering more distant homologs.

The emergence of new motions does not however systematically lead to an increased motion complexity. The murine MCL1 gives an illustrative example of apparent motion simplification upon expansion (**Fig. 3.2d-f**, in red, and **Fig. 3.2g**). At l_{80}^{80}, almost 30 components are needed to explain the variability observed over the couple of hundreds conformations in the ensemble. They represent local deformations of the inter-helical loops and the extremities (**Fig. 3.2g** and **Fig. A.3**). Extending the ensemble to distant members of the Bcl-2 family brings in about 50 new conformations (**Fig. 3.2f**). They reveal a new extended state the protein BAX adopts upon assembling into domain-swapped dimers [283]. The large amplitude transition between the compact conformation and the extended one takes a big part in the variance, resulting in a drastically reduced motion complexity (**Fig. 3.2d**). The benzaldehyde lyase BAL gives another example (**Fig. 3.2d-f**, in blue) where the transition to a new state, adopted by the distant homolog actinobacterial 2-hydroxyacyl-CoA lyase [284], dominates the variance (**Fig. A.3**). The conformational variability transforms from small (<1Å) seemingly random fluctuations to a one-dimensional motion.

Overall, about a third of the ensembles undergo an apparent motion simplification upon expansion (**Fig. 3.2c** and **Fig. A.4A**). They likely represent protein families where distant homologs exhibit novel distinct states. The larger the deviations of these novel states with respect to the other ones, the higher the contribution of the corresponding motions to the variance. To mitigate this variance-dependent effect, we repeated the analysis on the correlation matrix. The latter estimates the extent to which the residues move in the same direction, regardless of the magnitude of their displacements. We found that the motion complexity still decreases in over 20% of the ensembles (**Fig. A.4B**). This result indicates that motion simplification does not merely reflect larger transitions "hiding" smaller rearrangements. A substantial fraction of protein families show evidence of more concerted residue movements between more distant homologs.

3.3.4 Beyond single chains and sequence similarity, the ABC superfamily as a case study

We explored the possibility of using DANCE to chart the conformational variability of remote homologs with low sequence similarity and variable chain composition. We focused on the ABC (ATP Binding Cassette) transporter superfamily. The ABC architecture comprises two nucleotide-binding domains (NBDs) and two transmembrane domains (TMDs) encoded by one or several polypeptidic chains (**Fig. 3.3a**). The NBDs are highly conserved across species and families, whereas the TMDs exhibit various scaffolds associated with heterogeneous transport functions [250]. We considered a collection of a few hundreds ABC protein experimental 3D structures [250], taking the single-chain murine P-glycoprotein as reference (**Fig. 3.3a**, 5KOYA).

We bypassed DANCE sequence extraction, clustering and alignment steps and directly gave it a pre-computed alignment built from structural similarities as input (see *Methods*). Relying on structure rather than sequence similarity and considering various oligomeric states provided a more comprehensive description of ABC transporters' functional motions and states (**Fig. 3.3** and **Movies S1-2**). The resulting ensemble comprises 188 conformations encompassing 295 protein chains, some of which have sequence identity below 30% or coverage lower than 50% (**Fig. 3.3a**). A set of 25 linear motions are required to explain the positional variance. By comparison, the sequence similarity-based 5KOYA-containing collection generated by DANCE at l_{50}^{30} contains only 71 conformations explained by only four linear motions. These motions are essentially identical to those extracted from the 61 conformations at l_{80}^{80} (**Fig. 3.3b**, RMSIP = 0.99).

Despite having different motion complexities, the sequence- and structure-based conformational collections have largely overlapping motion subspaces (**Fig. 3.3b**, RMSIP $\sim$ 0.7). In particular, they all share the same most contributing motion describing the transition between the transporter inward-closed and inward-open forms (**Fig. A.5**). This functional transition controls the substrate access to the transporter's central binding pocket. It explains 45 to 70% of the variance on its own and involves over two-thirds of the residues. The structure similarity-based collection represents a quasi-continuum of increasingly open states (**Fig. 3.3c**, in blue, and **Movie S1**) between two extreme dimeric forms, one from the human lysosomal cobalamin exporter ABCD4 where the two NBDs are in contact and the other from *Salmonella typhimurium*'s lipid A transporter MsbA with a widely open cavity. The overwhelming majority of conformations are regularly spaced

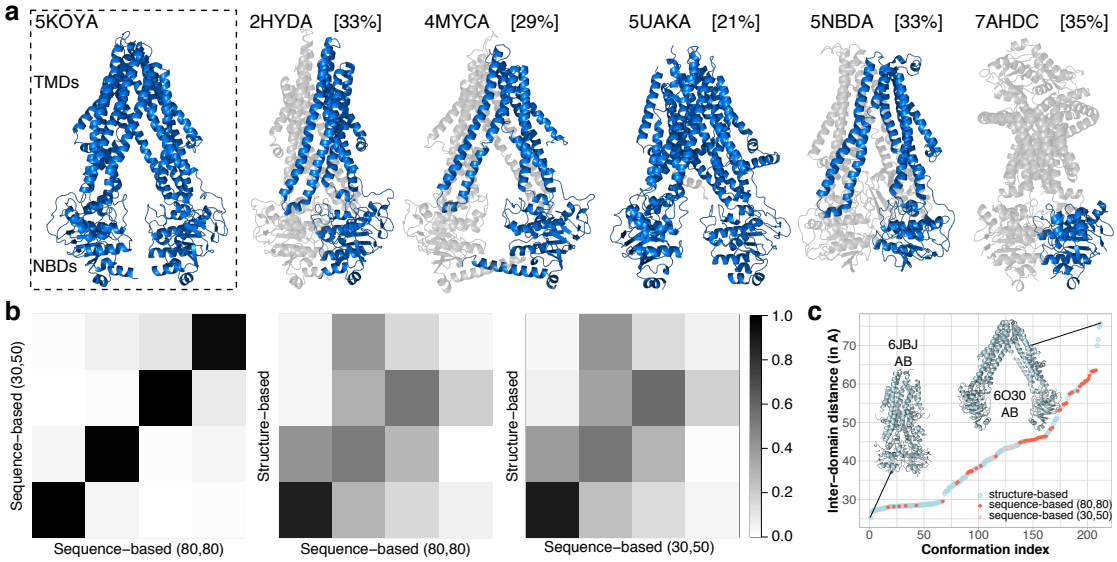


Figure 3.3: **ABC transporters' conformational variability. a.** Examples of protein structures from the ABC structure similarity-based conformational collection. The reference chain (5KOYA) is on the left, where we indicate the location of the two NBDs (∼500 residues) and two TMDs (∼700 residues). Within each of the other structures, we highlight one chain in marine, give its percentage of identity with the reference in squared brackets, and display the remaining chains in transparent grey. The six marine chains were assigned to six different collections by DANCE's default sequence similarity-based end-to-end protocol at l_{50}^{30}. **b.** Comparison of motion subspaces extracted from the sequence-based ensembles at l_{80}^{80} (61 conformations) and l_{50}^{30} (71 conformations) and the structure-based one (188 conformations). Each matrix shows the absolute pairwise scalar products computed for the first four PCA modes. The corresponding RMSIP are 0.99, 0.71 and 0.73. **c.** Distance between the geometric centres of the two NBDs (in Å). The conformations are ordered along the x-axis from the most closed one to the most open one.

by inter-NBD distance increments smaller than 1 Å. By contrast, the sequence similarity-based collections populate sparse regions of this continuous transition, with a high concentration of semi-open and open states (**Fig. 3.3c**, in pink and red, and **Movie S2**).

3.3.5 Classical manifold learning techniques can generate highly accurate conformations

Beyond describing the observed conformational variability, we evaluated the ability of several popular manifold learning techniques to generate unseen conformations. To do so, we identified a set of ten conformational ensembles with very different degrees of motion complexity (**Fig. 3.4a** and **Table A.3**). They comprise between 20 and over 3 300 conformations and their reference chains contain 80 to 1 200 residues. They represent proteins or protein families displaying substantial ($\geq$ 5 Å) and functionally relevant conformational changes, namely adenylate kinase (ADK) [285, 286], MurD [287, 243], the calcium pump ATPase [288, 289], the ABC transporters [290, 250], the small heat shock protein αB crystallin (Crys) [291, 292], the heat shock protein HSP90 [293, 294], calmodulin (CALM) [295, 296], kinases (KIN) [297, 298], RAS [299, 249], and the HIV capsid protein (CAP) [300, 301]. Most of them have been extensively characterized by experimental structure determination techniques or computational methods for simulating protein dynamics. Targeting their motions or their specific conformations bears a therapeutic interest.

We chose the linear PCA as baseline and we considered four non-linear techniques, namely kernel PCA (kPCA) [271, 272], UMAP [273], isoMAP [302] and t-SNE [303]. While all techniques allow for projecting the conformations in a low-dimensional space, only PCA, kPCA and UMAP allow for reconstructing conformations from the projections through an inverse transform. Furthermore, UMAP is limited to a narrow range of dimensions and, as a consequence, we could apply it only to a subset of the benchmark (see *Methods*). Hence, we primarily focus on the comparison between PCA and kPCA in the following. We tested three different kernels for kPCA, namely the sigmoid, polynomial and radial basis function (RBF) kernels. Within each ensemble, we first learned low-dimensional representations of a subset of conformations used as training samples. We then projected the test conformations, not seen during training, to the learned representation space, and mapped the projections back to the original 3D Cartesian space. The mapping is determined analytically in the case of linear PCA and learned in the case of kPCA and UMAP (see *Methods*). We evaluated the quality of the 3D reconstructions by

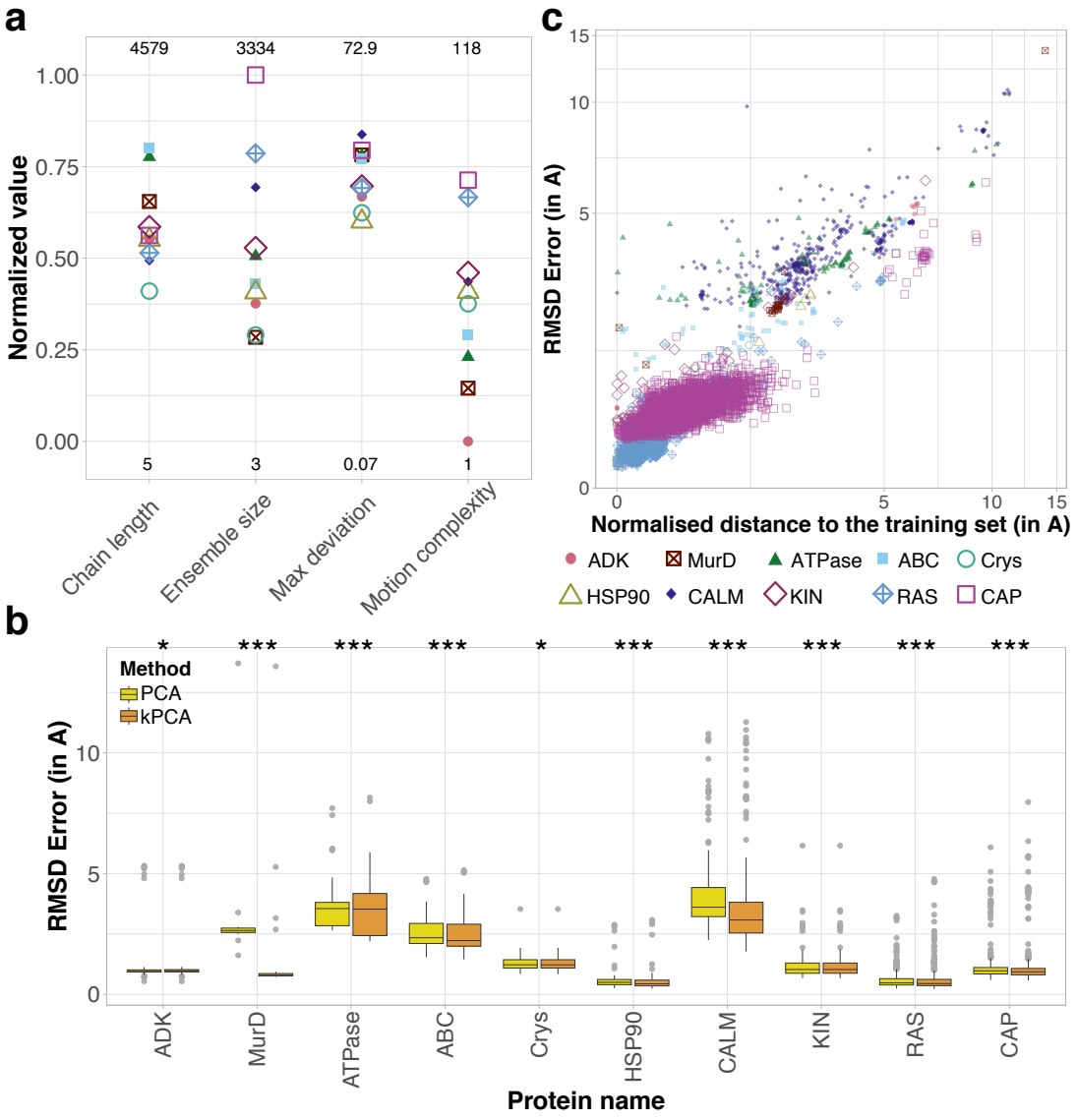


Figure 3.4: **Assessment of classical manifold learning techniques. a.** Properties of the benchmark set. For each property y, we computed its normalised value as $\frac{log(y)-\min(log(y))}{\max(log(y))-\min(log(y))}$. The minimum and maximum are determined over the whole l_{80}^{80} database. They are given at the bottom and on the top, respectively. **b.** Distributions of the RMSD reconstruction errors (in Å) for each ensemble in the benchmark set. We systematically reconstructed each conformation through a leave-one-cluster-out cross-validation procedure (see *Methods*). We set the two hyper-parameters of the kPCA (RBF kernel) to the values yielding the best reconstruction, for each ensemble. The protein names in the x-axis are ordered according to motion complexity. The stars indicate the statistical significance of the better performance of kPCA compared to linear PCA (one-sided paired t-test; *: p-val $< 1e^{-2}$; ***: p-val$< 1e^{-5}$). **c.** RMSD reconstruction error in function of the distance to the training set's convex hull in the PCA representation space.

computing their RMS deviations from the original conformations. We found that both PCA and kPCA (with RBF kernel) produced high-accuracy reconstructions (RMSD error below 2Å) for almost all proteins (**Fig. 3.4b**). The error distribution median and width vary from one protein to another and do not depend on motion complexity. For instance, all reconstructed conformations of HSP90 deviate by less than 3 Å from the original ones, while the reconstruction error can be as high as 14 Å for MurD. The distributions are overall shifted toward higher reconstruction errors for ATPase and ABC, likely due to their large size ($\sim$ 1 000 amino acids compared to less than 500 for the other proteins, **Table A.3**), and for CALM, likely due to the large amplitude of its motions (average RMSD = 10.38 ± 4.23 Å, **Table A.3**). The nonlinear kPCA performed significantly better than the linear PCA for all proteins from the benchmark. It allows increasing the percentage of high-quality reconstructions (RMSD error<2Å) from 5 to 82% for MurD and from 18 to 26% for ABC (**Table A.4**). Nevertheless, the reconstruction accuracy of kPCA varies greatly depending on the values of the two hyperparameters controlling the kernel width and the amount of regularisation (**Fig. A.6**). The optimal values vary from one system to another and determining them *a priori* is not trivial. The sigmoid and polynomial kernels may be better suited than RBF for some of the proteins, but the results are overall similar (**Fig. A.7** and **Table A.5**). By contrast, UMAP consistently produced reconstructions of substantially lower accuracy than PCA and kPCA (**Fig. A.7** and **Table A.5**). Moreover, its runtime was 100 to 100K times longer, depending on the representation space dimension.

3.3.6 Reconstruction accuracy strongly depends on the distance to the training set

The quality of the predictions strongly correlates with the distance between the test conformation and the training set's convex hull in the low-dimensional representation space (**Fig. 3.4c**). The linear PCA produces highly accurate reconstructions, with an RMSD error smaller than 2 Å, only for conformations lying in a close vicinity to the training set's convex hull (distance smaller than 3 Å). We observed a similar tendency for kPCA (**Fig. A.8**). This dependence can be appreciated by visualising how the conformations cluster in the representation space (**Fig. A.9**). For instance, the most poorly reconstructed MurD conformation forms a singleton located far away from all other conformations, particularly along the first most important principal component (**Fig. A.9B**, dark dot). For this protein, the kPCA performed substantially better than the PCA thanks to a better reconstruction of

the most populated cluster (**Fig. A.9B**, light squares). Hence, the further away from the training set, the more difficult the task. In addition, the overwhelming majority of conformations lie outside of the training set's convex hull. This observation agrees with a recent study showing that interpolation almost surely never happens with high dimensional datasets [304]. The 14 conformations (out of 4 892 in total) located inside come from ADK, CALM, KIN, RAS and CAP and are all reconstructed with high accuracy, the RMSD errors ranging from 0.2 to 2.9 Å.

3.3.7 Stereochemical quality and biological significance of the generated conformations

We assessed the physical realism of the generated conformations with PROCHECK, a popular software for checking the stereochemical quality of protein conformations, by comparing them with expected statistics [305]. The PCA- and kPCA-generated conformations displayed proportions of residues in the most favoured (or core) regions of the Ramachandran plot comparable with the experimental conformations (**Fig. A.10**). In particular, most of the conformations generated by kPCA for ADK, MurD, Crys, HSP90, RAS and CAP had more than 90% of their residues in the most favoured regions. Some of the generated conformations were even of higher stereochemical quality than their experimental counterparts. For instance, for the protein RAS, the linear PCA reconstruction greatly improved over the crystallographic structure 1PLL (chain A), from 63.6% to 94.4% residues in the most favoured Ramachandran regions. The secondary structures in the generated conformation are visibly better defined than in the experimental one (**Fig. A.11**). In this case, the PCA was able to denoise a poorly resolved conformation by learning from the other conformations in the collection. The conformations generated for CALM have the lowest stereochemical quality (**Fig. A.10**), in line with their large RMSD errors (**Fig. 3.4b**). The conformations generated with UMAP have very poor quality across all proteins to which we applied it (**Fig. A.10**, in green blue).

We further probed the biological significance of the representation spaces learnt by PCA and kPCA by investigating whether linear interpolations between extreme states in these spaces could recapitulate known intermediate conformations. We focused on ATPase as a case study and we chose the centres of clusters 0 and 4 as the end points (**Fig. A.9C**). We first learnt a low-dimensional representation space using all conformations from the two clusters, and we then generated 50 regularly spaced intermediate conformations along the trajectory between them. The generated conformations approximate known intermediates with RMSD errors as low as 3.6Å

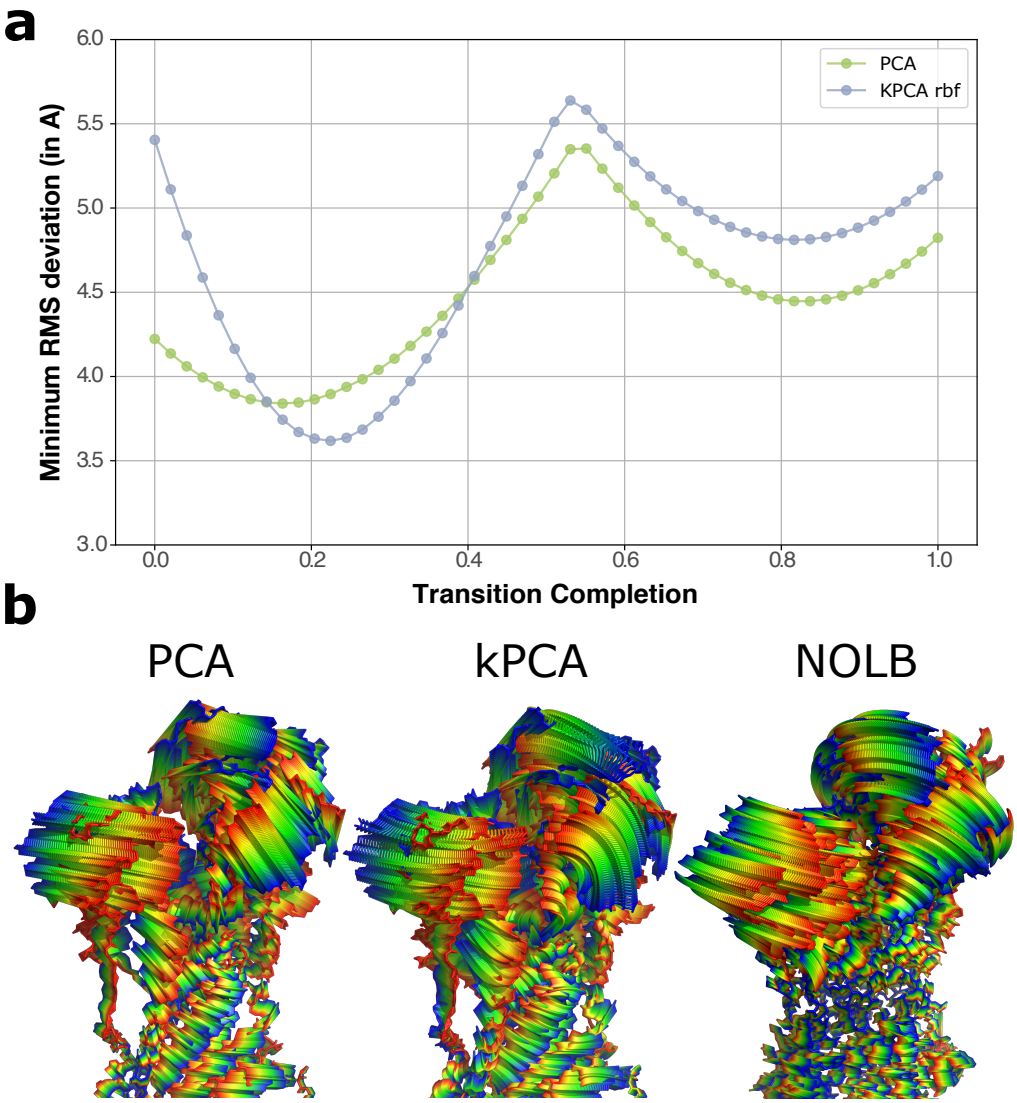


Figure 3.5: **Interpolation trajectories for ATPase.** The interpolation trajectories were computed between clusters 0 and 4, as depicted on Figure S9. For the kPCA, we used the RBF kernel with hyperparameters $\sigma = 130$ and $\alpha = 1 \times 10^{-10}$. **a.** For each of the 50 conformations generated along the PCA and kPCA trajectories, we report its minimum RMS deviation (in Å) to the known experimental intermediate conformations from clusters 1, 2 and 3. **b.** The conformations generated by PCA (left) and kPCA (middle) are colored according to the transition completion, from blue to red. We compare them with the transition computed directly in the ambient space by NOLB between the conformations 1KJUA from cluster 0 and 1T5SA from cluster 4. NOLB extrapolates motions computed from instantaneous linear and angular velocities, defined with the normal mode analysis, to large amplitudes (see *Methods*).

in the first half of the trajectory and 3.8Å in the second half (**Fig. 3.5a**). These results suggest that interpolating between known states in the learnt representation space can be a valid strategy to generate plausible intermediate conformations. In addition, one can visually appreciate the non-linear nature of the trajectories computed with kPCA compared to the linear PCA (**Fig. 3.5b**, compared left and middle panels). They bear some resemblance with trajectories computed using non-linear normal mode analysis [306, 113, 281] (**Fig. 3.5b**, compared middle and right panels).

3.3.8 Influence of data uncertainty handling and reference conformation choice

We assessed the influence of accounting for uncertainty in the data by assigning a weight to each position proportional to the number of conformations where it was resolved (see *Methods*). In principle, this operation may impact the conformations' superimposition and, as a consequence, their final coordinates, as well as the extracted motions. In practice, 95% of the ∼35 000 ensembles at l_{80}^{80} – excluding singletons and pairs, were not substantially altered by introducing position-wise uncertainty weights (**Fig. A.12**). They displayed the same displacement amplitude (± 1 Å) and motion complexity (± 1 mode). When the weights were impactful, they effectively lowered the importance of large deviations in uncertain regions, *i.e.*, poorly covered by the conformations, and prevented the associated motions, typically highly localised, from dominating the variance (**Fig. A.12**, red dots). Hence, the uncertainty weights tended to induce smaller deviations (**Fig. A.12A**), increased motion complexities (**Fig. A.12B**), and less dominant and more collective main motions (**Fig. A.12C-D**).

In addition, we performed two experiments probing the impact of choosing a different reference conformation. In the first one, we inverted the priority rules used to resolve ambiguities in the definition of the consensus sequence (see *Methods*). At a given position, in case of ambiguity, we would prefer a gap over an amino acid, thus favouring shorter reference conformations over longer ones, and a less frequent amino acid over a more frequent one, according to BLOSUM62 scores. Inverting the priority rules led to a different choice of reference in about 20% of the ∼35 000 collections. The displacement amplitude remained the same (± 1 Å) in all cases and the motion complexity deviated by more than one mode in only one case (TrwK protein, from 6 to 4 modes). This analysis shows that changing the priority rules has a negligible impact on the results. In the second experiment, we applied a much

more drastic change. Namely, we chose as alternative reference the conformation maximising the RMS deviation from the default reference. Moreover, we centred the data on the reference conformation, instead of the average conformation, prior to extracting the motions (see *Methods*). As expected, this setup yielded the most contrasted results, with about 57% of the ∼35 000 collections being impacted (**Fig. A.13**). It almost never happened that an ensemble consistently displayed a high motion complexity or a weakly contributing main motion for both references (**Fig. A.13B-C**). This result suggests that the ensembles exhibiting complex conformational rearrangements (e.g., loop deformations) among a bulk of conformations also include a few conformations comparatively far from all the others. The motions simplify when performing the PCA from the perspective of this minority. Normalising out the variance to focus on inter-residue correlations attenuates this effect (**Fig. A.14**).

3.4 Discussion

This work proposes a new perspective on the variability of protein 3D conformations. It provides the community with conformational collections representing the multiple protein states available in the PDB and a fully automated versatile computational pipeline to build custom collections. In doing so, it contributes to the representation and managing of multiple conformational models of proteins. It enhances access and understanding of protein functional states and motions and facilitates predictive methods benchmarking. Both DANCE pipeline and the produced PDB-wide data are readily usable in other studies.

We chose to rely on classical principal component analysis because of its intuitive geometrical interpretation. It allows describing protein conformational variability with a limited set of orthogonal vectors interpretable as linear motions. By default, DANCE reports the number of PCA components required to explain 50%, 80%, 85%, 90%, 95%, and 99% of the total positional variance, thus providing a multi-resolution description of the complexity of the motions explaining the observed conformational diversity. We found that a few linear motions suffice to explain over 90% of the positional variance observed in the vast majority of the conformational collections. The high complexity exhibited by a few protein families may reflect nonlinear structural deformations or seemingly random fluctuations. For instance, protein kinases exhibit highly complex loop conformational rearrangements despite a well-conserved overall fold and only two metastable functional states. Our analysis helps to identify such cases to prioritise their in-depth characterisation with more

sophisticated nonlinear dimensionality reduction techniques.

We designed DANCE for dealing primarily with single polypeptidic chains grouped based on sequence similarity. DANCE allows exploring different custom levels of sequence identity and coverage, thus providing a versatile framework for grouping the input 3D structures. Users who would like to save time may bypass the creation of the clusters and directly start from the pre-computed and weekly-updated clusters available through the RCSB PDB website. In addition, by default, DANCE analysis encompasses all polypeptidic chains found in the input 3D structures. These chains may be in different contexts and the motions extracted from the collections may be associated with the binding to a partner, as for BAX from the Bcl-2 family for instance. To ease interpretability, DANCE offers the users the possibility to restrict the context by excluding the protein chains engaged in oligomeric assemblies. Purely monomeric states represent about 15% of the $\sim$ 750K protein chains available from the PDB. Future improvements will include labelling complexes involving small molecules and accounting for them in the clustering. Furthermore, to go beyond sequence-based homology and the single-chain perspective, we have provided a proof-of-concept application study of DANCE's usefulness for comprehensively describing continuous motions shared across very distant homologs comprising different numbers of chains. We showed that ABC proteins with a wide diversity of substrates and transport mechanisms share a highly collective high amplitude opening/closing motion underlying their functioning.

In addition, our work goes beyond a descriptive analysis by showing that classical manifold learning techniques can generate plausible conformations in the vicinity of the training set. These conformations could serve as starting points for further conformational exploration, *e.g.* with molecular dynamics simulations, or as targets in drug discovery campaigns. A potential strategy would be to give them as templates to RoseTTAFold All-Atom [307] with a putative drug to guide the folding. The interpolation trajectories could provide insights into functional transitions involving substantial secondary structure rearrangements (e.g. membrane fusion proteins). The latter are particularly challenging to deal with for physics-based approaches, such as normal mode analysis [306]. Finally, our results can serve as baselines for evaluating more sophisticated approaches for predicting alternative conformations.

DANCE superimposes the conformations onto representative references and describes conformational variability as a set of linear motions of these references. This approach offers a multi-view perspective on a given collection of conformations,

easing interpretability and allowing for augmenting data in a learning context. Nevertheless, radical differences between conformations, such as fold changes, might confound the superimposition. Another limitation comes from the dependency of the superimposition on the multiple sequence alignment heuristic. Ambiguities arising from sequence similarities might result in suboptimal 3D coordinates matching and, thus, in large deviations. Future improvements will explore multi-reference or reference-free probabilistic frameworks and more refined accounts of data uncertainty [308, 309, 310, 311, 312].

Data availability

We provide public access to the conformational collections compiled by DANCE from the PDB at two levels of sequence similarity, namely l_{80}^{80} and l_{50}^{30} on Figshare [257].This repository also contains the structural similarity-based ABC transporter conformational collection along with the supplementary **Movies S1** and **S2**. In addition, we provide detailed information about the benchmark set and the assessment of PCA and kPCA.

Code availability

DANCE source codes are written in C/C++ and Python and are publicly available on GitHub at `https://github.com/PhyloSofS-Team/DANCE`. This repository also contains a Python wrapper allowing users to seamlessly run DANCE full pipeline. In addition, we provide example input 3D structures.

Chapter 4

SeaMoon

This chapter is based on the preprint *SeaMoon: Prediction of Molecular Motions Based on Language Models*, posted on September 25, 2024, on BioRχiv, freely accessible at the following link: `https://www.biorxiv.org/content/10.1101/2024.09.23.614585v1`.

SeaMoon: from protein language models to continuous structural heterogeneity

Valentin Lombard[1], Dan Timsit[1], Sergei Grudinin[2*], Elodie Laine[1,3*]

[1]Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France
[2]Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[3]Institut Universitaire de France (IUF)

Corresponding author(s): Sergei Grudinin (sergei.grudinin@univ-grenoble-alpes.fr), Elodie Laine (elodie.laine@sorbonne-universite.fr)

Abstract

How protein move and deform determines their interactions with the environment and is thus of utmost importance for cellular functioning. Following the revolution in single protein 3D structure prediction, researchers have focused on repurposing or developing deep learning models for sampling alternative protein conformations. In this work, we explored whether continuous compact representations of protein motions could be predicted directly from protein sequences, without exploiting nor sampling protein structures. Our approach, called SeaMoon, leverages protein Language Model (pLM) embeddings as input to a lightweight ($\sim$1M trainable parameters) convolutional neural network. SeaMoon achieves a success rate of up to 40% when assessed against $\sim$ 1 000 collections of experimental conformations exhibiting a wide range of motions. SeaMoon capture motions not accessible to the normal mode analysis, an unsupervised physics-based method relying solely on a protein structure's 3D geometry, and generalises to proteins that do not have any detectable sequence similarity to the training set. SeaMoon is easily retrainable with novel or updated pLMs.

4.1 Introduction

Proteins coordinate and regulate all biological processes by adapting their 3D shapes to their environment and cellular partners. Deciphering the complexities of how proteins move and deform in solution is thus of utmost importance for understanding the cellular machinery. Yet, despite spectacular advances in protein structure determination and prediction, comprehending protein conformational heterogeneity remains challenging [174, 238, 63].

Many recent approaches have concentrated on repurposing the protein structure prediction neural network AlphaFold2 [153] to generate conformational diversity [313]. Guiding the predictions with state-annotated templates proved successful for modelling the multiple functional states of a couple of protein families [241, 240]. In addition, massive sampling strategies have shown promising results for protein complexes [223] [224, 239] with notable success in the blind CASP15-CAPRI assessment [314]. While they can be deployed seamlessly with parallelized implementations [315], they remain highly resource-intensive.

Other strategies have explored promoting diversity by modulating and disentangling evolutionary signals [316]. The rationale is that amino acid co-variations in evolution reflect 3D structural constraints [317, 318, 319, 320, 321, 322, 323]. These evolutionary patterns can be extracted directly from alignments of evolutionary related sequences, or, as shown more recently, by modeling raw sequences at scale with protein language models [324, 140, 138]. Inputting shallow, masked, corrupted or sub-sampled alignments to AlphaFold2 allowed for modelling distinct conformations for a few protein families [325, 221, 219, 220]. Nevertheless, contradictory findings have highlighted difficulties in rationalising the effectiveness of these modifications and interpreting them, particularly for metamorphic proteins [326, 218, 242].

More classically, physics-based molecular dynamics (MD) is a method of choice to probe protein conformational landscapes [327]. Nonetheless, the time scales amenable to MD simulations on standard hardware remain much smaller than those spanned by slow molecular processes [328]. This limitation has stimulated the development of hybrid approaches combining MD with machine learning (ML) toward accelerating or enhancing sampling [329]. Deep neural networks can help to identify collective variables from MD simulations as part of importance-sampling strategies [328, 330, 331, 332, 333]. Or they may directly generate conformations

according to a probability distribution learnt from MD trajectories or sets of experimental structures [234, 334, 243, 335]. Diffusion-based architectures [187, 234, 233] and the more general flow-matching framework [235] provide highly efficient and flexible means to generate diverse conformations conditioned on cellular partners and ligands. Nevertheless, they are prone to hallucination, and models trained across protein families still fail to approximate solution ensembles [187].

On the other hand, the normal mode analysis (NMA) represents a data- and compute-inexpensive unsupervised alternative for accessing large-scale, shape-changing protein motions [306]. In particular, the NOLB method predicts protein functional transitions in real-time by deforming single structures along a few collective coordinates inferred with the NMA [281, 113]. The generated conformations are physically plausible and stereochemically realistic. However, the results strongly depend on the 3D geometry of the starting structure, and although some of the initial topological constraints can be easily alleviated [114], the NMA remains unsuitable for modelling extensive secondary structure rearrangements.

Training and benchmarking predictive methods is difficult due to the sparsity and inhomogeneity of the available experimental data [5]. X-ray crystallography, cryogenic-electron microscopy (cryo-EM), and nuclear magnetic resonance spectroscopy (NMR) have provided invaluable insights into protein diverse conformational states [244, 238], but only for a relatively small number of proteins [225]. Small-angle X-ray or neutron scattering (SAXS, SANS) and high-speed atomic force microscopy (HS-AFM) techniques allow for directly probing continuous protein heterogeneity, but with limited structural resolution [336, 337, 338].

Ongoing community-wide efforts aim at revealing the full potential of the available structural data by collecting, clustering, curating, visualising and functionally annotating experimental protein structures together with high-quality predicted models [245, 244, 246, 247, 248, 249, 250, 251]. For instance, the DANCE method produces movie-like visual narratives and compact continuous representations of protein conformational diversity, interpreted as *linear motions*, from static 3D snapshots [339]. DANCE application to the Protein Data Bank (PDB) [5] revealed that the conformations observed for most protein families lie on a low-dimensional *manifold*. Classical dimensionality reduction techniques can learn this manifold and generate unseen conformations with reasonable accuracy, albeit only in close vicinity of the training set [339].

Here, we explored the possibility of predicting protein motions directly from amino acid sequences without exploiting nor sampling protein 3D structures. To do so, we leveraged protein Language Models (pLMs) pre-trained through self-supervision over large databases of protein-related data. Our approach, SEAquencetoMOtioON or SeaMoon, is a 1D convolutional neural network inputting a protein sequence pLM embedding and outputting a set of 3D displacement vectors (**Fig. 4.1**). The latter define protein residues' relative motion amplitudes and directions. We tested whether SeaMoon could capture the *linear motion manifold* underlying experimentally resolved conformations across thousands of diverse protein families [339]. To this end, we devised an objective function invariant to global translations, rotations, and dilatations in 3D space. SeaMoon achieved a success rate similar to the normal mode analysis (NMA) when inputting purely sequence-based pLM embeddings [138] without any knowledge about protein 3D structures. It could generalise to proteins without any detectable sequence similarity to the training set and capture motions not directly accessible from protein 3D geometry. Injecting implicit structural knowledge with sequence-structure bilingual or multimodal pLMs [167, 162] further boosted the performance. This work establishes a community baseline and paves the way for developing evolutionary- and physics-informed neural networks to predict continuous protein motions.

4.2 Results

The approach introduced in this work, SeaMoon, predicts continuous representations of protein motions with a convolutional neural network inputting pLM sequence embeddings (**Fig. 4.1**). We considered the purely sequence-based pLM ESM2 [340] and two structure-aware pLMs, namely ESM3 [167] and ProstT5 [162]. ESM3 is the largest model (**Table B.1**), and it can condition on and reconstruct several protein sequence and structural properties. ProstT5, the smallest model (**Table B.1**), is a fine-tuned version of the sequence-only model T5 that translates amino acid sequences into sequences of discrete structural states and reciprocally. We trained and tested SeaMoon on over $\sim$ 17 000 experimental conformational collections representing a non-redundant set of the PDB at 80% sequence similarity. We used the principal components extracted from these collections as ground-truth linear motions to which we compared SeaMoon predicted 3D vectors. The latter are not anchored on a particular conformation and may be in any arbitrary orientation. To allow for a fair comparison, we determined the optimal rotation and scaling between the ground-truth and predicted vectors before computing the error between them

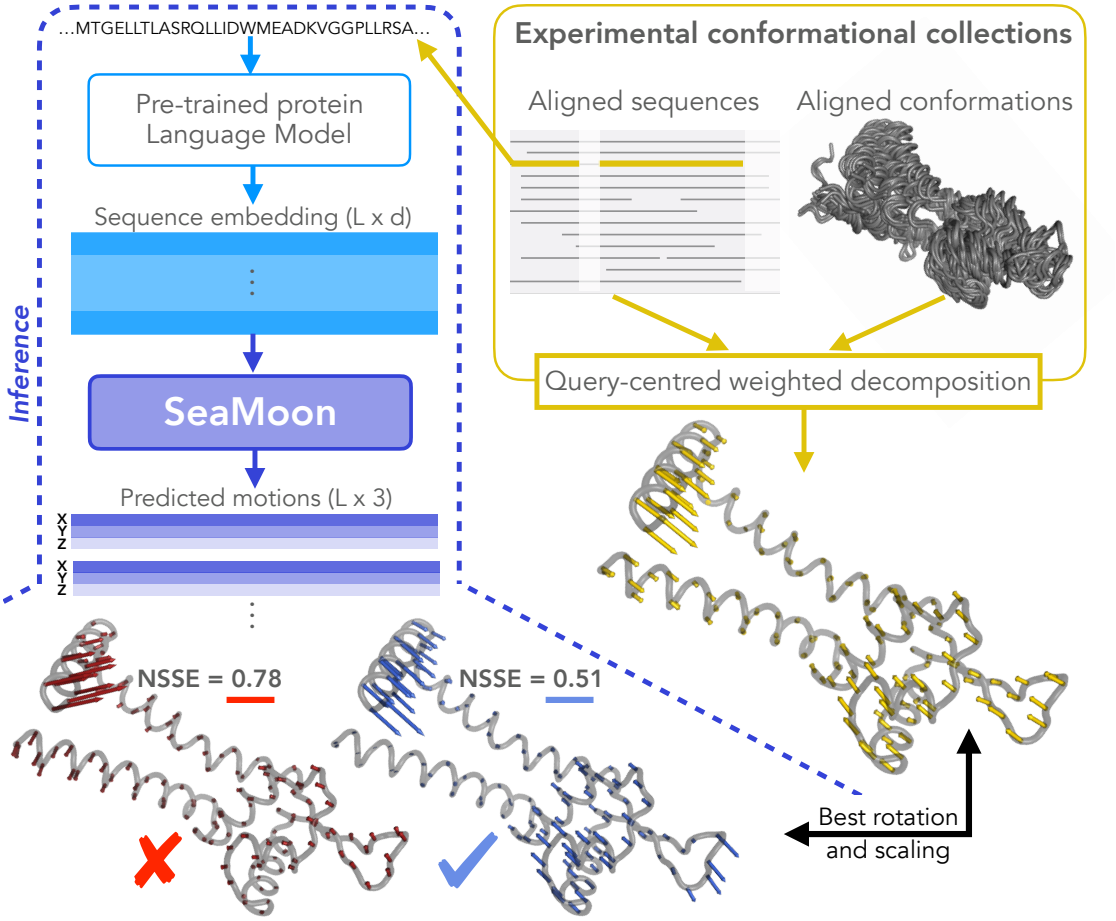


Figure 4.1: **Outline of SeaMoon's approach.** SeaMoon takes as input a high-dimensional $L \times d$ matrix representation of a protein sequence of length L computed by a pre-trained pLM. It outputs a set of 3D vectors of length L representing linear motions. The training procedure regresses these output motions (blue and red arrows) against ground-truth ones (yellow arrows) extracted from experimental conformational collections through principal component analysis. For this, SeaMoon identifies the transformation (rotation and scaling) minimising their discrepancy, computed as a sum-of-squares error (SSE). We consider predictions with a normalised error (NSSE) smaller than 0.6 as acceptable. We show the query protein 3D structure only for illustrating the motions, it is not used by SeaMoon nor by the pLM generating the input embeddings..

(see *Methods* for details). Based on visual inspection, we considered predictions as acceptable when their normalised sum-of-squares error (NSSE) was smaller than

0.6 (**Fig. 4.1**). See **Fig. B.1** for illustrative examples of different error levels. By comparison, random predictions typically display errors above 0.9 (**Fig. B.2**). SeaMoon is highly computationally efficient. It took 12s to predict 3 motions for each of 1 121 test proteins on a desk computer equipped with Intel Xeon W-2245 @ 3.90 GHz.

4.2.1 SeaMoon predicts motions from sequences across diverse protein families

SeaMoon predicted at least one acceptable linear motion for each of 300 test proteins from the purely sequence-based ESM2 embeddings (**Table 4.1** and **Fig. 4.2A**). Its performance was comparable to that of the purely geometry-based unsupervised NMA. SeaMoon success rate improved by 25-40% when inputting structurally-informed embeddings computed by ESM3 or ProstT5, outperforming the NMA by a large margin (**Table 4.1** and **Fig. 4.2A**). ProstT5, with the smallest number of parameters and embedding dimensions (**Table B.1**), yielded the best overall performance (**Fig. 4.2A**, paired Wilcoxon signed-rank test p-values $< 10^{-6}$ and $< 10^{-9}$ with respect to ESM3 and ESM2, respectively). In addition, we observed a boost in performance by up to 10% upon stimulating the model to learn a one-sequence-to-many-motions mapping (**Table 4.1** and **Fig. 4.2A**). More specifically, we augmented the training data by using multiple (up to 5) reference conformations per experimental collection (**Table B.2**). While the pLM embeddings within a collection should be highly similar, the extracted motions may differ substantially from one reference to another [339]. The positive impact of this data augmentation strategy was most visible for the ESM-based version of SeaMoon (**Table 4.1** and **Fig. 4.2A**).

SeaMoon effectively generalised to unseen proteins across diverse families (**Table 4.1**, **Fig. 4.2B**, **Fig. B.4 and Fig. B.5**). It produced high-quality predictions at different levels of similarity to the training set, which we can interpret as varying difficulty levels. For instance, SeaMoon-ESM2(x5) almost perfectly recapitulated the motions of antibodies (**Fig. B.5A**), a class of proteins well represented in both train and test sets. Beyond such easy cases, SeaMoon-ESM2(x5) could transfer knowledge between proteins with similar 3D folds but highly divergent sequences. The ATP-binding cassette (ABC) transporter superfamily provides an illustrative example of this intermediate difficulty (**Fig. B.5B**). SeaMoon-ESM2(x5) accurately predicted the opening-closing motion of a putative ABC transporter from *Campylobacter jejuni* (**Fig. B.5B**, 5T1PE, $NSSE = 0.33$) that does not have any

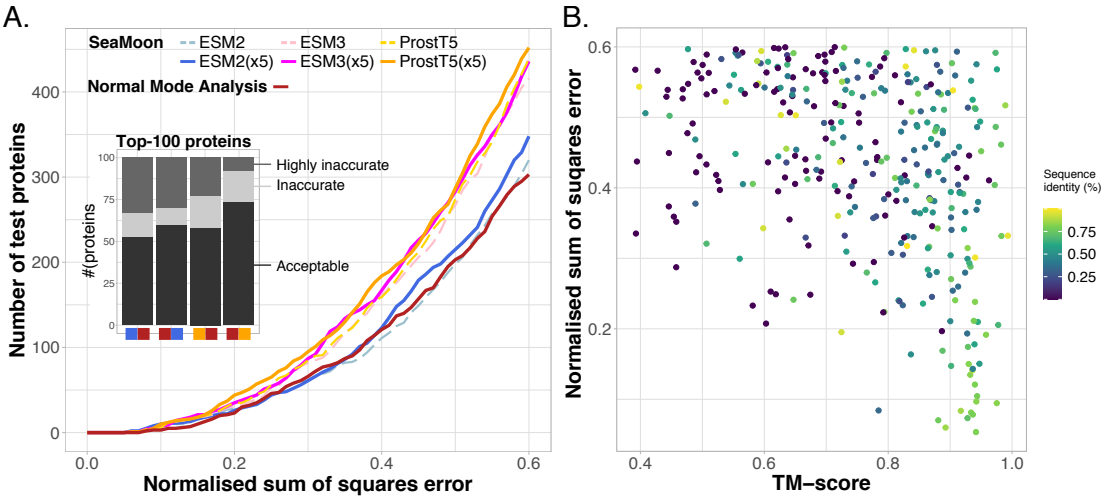


Figure 4.2: **SeaMoon performance and generalisation capability.** We report the NSSE of the best match between 3 predictions and 3 ground-truth motions for each of the 1 121 test proteins. **A.** Cumulative NSSE for six different versions of SeaMoon and for the NMA. We tested three pLMs, namely ESM2, ESM3 and ProstT5, and a data augmentation strategy with 5 training samples per experimental collection (x5). We cropped the plot at $NSSE = 0.6$ for ease of visualisation; see Fig. B.3 for the full curves. Inset: Agreement between a selection of methods. For instance, the first bar stack gives the numbers of proteins for which the NMA (right red square) produced acceptable ($NSSE < 0.6$), inaccurate ($0.6 < NSSE < 0.75$) or highly inaccurate ($NSSE > 0.75$) predictions among the top-100 proteins best-predicted by SeaMoon-ESM2(x5) (left blue square). **B.** NSSE computed for SeaMoon-ESM2(x5) in function of sequence and structural similarity to the training set.

detectable sequence similarity with the training set. This motion is characteristic of the "Venus Fly-trap" mechanism for transporting sugars [341] and is shared with a structurally similar ABC transporter from the training set (**Fig. B.5B**, 7C68B, TM-score = 0.83). At the most difficult level, SeaMoon-ESM2(x5) successfully captured the motions of proteins completely unrelated to the training set, such as the benzoyl-coenzyme A reductase from *Geobacter metallireducens* (**Fig. B.5C**, 4Z3ZF, $NSSE = 0.37$).

4.2.2 SeaMoon complementary to the normal mode analysis

We investigated the extent of the agreement between the purely sequence-based version of SeaMoon and the purely geometry-based NMA (**Fig. 4.2A**, inset, and

Table 4.1: **Performance and dependence on the similarity to the training set**

Method	Protocol	Number of proteins w. acceptable predictions	Correlation w. TM-Score	Correlation w. sequence id.
SeaMoon	ESM2	320 (29%)	-0.35	-0.20
	ESM2(x5)	348 (31%)	-0.39	-0.26
	ESM3	416 (37%)	-0.31	-0.18
	ESM3(x5)	436 (39%)	-0.38	-0.22
	ProstT5	439 (39%)	-0.32	-0.12
	ProstT5(x5)	**452 (40%)**	-0.37	-0.20
NMA		303 (27%)	-0.09	0.03

We consider predictions as acceptable if their normalised sum-of-squares error is smaller than 0.6. The highest success rate is highlighted in bold.

Fig. B.6). Among the top-100 proteins best-predicted by SeaMoon-ESM2(x5), about half exhibit motions accessible to the NMA (**Fig. 4.2A**, inset). Most of these motions involve a large portion of the protein (median collectivity $\kappa = 0.69$) and correspond to large conformational changes (median deviation of 5.1Å). They include functional opening-closing motions of virulence factors, thermophilic proteins, metalloenzymes, periplasmic binding proteins, dehydrogenases, glutamate receptors, and antibodies (see **Fig. B.7** for illustrative examples). On the other hand, the NMA performed extremely poorly for a third of SeaMoon-ESM2(x5) top-100 ($NSSE > 0.75$, see **Fig. 4.2A**, inset). The associated motions tend to be localised with median collectivity $\kappa = 0.20$.

The bacterial toxins PemK and protective antigen (PA) from anthrax illustrate SeaMoon's capability to go beyond the NMA physics-based inference for highly localised motions and fold-switching deformations (**Fig. 4.3**). SeaMoon-ESM2(x5) captured the PemK's loop L12 motion with high precision (**Fig. 4.3A**, $NSSE = 0.24$) whereas the NMA failed to delineate the mobile region in the protein and to infer its direction of movement (**Fig. 4.3A**, in red). This highly localised motion ($\kappa = 0.17$) plays a decisive role in regulating PemK RNAse activity by promoting the formation of the PemK-PemI toxin-antitoxin [342]. In the anthrax protective antigen, SeaMoon-ESM2(x5) accurately predicted the relative motion amplitudes and directions of an 80 residue-long region that detaches from the rest of the protein upon forming an heptameric pore **Fig. 4.3B**). By contrast, the NMA predicted a breathing motion poorly approximating the ground-truth one (**Fig. 4.3B**), likely

due to its assumption that proteins behave as elastic networks. PA's ∼30Å-large conformational transition is essential for the translocation of the bacterium's edema and lethal factors to the host cell [343]. PemK and PA do not have any detectable sequence similarity to the training set. SeaMoon likely leveraged information coming from training proteins with similar folds and functions from other bacteria [344, 345].

Reciprocally, SeaMoon covered 60% of the top-100 proteins best-predicted by the NMA with ESM2 embeddings, and up to 75% with ProstT5 embeddings (**Fig. 4.2A**, inset, and **Fig. B.6**). Using implicit structural knowledge allowed recovering elastic motions such as that exhibited by the mammalian plexin A4 ectodomain (**Fig. B.8**, $NSSE = 0.28$). Taken together, SeaMoon-ProstT5(x5) and the NMA approximated the motions of 554 test proteins (out of 1121, 49%) with reasonable accuracy (**Table 4.1**). This result suggests that combining SeaMoon transfer learning approach with the physics- and geometry-based NMA could be a valuable strategy.

4.2.3 SeaMoon can recapitulate entire motion subspaces

Beyond assessing individual predictions, we evaluated the global similarities between predicted and ground-truth 3-motion subspaces focusing on the test proteins for which SeaMoon produced at least one acceptable prediction (**Table 4.1**). We found that SeaMoon motion subspaces were fairly similar to the ground-truth ones, with a Root Mean Square Inner Product (RMSIP) [126, 121, 122] higher than 0.5, for almost two thirds of these proteins. We observed an excellent correspondence for a dozen proteins, *e.g.*, the *Mycobacterium* phage Ogopogo major capsid protein (**Fig. 4.4** and **Fig. B.9**). The purely sequence-based SeaMoon-ESM2(x5) achieved an RMSIP of 0.75 on this protein, and the structure-aware SeaMoon-ProstT5(x5) reached 0.82. SeaMoon-ProstT5(x5) first, second and third predicted motions had a Pearson correlation of 0.93, 0.73 and 0.75 with the first, third and second ground-truth principal components, respectively (**Fig. 4.4A**). The associated NSSE were all smaller than 0.5 (**Fig. 4.4B**). By inspecting the training set, we could identify several major capsid proteins from other bacteriophages sharing the same HK97-like fold as the Ogopogo one (TM-score up to 0.78), despite relatively low sequence similarity (up to 34%). The ability of SeaMoon to recapitulate the Ogopogo protein entire motion subspace with reasonable accuracy likely reflects the high conservation of major capsid protein dynamics upon forming icosahedral shells [346].

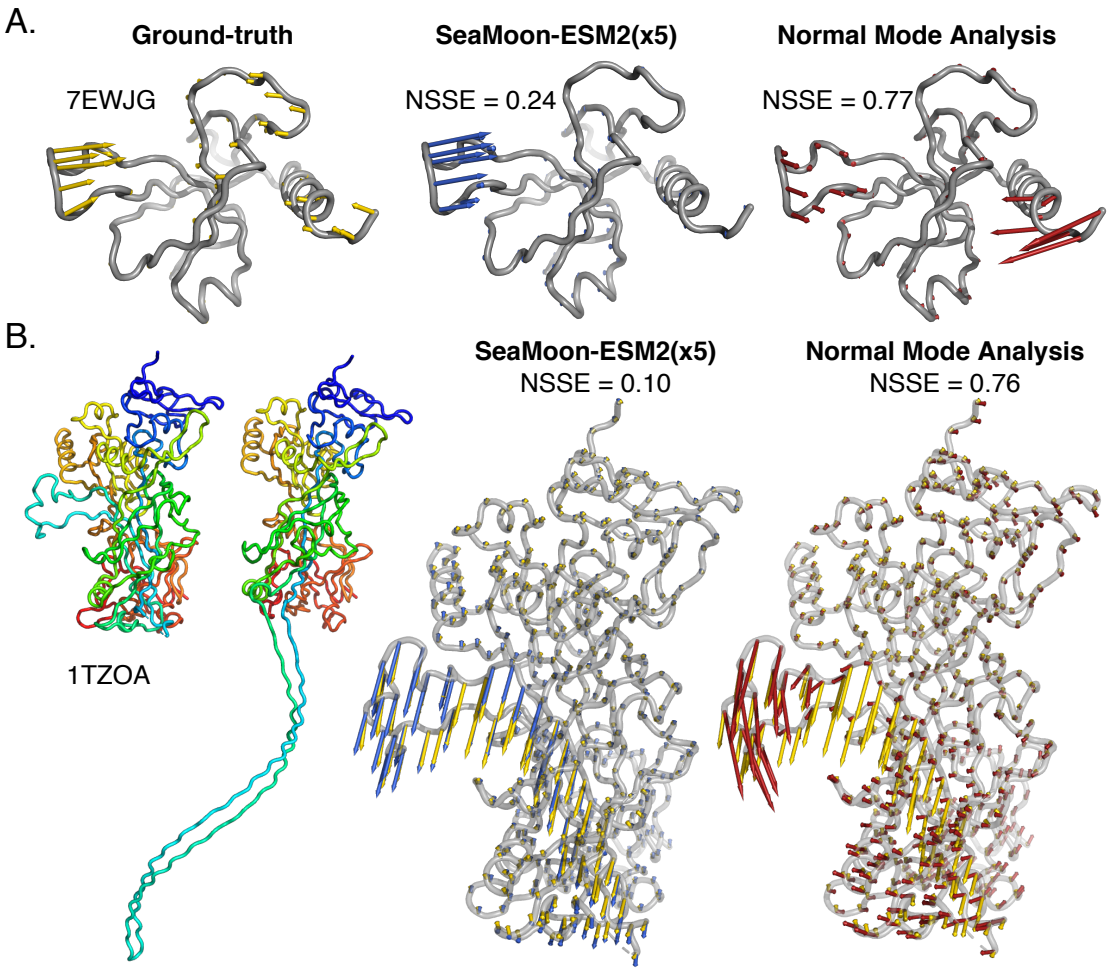


Figure 4.3: **Examples of motions well predicted by SeaMoon and not by the NMA.** The arrows depicted in yellow, blue and red on the grey 3D structures represent the ground-truth motions and the best-matching predictions from SeaMoon-ESM2(x5) and the NMA, respectively. **A.** Bacterial toxin PemK (PDB code: 7EWJ, chain G) from the test set. It does not have any detectable sequence similarity to the training set **B.** Anthrax protective antigen (PDB code: 1TZO, chain A) from the validation set. We show the two most extreme conformations of the collection on the left, colored according to the residue index, from the N-terminus in blue, to the C-terminus in red. The closest homolog from the training set shares 35% sequence similarity.

4.2.4 Contributions of the inputs and design choices

We investigated the contribution of SeaMoon inputs, architecture and objective function to its success rate through an ablation study, starting from SeaMoon-

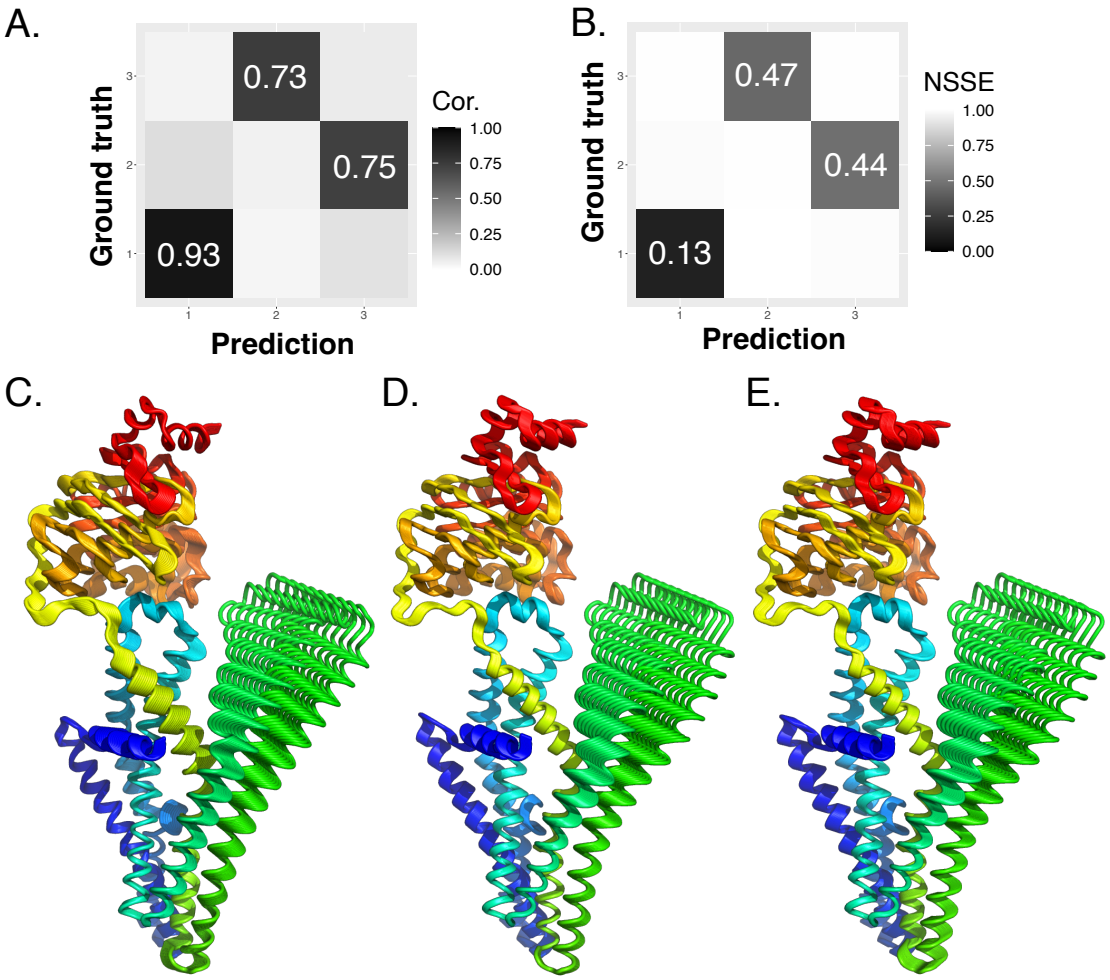


Figure 4.4: **Motion subspace comparison and deformation trajectories. A-B.** Ogopogo major capsid protein motion subspace. PDB code: 8ECN, chain B. **A.** Pairwise similarities measured as Pearson correlations between the ground-truth motions and SeaMoon-ProstT5(x5) predictions. **B.** Pairwise discrepancies measured as NSSE. **C-E.** Trajectories of a human ABC transporter (PDB code: 7D7R, chain A) deformed along its first ground-truth principal component (C) and the best-matching SeaMoon-ProstT5(x5) prediction (D-E). **D.** The prediction is optimally aligned with the ground truth. **E.** The orientation of the prediction minimises the protein conformation's angular velocity. Each trajectory comprises 10 conformations coloured from blue at the N-terminus to red at the C-terminus.

ProstT5 baseline model (**Table B.3** and **Fig. B.10**). Inputting random matrices instead of pre-trained pLM embeddings or using only positional encoding had the most drastic impacts. Still, we observed that the network can produce accurate

predictions for over 100 proteins in this extreme situation (**Fig. B.10**, in grey). Annihilating sequence embedding context by setting all convolutional filter sizes to 1 also had a dramatic impact, reducing to success rate from 40 to 25% (**Table B.3** and **Fig. B.10**). Moreover, a 7-layer transformer architecture (see *Methods*) underperformed SeaMoon's convolutional neural network, despite having roughly the same number of free parameters (**Fig. B.10**, in brown). Finally, disabling either sign flip or reflection (*i.e.*, pseudo-rotation) or permutation when computing the loss degraded the performance by 6 to 15% (**Fig. B.10**, in light green). This result underlines the utility of implementing a permissive and flexible comparison of the predicted and ground-truth motions during training.

4.2.5 SeaMoon practical utility to deform protein structures

SeaMoon does not use any explicit 3D structural information during inference. Its predictions are independent of the global orientation of any protein conformation, making it impractical to directly use them to deform protein structures. To partially overcome this limitation, we propose an unsupervised procedure to orient SeaMoon predicted vectors with respect to a given protein 3D conformation. This method exploits the rotational constraints of the ground-truth principal components. Namely, the total angular velocity of the reference conformation subjected to a ground-truth principal component is zero (see *Methods*). Therefore, we determine the rotation that must be applied to the predicted motion vectors to minimize the total angular velocity of a target conformation.

This strategy proved successful for the vast majority of SeaMoon's highly accurate predictions. SeaMoon-ProstT5(x5) predicted motion vectors, oriented to minimise angular velocity, exhibit an acceptable error (< 0.6) in 85% of cases where the optimal alignment with the ground truth results in $NSSE < 0.3$. This result indicates that predictions that approximate well the ground-truth principal components also preserve their properties. The human ABC transporter sub-family B member 6 gives an illustrative example where the third predicted motion vector approximates the first ground-truth principal component with $NSSE = 0.20$ upon optimal alignment and 0.22 upon angular velocity minimisation (**Fig. 4.4C-E**). Overall, the procedure allowed for correctly orienting acceptable predictions for 215 test proteins.

Note that this post-processing increases computing time significantly, from 12s to 24m over the 1 121 test proteins on a desk computer equipped with Intel Xeon

W-2245 @ 3.90 GHz.

4.3 Discussion

This proof-of-concept study explores the extent to which protein sequences encode functional motions. SeaMoon reconstructs these motions within an invariant subspace directly from sequence-based pLM embeddings. Our results indicate that incorporating structure-aware input embeddings significantly improves the success rate. Moreover, they highlight SeaMoon's ability to transfer knowledge about motions across distant homologs, leveraging the universal representation space of pLMs. However, the framework's capacity to predict entirely novel motions has yet to be fully assessed.

SeaMoon's transfer learning approach complements unsupervised methods that rely solely on the 3D geometry of protein structures, such as Normal Mode Analysis (NMA). Future work will focus on integrating these two sources of information into a unified, end-to-end framework. Incorporating explicit structural information for a target protein could resolve the ambiguity in orienting predicted motions without requiring ground-truth knowledge.

One current limitation is the scarcity of functional motions in the training set, raising concerns about its accuracy and completeness. Both SeaMoon and NMA struggle to predict certain motions, suggesting that these may lack biological or physical relevance. Conversely, SeaMoon could be used to assess the evolutionary conservation of motions. Another limitation of the current approach is its reliance on a linear description of protein motion subspaces. Linear principal components are insufficient for describing complex loop deformations or large rearrangements of secondary structures. Introducing non-linearity could yield more realistic motion predictions. Future work will address these issues, potentially augmenting the training set with *in silico* generated data, such as motions derived from MD and NMA simulations, or protein conformations predicted by AlphaFold.

Despite these limitations, the current findings offer valuable insights for integrative structural biology. SeaMoon provides a compact representation of continuous structural heterogeneity in proteins, enabling the sampling of conformations through a generative model. Additionally, the estimated motion subspaces can be used to compute protein conformational entropy. Lastly, our framework is highly versatile, featuring a lightweight, trainable deep learning architecture that does not depend on fine-tuning a large pre-trained model. This flexibility allows users to easily adapt the system to new input pLM embeddings without modifying the model

architecture.

4.4 Methods

4.4.1 Datasets

To generate training data, we constructed a non-redundant set of conformational collections representing the whole PDB (as of June 2023) using DANCE [339]. To ensure high quality of the data, we replaced the raw PDB coordinates with their updated and optimised versions from PDB-REDO whenever possible [269]. We used a stringent setup where each conformational collection is specific to a set of close homologs. Specifically, any two protein chains belonging to the same collection share at least 80% sequence identity and coverage. We filtered out the collections with too few or too many data points. Namely, we asked for at least 4 and at most 500 conformations and a representative protein chain comprising between 30 and 1 000 residues. We further retained only Cα atoms (option `-c`) and used coordinate weights to account for uncertainty (option `-w`).

For each collection, DANCE extracted the $K = 3$ principal components contributing the most to its total positional variance [339]. We interpret these components as the main linear motions explaining the collection's conformational diversity. Namely, the kth principal component defines a set of 3D displacement vectors $\{\vec{x}_{ik}^{\mathrm{GT}}, i = 1, 2, ...L\}$ for the L protein residues' Cα atoms. We normalised these vectors to facilitate their comparison across different proteins, such that $\sum_{i=1}^{L} \|\vec{x}_{ik}^{\mathrm{GT}}\|^2 = L$. We further applied three filtering criteria with the aim of excluding collections with low diversity or highly non-linear complex deformations: (*i*) maximum Root Mean Squared Deviation (RMSD) between any two conformations of at least 2 Å, (*ii*) first principal component (main linear motion) contributing at least 80% of the total variance and (*iii*) involving at least 12 residues, *i.e.*, $L \times \kappa \geq 12$, where κ is the collectivity of the principal component (see definition below). This operation resulted in 7 339 collections, randomly split between train (70%), validation (15%) and test (15%) sets.

DANCE makes use of a reference conformation to superimpose the Cα atoms' 3D coordinates and centre them prior to extracting motions with PCA. By default, the reference corresponds to the protein chain with the most representative amino acid sequence [339]. In order to augment the data, we defined up to 4 alternative reference conformations, in addition to the default one (option `-n 5`). At each iteration, DANCE chose the new reference conformation as the one displaying

the highest RMSD from the previous one. This strategy maximises the impact of changing the reference and thus the diversity of the extracted motions.

4.4.2 Model Specifications

Input features

SeaMoon takes as input embeddings computed from pre-trained pLMs, namely Evolutionary Scale Models ESM2-T33-650M-UR50 [340] and ESM3-small (1.4B) [167], as well as Protein sequence-structure T5 [162]. ESM2-T33-650M-UR50 is a BERT [147] style 650-million-parameter encoder-only transformer architecture trained on all clusters from Uniref50 [149, 148], a version of UniProt [236] clustered at 50% sequence similarity, augmented by sampling sequences from the Uniref90 clusters of the representative chains (excluding artificial sequences). ESM3-small (1.4B) is a transformer-based [135] all-to-all generative architecture that both conditions on and generates a variety of different tracks representing protein sequence, secondary and tertiary structure, solvent accessibility and function. It was trained on over 2.5 billion natural proteins collected from sequence and structure databases, including UniRef, MGnify [347], OAS [348] and the PDB [5], augmented with synthetic sequences generated by an inverse folding model [167]. Protein sequence-structure T5 is a bilingual pLM trained on a high-quality clustered version of the AlphaFold Protein Structure Database [349, 175] to translate 1D sequences of amino acids into 1D sequences of 3Di tokens representing 3D structural states [163] and vice versa. The 3Di alphabet, introduced by the 3D-alignment method Foldseek [163], describes tertiary contacts between protein residues and their nearest neighbours. This 1D discretised representation of 3D structures is sensitive to fold change but robust to conformational rearrangements. Protein sequence-structure T5 expands on ProtT5-XL-U50 [140], an encoder-decoder transformer architecture [156] trained on reconstructing corrupted amino acids from the Big Fantastic Database [350] and UniRef50. Throughout the text, we refer to these pLMs as ESM2, ESM3 and ProstT5, respectively. We used the pre-trained pLMs as is, without fine-tuning their weights, and we gave them only amino acid sequences as input.

Model's architecture

SeaMoon's architecture is a convolutional neural network [351] taking as input a sequence embedding of dimensions $L \times d$, with L the number of protein residues and

d the representation dimension of the chosen pLM, namely 1 280 for ESM2, 1 536 for ESM3, and 1 024 for ProstT5, and outputting K predicted tensors of dimensions $L \times 3$. It comprises a linear layer followed by two hidden 1-dimensional convolutional layers with filter sizes of 15 and 31, respectively, and finally K parallel linear layers (**Table B.1**). SeaMoon's convolutional architecture allows handling sequences of any arbitrary length L and preserving this dimension throughout the network. All layers were linked through the LeakyReLu activation function [352], as well as 80% dropout [353]. We experimented with other types of architectures, including those based on sequence transformers, and chose the one based on CNNs as it demonstrated the maximum accuracy at a reasonable number of trained parameters. Please see **Table B.3** and **Fig. B.10** for more details. We implemented the models in PyTorch [354] v2.1.0 using Python 3.11.9.

By design, the SeaMoon model predicts the K motion tensors in a latent space that is invariant to the protein's actual 3D orientation. To align these predictions with a given 3D conformation, additional information, such as the ground-truth motions, is required, as explained below.

Loss function

We aim to minimise the discrepancy between the predicted tensor X and the ground-truth tensor X^{GT}, both of dimensions $L \times K \times 3$, expressed as a weighted aligned sum-of-squares error loss,

$$\mathcal{L} = \frac{1}{L} \min_{R,S,P} \left(\sum_{i=1}^{L} w_i \| R(PX_i^{\mathrm{GT}})^T - (SX_i)^T \|_F^2 \right), \tag{4.1}$$

where X_i defines the set of K 3D displacements vectors $\{\vec{x}_{ik} \equiv (X_{i,k,.})^T,\ k = 1, 2, ...K\}$ predicted for the Cα atom of residue i, X_i^{GT} defines the corresponding ground-truth 3D displacement vector set, $\| \cdot \|_F$ designates the Frobenius norm, and w_i is a weight reflecting the confidence in the ground-truth data for residue i [339]. It is computed as the proportion of conformations in the experimental collection with resolved 3D coordinates for residue i. The matrices R, of dimension 3×3, and P, of dimension $K \times K$, allow for rotating and permuting the ground-truth vectors to optimally align them with the predicted ones. We chose to apply the transformations to the ground-truth vectors for gradient stability. We allow for rotations R because SeaMoon relies solely on a protein sequence embedding as input. Its predictions are not anchored in a particular 3D structure and hence, they may be in any arbitrary orientation. We allow for permutation P to stimulate

knowledge transfer across conformational collections. The rationale is that a motion may be shared between two collections without necessarily contributing to their positional variance to the same extent. Additionally, we allow for scaling predictions with the $K \times K$ diagonal matrix S, so that SeaMoon can focus on predicting only the relative motion amplitudes between the amino acid residues.

In practice, we first jointly determine the optimal permutation P and rotation R of the ground-truth 3D vectors. We test all possible permutations, and, for each, we determine the best rotation by solving the orthogonal Procrustes problem [355, 356]. We shall note that the optimal solution may be a pseudo-rotation, *i.e.*, $\det(R) = -1$, which corresponds to the combination of a rotation and an inversion. The loss can then be reformulated as,

$$\mathcal{L} = \frac{1}{L} \min_{S} \left(\sum_{k=1}^{K} \sum_{i=1}^{L} w_i \|\vec{x}_{ik}^{\text{GT-trans}} - S_{kk}\vec{x}_{ik}\|^2 \right), \tag{4.2}$$

where $\vec{x}_{ik}^{\text{GT-trans}}$ is the ground-truth 3D displacement vector for residue i matching the predicted 3D vector $\vec{x}_{ik}$ and aligned with it, and $S_{kk} \in \mathbb{R}$ is the kth scaling coefficient, *i.e.* the kth non-null term of the diagonal scaling matrix S. The optimal value for S_{kk} is computed as,

$$S_{kk} = \frac{\sum_{i=1}^{L} w_i (\vec{x}_{ik}^{\text{GT-trans}})^T \vec{x}_{ik}}{\sum_{i=1}^{L} w_i \|\vec{x}_{ik}\|^2}. \tag{4.3}$$

Training

We trained six models (**Table B.2**) to predict $K = 3$ motions using the Adam optimizer [131] with a learning rate of 1e-02. We used a batch size of 64 input sequences and employed padding to accommodate sequences of variable sizes in the same batch. We trained for 500 epochs and kept the best model according to the performance on the validation set.

Inference

We provide an unsupervised procedure to orient SeaMoon's predicted motions with respect to a target 3D conformation $\vec{C}_i$ during inference. This approach relies on the assumption that correct predictions comply with the same rotational constraints as ground-truth motions (see *Supplementary Methods*). Specifically, these constraints state that the cross products between the positional 3D vectors of the reference conformation C^0 and the 3D displacement vectors defined by a

ground-truth principal component X_k^{GT} result in a null vector,

$$\sum_{i=1}^{L} \vec{C}_i^{\,0} \times \vec{x}_{ik}^{\mathrm{GT}} = \vec{0}. \tag{4.4}$$

Assuming that the motion tensor X_k predicted by SeaMoon preserves this property, we determine the rotation R that minimises the following cross-product,

$$\sum_{i=1}^{L} \vec{C}_i \times R\vec{x}_{ik} = \vec{0}. \tag{4.5}$$

This problem has at most four solutions and we solve it exactly using the symbolic package *wolframclient* in Python. See *Supplementary Methods* for a detailed explanation. In practice, we observe that these four solutions reduce to two pairs of highly similar rotations.

4.4.3 Evaluation

We assessed SeaMoon predictions on each test protein from two different perspectives. In the first assessment, we considered all $K \times K$ pairs of predicted and ground-truth motions and estimated the discrepancy between the two motions within each pair after optimally rotating and scaling them. We focused on the best matching pair for computing success rates and illustrating the results. In the second assessment, we considered the predicted and ground-truth motion subspaces at once and estimated their permutation-, rotation- and scaling-invariant global similarity. In addition, we estimated discrepancies and similarities between individual predicted and ground-truth motions after globally matching and aligning the subspaces. We detail our evaluation metrics and procedures in the following.

Normalised sum-of-squares error

At inference time, we estimate the discrepancy between the kth predicted motion and the lth ground-truth principal component by computing their weighted sum-of-squares error under optimal rotation R^{opt} and scaling s^{opt},

$$SSE = \frac{1}{L} \sum_{i=1}^{L} w_i \|\vec{x}_{il}^{\text{GT-trans}} - s^{opt}\vec{x}_{ik}\|^2, \tag{4.6}$$

$$\text{with } \vec{x}_{il}^{\text{GT-trans}} = R^{opt}\vec{x}_{il}^{\mathrm{GT}} \tag{4.7}$$

In the best-case scenario, the prediction is colinear to the transformed ground-truth, $\vec{x}_{il}^{\text{GT-trans}} = c\vec{x}_{ik}$, $c \in \mathbb{R}$, such that $(\vec{x}_{il}^{\text{GT-trans}})^T \vec{x}_{ik} = \|\vec{x}_{il}^{\text{GT-trans}}\|\|\vec{x}_{ik}\| = c\|\vec{x}_{ik}\|^2$, $\forall i \in 1, 2, ...L$. By virtue of Eq. (4.3), the scaling coefficient s^{opt} will be equal to c, and thus, the error will be null,

$$SSE_{min} = \frac{1}{L}\sum_{i=1}^{L} w_i \|\vec{x}_{il}^{\text{GT-trans}} - c\vec{x}_{ik}\|^2 = \frac{1}{L}\sum_{i=1}^{L} w_i \|c\vec{x}_{ik} - c\vec{x}_{ik}\|^2 = 0. \tag{4.8}$$

In the worst-case scenario, the prediction is orthogonal to the ground truth, such that $(\vec{x}_{il}^{\text{GT-trans}})^T \vec{x}_{ik} = 0$, $\forall i \in 1, 2, ...L$. The scaling coefficient will be null and, hence, this situation is equivalent to having a null prediction,

$$SSE_{max} = \frac{1}{L}\sum_{i=1}^{L} w_i \|\vec{x}_{il}^{\text{GT-trans}} - \vec{0}\|^2 = \frac{1}{L}\sum_{i=1}^{L} w_i \|\vec{x}_{il}^{\text{GT-trans}}\|^2. \tag{4.9}$$

The value of the raw error depends on the uncertainty of the ground-truth data. If all conformations in the collection have resolved 3D coordinates for all protein residues, then $w_i = 1$, $\forall i = 1, 2, ..., L$ and the maximum error is $SSE_{max} = \frac{1}{L}\sum_{i=1}^{L} \|\vec{x}_{il}^{\text{GT-trans}}\|^2 = \frac{L}{L} = 1$. As uncertainty in the ground-truth data increases, the associated errors will become smaller. To ensure a fair assessment of the predictions across proteins, we normalise the raw errors,

$$NSSE = \frac{SSE}{SSE_{max}}. \tag{4.10}$$

Estimation of sum-of-squares errors for random vectors

To compare SeaMoon results with a random baseline, we selected 14 ground-truth principal components from the test set. We focused on proteins with maximum confidence, *i.e.*, for which $w_i = 1$, $\forall i = 1, 2, ..., L$. We started with a set of 10 components chosen randomly. We then added the most localised component (collectivity $\kappa = 0.06$), the most collective one ($\kappa = 0.85$), a component from the smallest protein (33 residues), and a component from the longest one (662 residues). We generated 1000 random predictions for each ground truth component and computed their sum-of-squares errors under optimal rotation and scaling.

Subspace comparison

We estimated the similarity between the $K \times 3$ subspaces spanned by SeaMoon predictions and the ground-truth principal components as their Root Mean Square

Inner Product (RMSIP) [126, 121, 122]. It is computed as an average of the normalised inner products of all the vectors in both subspaces,

$$\text{RMSIP} = \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{i=1}^{L} \frac{(\vec{x}_{ik}^{\text{GT}})^T \vec{x}_{il}^{\text{ortho}}}{\|\vec{x}_{ik}^{\text{GT}}\| \|\vec{x}_{il}^{\text{ortho}}\|} \right), \tag{4.11}$$

where $\vec{x}_{il}^{\text{ortho}}$ is obtained by orthogonalising SeaMoon predictions using the Gram–Schmidt process. This operation ensures that the RMSIP ranges from zero for mutually orthogonalising subspaces to one for identical subspaces and avoids artificially inflating the RMSIP due to redundancy in the predicted motions. We should stress that in practice, this redundancy is limited and the motions predicted for a given protein never collapse (**Fig. B.11**). A RMSIP score of 0.70 is considered an excellent correspondence while a score of 0.50 is considered fair [126].

While the RMSIP is invariant to permutations and rotations, the individual inner products, reflecting similarities between pairs of motions, are not. For interpretability purposes, we maximised these pairwise similarities through the following procedure:

1. compute the NSSE for all pairs of predictions and ground-truth principal components, under optimal rotation and scaling, as in Eq. (4.7),
2. orthogonalise the predictions in the order of their losses, from the best-matching prediction to the worst-matching one,
3. determine the optimal global rotation of the ordered set of matching ground-truth components onto the ordered set of orthogonalised predictions,
4. compute all pairwise normalised inner products and the corresponding RMSIP, and all pairwise NSSE under optimal scaling.

4.4.4 Comparison with the normal mode analysis

We compared SeaMoon performance with the physics-based unsupervised normal mode analysis (NMA) [306]. The NMA takes as input a protein 3D structure and builds an elastic network model where the nodes represent the atoms and the edges represent springs linking atoms located close to each other in 3D space. The normal modes are obtained by diagonalizing the mass-weighted Hessian matrix of the potential energy of this network. We used the highly efficient NOLB method [113] to extract the first $K = 3$ normal modes from the test protein 3D conformations. We retained only the Cα atoms, as for the principal component analysis, and

defined the edges in the elastic network using a distance cutoff of 10Å. We enhanced the elastic network dynamical potential by excluding edges corresponding to small contact areas between protein segments. We detected them as disconnected patches in the contact map using HOPMA [114]. Contrary to SeaMoon predictions, the orientation of the NMA predictions is not arbitrary and thus, we do not need to align the ground-truth components onto them.

4.4.5 Protein properties

Sequence and structure similarity

We estimated sequence similarity between train and test proteins using MMseqs2 [258] with default settings. We used TM-align (version 20220412) to perform all-to-all pairwise structural alignments between train and test protein conformations and compute TM-scores [357]. TM-score measures the topological similarity of protein structures. It ranges between 0 and 1, and a score higher than 0.5 assume roughly the same fold.

Motion contribution and collectivity

We estimate the contribution of the $L \times 3$ ground-truth principal component X_k^{GT} to the total positional variance as its normalised eigenvalue, $\frac{\lambda_k}{\sum_l \lambda_l}$. We estimate the collectivity [263, 264] of the $L \times 3$ predicted or ground-truth motion tensor X_k as,

$$\kappa(X_k) = \frac{1}{L} \exp\left(-\sum_{i=1}^{L} \sum_{j=1}^{3} X_{ijk}^2 \log X_{ijk}^2 \right), \tag{4.12}$$

with L the number of residues. If $\kappa(\mathbf{v}) = 1$, then the corresponding motion is maximally collective and has all the atomic displacements identical. In case of an extremely localised motion, where only one single atom is affected, the collectivity is minimal and equals to $1/L$.

Data and code availability

The source code and model weights of this work are freely available at `https://github.com/PhyloSofS-Team/seamoon`. The data used for development and evaluation of SeaMoon are freely available at Zenodo [358].

Chapter 5

Final words

This thesis took place during a highly dynamic period for the application of deep learning in structural biology. This is reflected by the fact that nearly half of the references in the bibliography date from 2021, the year this PhD began, or were published afterward. As mentioned several times in this manuscript, now that the problem of protein structure prediction has largely been solved, predicting their dynamics and alternative conformations naturally emerges as a next challenge. The work presented in this manuscript offers, in part, a response to this issue.

Key findings

Chapter 3 proposed a pipeline for generating collections of conformations from which linear motions are extracted. This pipeline was applied to the entire set of resolved structures available in the PDB, with clustering at different levels of identity and coverage. We observed that, in the vast majority of cases, protein motions are contained within a low-dimensional manifold. We assessed different manifold learning methods for the task of reconstructing unseen conformations on a representative benchmark and showed that classical manifold learning methods can generate accurate conformations. Both the data and the method have been made available to the community, can be easily updated with new experimental data, and are easily applicable to custom datasets. I believe these data are of great interest to all methods aiming to predict alternative conformations or deformations, whether for experimental validation or for training deep learning methods.

The findings from Chapter 3 motivated us to develop a deep learning method introduced in Chapter 4. This method predicts the linear motion manifold of a

protein from a protein sequence embedding generated by a Protein Language Model. This method uses a lightweight convolutional network to translate the embedding into probable deformation directions. It is flexible, as it can be easily retrained with different embeddings, works natively with sequences of arbitrary length, and can be configured to predict a desired number of modes. However, it faces significant limitations in terms of applicability, as the modes are predicted in a space that must be aligned with the final structure, and finding the optimal alignment proves to be non-trivial during inference. This limitation will drive future developments. Nevertheless, an alignment method has been proposed by minimizing the torque of the predicted deformation.

Future perspectives

My work will focus on overcoming the limitations of SeaMoon through architectural changes. The current idea is to design a network that predicts deformations in a local representation, invariant to rotations and translations in 3D space. The transition from the internal representation to 3D space should be unambiguous. A message-passing graph neural network, representing the protein as a graph, as described in Section 2.1, would theoretically allow predictions in a local frame at each C_α atom, while efficiently incorporating structural information, which has already proven crucial in the experiments conducted in Chapter 4, where we observed that the structure-informed ProstT5 embedding outperformed the ESM2 embedding. Transitioning to this type of architecture will also simplify the loss function, as the need for alignment optimization will no longer be required.

Bibliography

[1] L. Pauling, R. B. Corey, and H. R. Branson. "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain". In: *Proceedings of the National Academy of Sciences of the United States of America* 37.4 (Apr. 1951), pp. 205–211. ISSN: 0027-8424. DOI: 10.1073/pnas.37.4.205.

[2] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. "Stereochemistry of polypeptide chain configurations". In: *Journal of Molecular Biology* 7 (July 1963), pp. 95–99. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(63)80023-6.

[3] C. B. Anfinsen. "Principles that govern the folding of protein chains". In: *Science (New York, N.Y.)* 181.4096 (July 20, 1973), pp. 223–230. ISSN: 0036-8075. DOI: 10.1126/science.181.4096.223.

[4] Warren L DeLano et al. "Pymol: An open-source molecular graphics tool". In: *CCP4 Newsl. Protein Crystallogr* 40.1 (2002). Publisher: Citeseer, pp. 82–92.

[5] Helen M. Berman et al. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (Jan. 1, 2000), pp. 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

[6] "Crystallography: Protein Data Bank". In: *Nature New Biology* 233.42 (Oct. 1, 1971). Publisher: Nature Publishing Group, pp. 223–223. ISSN: 2058-1092. DOI: 10.1038/newbio233223b0.

[7] F. C. Bernstein et al. "The Protein Data Bank: a computer-based archival file for macromolecular structures". In: *Journal of Molecular Biology* 112.3 (May 25, 1977), pp. 535–542. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(77)80200-3.

[8] Helen Berman, Kim Henrick, and Haruki Nakamura. “Announcing the worldwide Protein Data Bank”. In: *Nature Structural Biology* 10.12 (Dec. 2003), p. 980. ISSN: 1072-8368. DOI: 10.1038/nsb1203-980.

[9] Helen Berman et al. “The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data”. In: *Nucleic Acids Research* 35 (Database issue Jan. 2007), pp. D301–303. ISSN: 1362-4962. DOI: 10.1093/nar/gkl971.

[10] Richard Van Noorden, Brendan Maher, and Regina Nuzzo. “The top 100 papers”. In: *Nature News* 514.7524 (Oct. 30, 2014). Cg_type: Nature News Section: News Feature, p. 550. DOI: 10.1038/514550a.

[11] Jodi Basner. *Impact Analysis of ”Berman HM et al., (2000), The Protein Data Bank”*. Clarivate Analytics, May 1, 2017.

[12] John D. Westbrook and Stephen K. Burley. “How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals”. In: *Structure (London, England: 1993)* 27.2 (Feb. 5, 2019), pp. 211–217. ISSN: 1878-4186. DOI: 10.1016/j.str.2018.11.007.

[13] Stephen K. Burley et al. “Impact of structural biology and the protein data bank on us fda new drug approvals of low molecular weight antineoplastic agents 2019–2023”. In: *Oncogene* 43.29 (July 2024). Publisher: Nature Publishing Group, pp. 2229–2243. ISSN: 1476-5594. DOI: 10.1038/s41388-024-03077-2.

[14] W. H. Bragg. “The Reflection of X-Rays by Crystals”. In: *Nature* 91.2280 (June 1913). Publisher: Nature Publishing Group, pp. 477–477. ISSN: 1476-4687. DOI: 10.1038/091477b0.

[15] D. C. Hodgkin. “The X-ray analysis of the structure of penicillin”. In: *Advancement of Science* 6.22 (July 1949), pp. 85–89. ISSN: 0001-866X.

[16] Dorothy Crowfoot Hodgkin et al. “Structure of Vitamin B12”. In: *Nature* 178.4524 (July 1956). Publisher: Nature Publishing Group, pp. 64–66. ISSN: 1476-4687. DOI: 10.1038/178064a0.

[17] J. C. Kendrew et al. “A three-dimensional model of the myoglobin molecule obtained by x-ray analysis”. In: *Nature* 181.4610 (Mar. 8, 1958), pp. 662–666. ISSN: 0028-0836. DOI: 10.1038/181662a0.

[18] Henry N. Chapman et al. “Femtosecond X-ray protein nanocrystallography”. In: *Nature* 470.7332 (Feb. 3, 2011), pp. 73–77. ISSN: 0028-0836. DOI: 10.1038/nature09750.

[19] Janet L. Smith, Robert F. Fischetti, and Masaki Yamamoto. “Micro-crystallography comes of age”. In: *Current opinion in structural biology* 22.5 (Oct. 2012), pp. 602–612. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2012.09.001.

[20] M S Smyth and J H J Martin. “x Ray crystallography”. In: *Molecular Pathology* 53.1 (Feb. 2000), pp. 8–14. ISSN: 1366-8714.

[21] Robert Huber. “Conformational flexibility and its functional significance in some protein molecules”. In: *Trends in Biochemical Sciences* 4.12 (Dec. 1, 1979), pp. 271–276. ISSN: 0968-0004. DOI: 10.1016/0968-0004(79)90298-6.

[22] Dagmar Ringe and Gregory A. Petsko. “Study of protein dynamics by X-ray diffraction”. In: *Methods in Enzymology*. Enzyme Structure Part L 131 (Jan. 1, 1986), pp. 389–433. DOI: 10.1016/0076-6879(86)31050-4.

[23] M. Vihinen. “Relationship of protein flexibility to thermostability”. In: *Protein Engineering* 1.6 (Dec. 1987), pp. 477–480. ISSN: 0269-2139. DOI: 10.1093/protein/1.6.477.

[24] O. Carugo and P. Argos. “Reliability of atomic displacement parameters in protein crystal structures”. In: *Acta Crystallographica. Section D, Biological Crystallography* 55 (Pt 2 Feb. 1999), pp. 473–478. ISSN: 0907-4449. DOI: 10.1107/s0907444998011688.

[25] Wayne A. Hendrickson. “Radiation damage in protein crystallography”. In: *Journal of Molecular Biology* 106.3 (Sept. 25, 1976), pp. 889–893. ISSN: 0022-2836. DOI: 10.1016/0022-2836(76)90271-0.

[26] J. R. Helliwell. “Protein crystal perfection and the nature of radiation damage”. In: *Journal of Crystal Growth* 90.1 (July 2, 1988), pp. 259–272. ISSN: 0022-0248. DOI: 10.1016/0022-0248(88)90322-3.

[27] R. Frisch and O. Stern. “Über die magnetische Ablenkung von Wasserstoffmolekülen und das magnetische Moment des Protons. I”. In: *Zeitschrift für Physik* 85.1 (Jan. 1, 1933), pp. 4–16. ISSN: 0044-3328. DOI: 10.1007/BF01330773.

[28] I. Estermann and O. Stern. "Über die magnetische Ablenkung von Wasserstoffmolekülen und das magnetische Moment des Protons. II". In: *Zeitschrift für Physik* 85.1 (Jan. 1, 1933), pp. 17–24. ISSN: 0044-3328. DOI: 10.1007/BF01330774.

[29] F. Bloch. "Nuclear Induction". In: *Physical Review* 70.7 (Oct. 1, 1946). Publisher: American Physical Society, pp. 460–474. DOI: 10.1103/PhysRev.70.460.

[30] E. M. Purcell, H. C. Torrey, and R. V. Pound. "Resonance Absorption by Nuclear Magnetic Moments in a Solid". In: *Physical Review* 69 (Jan. 1, 1946). Publisher: APS ADS Bibcode: 1946PhRv...69...37P, pp. 37–38. ISSN: 1536-6065. DOI: 10.1103/PhysRev.69.37.

[31] R. C. Ferguson and W. D. Phillips. "High-Resolution Nuclear Magnetic Resonance Spectroscopy". In: *Science* 157.3786 (July 21, 1967). Publisher: American Association for the Advancement of Science, pp. 257–267. DOI: 10.1126/science.157.3786.257.

[32] Jean Jeener et al. "Investigation of Exchange Process by Two-Dimensional NMR Spectroscopy". In: *The Journal of Chemical Physics* 71 (Dec. 1, 1979), pp. 4546–4553. DOI: 10.1063/1.438208.

[33] Kurt Wüthrich. "NMR with Proteins and Nucleic Acids". In: *Europhysics News* 17.1 (1986), pp. 11–13. ISSN: 0531-7479, 1432-1092. DOI: 10.1051/epn/19861701011.

[34] Natalie K Goto and Lewis E Kay. "New developments in isotope labeling strategies for protein solution NMR spectroscopy". In: *Current Opinion in Structural Biology* 10.5 (Oct. 2000), pp. 585–592. ISSN: 0959440X. DOI: 10.1016/S0959-440X(00)00135-4.

[35] Yajun Jiang and Charalampos G. Kalodimos. "NMR Studies of Large Proteins". In: *Journal of Molecular Biology*. John Kendrew's 100th Anniversary Special Edition 429.17 (Aug. 18, 2017), pp. 2667–2676. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2017.07.007.

[36] Lewis E. Kay et al. "Three-dimensional triple-resonance NMR Spectroscopy of isotopically enriched proteins". In: *Journal of Magnetic Resonance*. Magnetic Moments 213.2 (Feb. 22, 1990), pp. 423–441. ISSN: 1090-7807. DOI: 10.1016/j.jmr.2011.09.004.

[37] K. H. Gardner and L. E. Kay. "The use of 2H, 13C, 15N multidimensional NMR to study the structure and dynamics of proteins". In: *Annual Review of Biophysics and Biomolecular Structure* 27 (1998), pp. 357–406. ISSN: 1056-8700. DOI: 10.1146/annurev.biophys.27.1.357.

[38] Konstantin Pervushin et al. "Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution". In: *Proceedings of the National Academy of Sciences* 94.23 (Nov. 11, 1997). Publisher: Proceedings of the National Academy of Sciences, pp. 12366–12371. DOI: 10.1073/pnas.94.23.12366.

[39] Remco Sprangers and Lewis E. Kay. "Quantitative dynamics and binding studies of the 20S proteasome by NMR". In: *Nature* 445.7128 (Feb. 2007). Publisher: Nature Publishing Group, pp. 618–622. ISSN: 1476-4687. DOI: 10.1038/nature05512.

[40] L. E. Kay, D. A. Torchia, and A. Bax. "Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease". In: *Biochemistry* 28.23 (Nov. 14, 1989), pp. 8972–8979. ISSN: 0006-2960. DOI: 10.1021/bi00449a003.

[41] Rieko Ishima and Dennis A. Torchia. "Protein dynamics from NMR". In: *Nature Structural Biology* 7.9 (Sept. 2000). Publisher: Nature Publishing Group, pp. 740–743. ISSN: 1545-9985. DOI: 10.1038/78963.

[42] Arthur G. III Palmer. "NMR Characterization of the Dynamics of Biomacromolecules". In: *Chemical Reviews* 104.8 (Aug. 1, 2004). Publisher: American Chemical Society, pp. 3623–3640. ISSN: 0009-2665. DOI: 10.1021/cr030413t.

[43] Anthony K. Mittermaier and Lewis E. Kay. "Observing biological dynamics at atomic resolution using NMR". In: *Trends in Biochemical Sciences* 34.12 (Dec. 1, 2009), pp. 601–611. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2009.07.004.

[44] Louis de Broglie. "Recherches sur la théorie des Quanta". PhD thesis. Migration - université en cours d'affectation, Nov. 25, 1924.

[45] C. Davisson and L. H. Germer. "The Scattering of Electrons by a Single Crystal of Nickel". In: *Nature* 119.2998 (Apr. 1927). Publisher: Nature Publishing Group, pp. 558–560. ISSN: 1476-4687. DOI: 10.1038/119558a0.

[46] C. Davisson and L. H. Germer. “Diffraction of Electrons by a Crystal of Nickel”. In: *Physical Review* 30.6 (Dec. 1, 1927). Publisher: American Physical Society, pp. 705–740. DOI: 10.1103/PhysRev.30.705.

[47] G. P. Thomson and A. Reid. “Diffraction of Cathode Rays by a Thin Film”. In: *Nature* 119.3007 (June 1927). Publisher: Nature Publishing Group, pp. 890–890. ISSN: 1476-4687. DOI: 10.1038/119890a0.

[48] M. Knoll and E. Ruska. “Das Elektronenmikroskop”. In: *Zeitschrift für Physik* 78.5 (May 1, 1932), pp. 318–339. ISSN: 0044-3328. DOI: 10.1007/BF01342199.

[49] E. Ruska. “Die elektronenmikroskopische Abbildung elektronenbestrahlter Oberflächen”. In: *Zeitschrift für Physik* 83.7 (July 1, 1933), pp. 492–497. ISSN: 0044-3328. DOI: 10.1007/BF01338960.

[50] E. Ruska. “Über ein magnetisches Objektiv für das Elektronenmikroskop”. In: *Zeitschrift für Physik* 89.1 (Jan. 1, 1934), pp. 90–128. ISSN: 0044-3328. DOI: 10.1007/BF01333236.

[51] L. Marton. “Electron Microscopy of Biological Objects”. In: *Physical Review* 46.6 (Sept. 15, 1934). Publisher: American Physical Society, pp. 527–528. DOI: 10.1103/PhysRev.46.527.

[52] G. E. Palade. “A small particulate component of the cytoplasm”. In: *The Journal of Biophysical and Biochemical Cytology* 1.1 (Jan. 1955), pp. 59–68. ISSN: 0095-9901. DOI: 10.1083/jcb.1.1.59.

[53] J. Dubochet and A.w. McDowall. “Vitrification of Pure Water for Electron Microscopy”. In: *Journal of Microscopy* 124.3 (1981), pp. 3–4. ISSN: 1365-2818. DOI: 10.1111/j.1365-2818.1981.tb02483.x.

[54] J. Dubochet et al. “Electron microscopy of frozen water and aqueous solutions”. In: *Journal of Microscopy* 128.3 (1982), pp. 219–237. ISSN: 1365-2818. DOI: 10.1111/j.1365-2818.1982.tb04625.x.

[55] J. Lepault, F. P. Booy, and J. Dubochet. “Electron microscopy of frozen biological suspensions”. In: *Journal of Microscopy* 129 (Pt 1 Jan. 1983), pp. 89–102. ISSN: 0022-2720. DOI: 10.1111/j.1365-2818.1983.tb04163.x.

[56] Xiao-chen Bai, Greg McMullan, and Sjors H. W. Scheres. “How cryo-EM is revolutionizing structural biology”. In: *Trends in Biochemical Sciences* 40.1 (Jan. 1, 2015). Publisher: Elsevier, pp. 49–57. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2014.10.005.

[57] Takanori Nakane et al. “Single-particle cryo-EM at atomic resolution”. In: *Nature* 587.7832 (Nov. 2020). Publisher: Nature Publishing Group, pp. 152–156. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2829-0.

[58] Ka Man Yip et al. “Atomic-resolution protein structure determination by cryo-EM”. In: *Nature* 587.7832 (Nov. 2020). Publisher: Nature Publishing Group, pp. 157–161. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2833-4.

[59] Slavica Jonić. “Computational methods for analyzing conformational variability of macromolecular complexes from cryo-electron microscopy images”. In: *Current Opinion in Structural Biology*. Theory and simulation • Macromolecular assemblies 43 (Apr. 1, 2017), pp. 114–121. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2016.12.011.

[60] M. Karplus and J. A. McCammon. “Dynamics of proteins: elements and function”. In: *Annual Review of Biochemistry* 52 (1983), pp. 263–300. ISSN: 0066-4154. DOI: 10.1146/annurev.bi.52.070183.001403.

[61] J. A. McCammon. “Protein dynamics”. In: *Reports on Progress in Physics* 47.1 (Jan. 1984), p. 1. ISSN: 0034-4885. DOI: 10.1088/0034-4885/47/1/001.

[62] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. “The energy landscapes and motions of proteins”. In: *Science (New York, N.Y.)* 254.5038 (Dec. 13, 1991), pp. 1598–1603. ISSN: 0036-8075. DOI: 10.1126/science.1749933.

[63] Katherine Henzler-Wildman and Dorothee Kern. “Dynamic personalities of proteins”. In: *Nature* 450.7172 (Dec. 2007). Publisher: Nature Publishing Group, pp. 964–972. ISSN: 1476-4687. DOI: 10.1038/nature06522.

[64] D. E. Koshland. “Application of a Theory of Enzyme Specificity to Protein Synthesis*”. In: *Proceedings of the National Academy of Sciences of the United States of America* 44.2 (Feb. 1958), pp. 98–104. ISSN: 0027-8424.

[65] Ahmet Bakan and Ivet Bahar. “The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding”. In: *Proceedings of the National Academy of Sciences* 106.34 (Aug. 25, 2009). Publisher: Proceedings of the National Academy of Sciences, pp. 14349–14354. DOI: 10.1073/pnas.0904214106.

[66] Gregorio Weber. “Ligand binding and internal equilibiums in proteins”. In: *Biochemistry* 11.5 (Feb. 29, 1972). Publisher: American Chemical Society, pp. 864–878. ISSN: 0006-2960. DOI: 10.1021/bi00755a028.

[67] Yan Zhang et al. “Intrinsic dynamics is evolutionarily optimized to enable allosteric behavior”. In: *Current Opinion in Structural Biology* 62 (June 2020), pp. 14–21. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2019.11.002.

[68] David D. Boehr, Ruth Nussinov, and Peter E. Wright. “The role of dynamic conformational ensembles in biomolecular recognition”. In: *Nature Chemical Biology* 5.11 (Nov. 2009), pp. 789–796. ISSN: 1552-4469. DOI: 10.1038/nchembio.232.

[69] Francis Gaudreault, Matthieu Chartier, and Rafael Najmanovich. “Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding”. In: *Bioinformatics* 28.18 (Sept. 15, 2012), pp. i423–i430. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts395.

[70] T Alber et al. “On the three-dimensional structure and catalytic mechanism of triose phosphate isomerase”. In: *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 293.1063 (June 1, 1981), pp. 159–171. ISSN: 1471-2970. DOI: 10.1098/rstb.1981.0069.

[71] Qinghua Liao et al. “Loop Motion in Triosephosphate Isomerase Is Not a Simple Open and Shut Case”. In: *Journal of the American Chemical Society* 140.46 (Nov. 21, 2018). Publisher: American Chemical Society, pp. 15889–15903. ISSN: 0002-7863. DOI: 10.1021/jacs.8b09378.

[72] Jack F. Kirsch et al. “Mechanism of action of aspartate aminotransferase proposed on the basis of its spatial structure”. In: *Journal of Molecular Biology* 174.3 (Apr. 15, 1984), pp. 497–525. ISSN: 0022-2836. DOI: 10.1016/0022-2836(84)90333-4.

[73] A. Keith Dunker et al. “What’s in a name? Why these proteins are intrinsically disordered”. In: *Intrinsically Disordered Proteins* 1.1 (Apr. 1, 2013), e24157. ISSN: 2169-0693. DOI: 10.4161/idp.24157.

[74] A. Keith Dunker et al. “Intrinsically disordered protein”. In: *Journal of Molecular Graphics and Modelling* 19.1 (Feb. 1, 2001), pp. 26–59. ISSN: 1093-3263. DOI: 10.1016/S1093-3263(00)00138-8.

[75] Alexander Cumberworth et al. “Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes”. In: *The Biochemical Journal* 454.3 (Sept. 15, 2013), pp. 361–369. ISSN: 1470-8728. DOI: 10.1042/BJ20130545.

[76] David S.W. Protter et al. “Intrinsically Disordered Regions Can Contribute Promiscuous Interactions to RNP Granule Assembly”. In: *Cell reports* 22.6 (Feb. 6, 2018), pp. 1401–1412. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2018.01.036.

[77] Pinak Chakrabarti and Devlina Chakravarty. “Intrinsically disordered proteins/regions and insight into their biomolecular interactions”. In: *Biophysical Chemistry* 283 (Apr. 2022), p. 106769. ISSN: 1873-4200. DOI: 10.1016/j.bpc.2022.106769.

[78] Peter Tompa. “Intrinsically unstructured proteins”. In: *Trends in Biochemical Sciences* 27.10 (Oct. 2002), pp. 527–533. ISSN: 0968-0004. DOI: 10.1016/s0968-0004(02)02169-2.

[79] Peter Tompa. “The interplay between structure and function in intrinsically unstructured proteins”. In: *FEBS Letters*. Budapest Special Issue 579.15 (June 13, 2005), pp. 3346–3354. ISSN: 0014-5793. DOI: 10.1016/j.febslet.2005.03.072.

[80] Samrat Mukhopadhyay. “The Dynamism of Intrinsically Disordered Proteins: Binding-Induced Folding, Amyloid Formation, and Phase Separation”. In: *The Journal of Physical Chemistry. B* 124.51 (Dec. 24, 2020), pp. 11541–11560. ISSN: 1520-5207. DOI: 10.1021/acs.jpcb.0c07598.

[81] H. Jane Dyson and Peter E. Wright. “Coupling of folding and binding for unstructured proteins”. In: *Current Opinion in Structural Biology* 12.1 (Feb. 2002), pp. 54–60. ISSN: 0959-440X. DOI: 10.1016/s0959-440x(02)00289-0.

[82] Frederik Lermyte. “Roles, Characteristics, and Analysis of Intrinsically Disordered Proteins: A Minireview”. In: *Life* 10.12 (Nov. 30, 2020), p. 320. ISSN: 2075-1729. DOI: 10.3390/life10120320.

[83] Gregory Petsko et al. *Protein Structure and Function*. Primers in Biology. Oxford, New York: Oxford University Press, May 29, 2008. 224 pp. ISBN: 978-0-19-955684-7.

[84] Ivet Bahar, Chakra Chennubhotla, and Dror Tobi. “Intrinsic Enzyme Dynamics in the Unbound State and Relation to Allosteric Regulation”. In: *Current opinion in structural biology* 17.6 (Dec. 2007), pp. 633–640. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2007.09.011.

[85] Ulrich Weininger et al. “Dynamics of Aromatic Side Chains in the Active Site of FKBP12”. In: *Biochemistry* 56.1 (Jan. 10, 2017). Publisher: American Chemical Society, pp. 334–343. ISSN: 0006-2960. DOI: 10.1021/acs.biochem.6b01157.

[86] N. Go, T. Noguti, and T. Nishikawa. “Dynamics of a small globular protein in terms of low-frequency vibrational modes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 80.12 (June 1983), pp. 3696–3700. ISSN: 0027-8424. DOI: 10.1073/pnas.80.12.3696.

[87] B. Brooks and M. Karplus. “Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor”. In: *Proceedings of the National Academy of Sciences of the United States of America* 80.21 (Nov. 1983), pp. 6571–6575. ISSN: 0027-8424. DOI: 10.1073/pnas.80.21.6571.

[88] Bernard Brooks and Martin Karplus. “Normal Modes for Specific Motions of Macromolecules: Application to the Hinge-Bending Mode of Lysozyme”. In: *Proceedings of the National Academy of Sciences of the United States of America* 82.15 (1985). Publisher: National Academy of Sciences, pp. 4995–4999. ISSN: 0027-8424.

[89] Qiang Cui et al. “A Normal Mode Analysis of Structural Plasticity in the Biomolecular Motor F1-ATPase”. In: *Journal of Molecular Biology* 340.2 (July 2, 2004), pp. 345–372. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2004.04.044.

[90] W. G. Krebs et al. “Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic”. In: *Proteins: Structure, Function, and Bioinformatics* 48.4 (2002). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.10168, pp. 682–695. ISSN: 1097-0134. DOI: 10.1002/prot.10168.

[91] Jianpeng Ma and Martin Karplus. “Ligand-induced conformational changes in *ras* p21: a normal mode and energy minimization analysis1”. In: *Journal of Molecular Biology* 274.1 (Nov. 21, 1997), pp. 114–131. ISSN: 0022-2836. DOI: 10.1006/jmbi.1997.1313.

[92] Dengming Ming et al. “How to describe protein motion without amino acid sequence and atomic coordinates”. In: *Proceedings of the National Academy of Sciences* 99.13 (June 25, 2002). Publisher: Proceedings of the National Academy of Sciences, pp. 8620–8625. DOI: 10.1073/pnas.082148899.

[93] Dengming Ming et al. “Simulation of F-Actin Filaments of Several Microns”. In: *Biophysical Journal* 85.1 (July 2003), pp. 27–35. ISSN: 00063495. DOI: 10.1016/S0006-3495(03)74451-8.

[94] Yasunobu Seno and Nobuhiro Gō. “Deoxymyoglobin studied by the conformational normal mode analysis: II. The conformational change upon oxygenation”. In: *Journal of Molecular Biology* 216.1 (Nov. 5, 1990), pp. 111–126. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80064-6.

[95] Yasunobu Seno and Nobuhiro Gō. “Deoxymyoglobin studied by the conformational normal mode analysis: I. Dynamics of globin and the heme-globin interaction”. In: *Journal of Molecular Biology* 216.1 (Nov. 5, 1990), pp. 95–109. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80063-4.

[96] Aline Thomas et al. “Analysis of the Low Frequency Normal Modes of the T-state of Aspartate Transcarbamylase”. In: *Journal of Molecular Biology* 257.5 (Apr. 19, 1996), pp. 1070–1087. ISSN: 0022-2836. DOI: 10.1006/jmbi.1996.0224.

[97] Aline Thomas et al. “Tertiary and quaternary conformational changes in aspartate transcarbamylase: a normal mode study”. In: *Proteins: Structure, Function, and Bioinformatics* 34.1 (1999), pp. 96–112. ISSN: 1097-0134. DOI: 10.1002/(SICI)1097-0134(19990101)34:1<96::AID-PROT8>3.0.CO;2-0.

[98] Ivet Bahar et al. “Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins”. In: *Chemical Reviews* 110.3 (Mar. 10, 2010). Publisher: American Chemical Society, pp. 1463–1497. ISSN: 0009-2665. DOI: 10.1021/cr900095e.

[99] Wenjun Zheng, Bernard R. Brooks, and D. Thirumalai. “Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations”. In: *Proceedings of the National Academy of Sciences* 103.20 (May 16, 2006). Publisher: Proceedings of the National Academy of Sciences, pp. 7664–7669. DOI: 10.1073/pnas.0510426103.

[100] Sandra Maguid, Sebastian Fernandez-Alberti, and Julian Echave. “Evolutionary conservation of protein vibrational dynamics”. In: *Gene*. Physical and Chemical Foundations of Bioinformatics Methods 422.1 (Oct. 1, 2008), pp. 7–13. ISSN: 0378-1119. DOI: 10.1016/j.gene.2008.06.002.

[101] Sandhya P Tiwari and Nathalie Reuter. "Conservation of intrinsic dynamics in proteins — what have computational models taught us?" In: *Current Opinion in Structural Biology*. Carbohydrates • Sequences and topology 50 (June 1, 2018), pp. 75–81. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2017.12.001.

[102] Alessandro Pandini et al. "Detecting similarities among distant homologous proteins by comparison of domain flexibilities". In: *Protein Engineering, Design and Selection* 20.6 (June 1, 2007), pp. 285–299. ISSN: 1741-0126. DOI: 10.1093/protein/gzm021.

[103] Philippe Durand, Georges Trinquier, and Yves-Henri Sanejouand. "A new approach for determining low-frequency normal modes in macromolecules". In: *Biopolymers* 34.6 (1994), pp. 759–771. ISSN: 1097-0282. DOI: 10.1002/bip.360340608.

[104] F. Tama et al. "Building-block approach for determining low-frequency normal modes of macromolecules". In: *Proteins* 41.1 (Oct. 1, 2000), pp. 1–7. ISSN: 0887-3585. DOI: 10.1002/1097-0134(20001001)41:1<1::aid-prot10>3.0.co;2-p.

[105] Monique M. Tirion. "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis". In: *Physical Review Letters* 77.9 (Aug. 26, 1996). Publisher: American Physical Society, pp. 1905–1908. DOI: 10.1103/PhysRevLett.77.1905.

[106] A R Atilgan et al. "Anisotropy of fluctuation dynamics of proteins with an elastic network model." In: *Biophysical Journal* 80.1 (Jan. 2001), pp. 505–515. ISSN: 0006-3495.

[107] Guang Song and Robert L. Jernigan. "An enhanced elastic network model to represent the motions of domain-swapped proteins". In: *Proteins* 63.1 (Apr. 1, 2006), pp. 197–209. ISSN: 1097-0134. DOI: 10.1002/prot.20836.

[108] Lei Yang, Guang Song, and Robert L. Jernigan. "How well can we understand large-scale protein motions using normal modes of elastic network models?" In: *Biophysical Journal* 93.3 (Aug. 1, 2007), pp. 920–929. ISSN: 0006-3495. DOI: 10.1529/biophysj.106.095927.

[109] Turkan Haliloglu, Ivet Bahar, and Burak Erman. "Gaussian Dynamics of Folded Proteins". In: *Physical Review Letters* 79.16 (Oct. 20, 1997). Publisher: American Physical Society, pp. 3090–3093. DOI: 10.1103/PhysRevLett.79.3090.

[110] I. Bahar, A. R. Atilgan, and B. Erman. "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential". In: *Folding & Design* 2.3 (1997), pp. 173–181. ISSN: 1359-0278. DOI: 10.1016/S1359-0278(97)00024-2.

[111] T. Haliloglu and I. Bahar. "Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data". In: *Proteins* 37.4 (Dec. 1, 1999), pp. 654–667. ISSN: 0887-3585. DOI: 10.1002/(sici)1097-0134(19991201)37:4<654::aid-prot15>3.0.co;2-j.

[112] Sibsankar Kundu et al. "Dynamics of proteins in crystals: comparison of experiment with simple models". In: *Biophysical Journal* 83.2 (Aug. 2002), pp. 723–732. ISSN: 0006-3495. DOI: 10.1016/S0006-3495(02)75203-X.

[113] Alexandre Hoffmann and Sergei Grudinin. "NOLB: Nonlinear Rigid Block Normal-Mode Analysis Method". In: *Journal of Chemical Theory and Computation* 13.5 (May 9, 2017), pp. 2123–2134. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/acs.jctc.7b00197.

[114] Elodie Laine and Sergei Grudinin. "HOPMA: Boosting Protein Functional Dynamics with Colored Contact Maps". In: *The Journal of Physical Chemistry B* 125.10 (Mar. 18, 2021). Publisher: American Chemical Society, pp. 2577–2588. ISSN: 1520-6106. DOI: 10.1021/acs.jpcb.0c11633.

[115] Jay W. Ponder and David A. Case. "Force Fields for Protein Simulations". In: *Advances in Protein Chemistry*. Vol. 66. Protein Simulations. Academic Press, Jan. 1, 2003, pp. 27–85. DOI: 10.1016/S0065-3233(03)66002-X.

[116] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. "Dynamics of folded proteins". In: *Nature* 267.5612 (June 1977). Publisher: Nature Publishing Group, pp. 585–590. ISSN: 1476-4687. DOI: 10.1038/267585a0.

[117] Mayar Ahmed, Alex M. Maldonado, and Jacob D. Durrant. "From Byte to Bench to Bedside: Molecular Dynamics Simulations and Drug Discovery". In: *ArXiv* (Nov. 28, 2023), arXiv:2311.16946v1. ISSN: 2331-8422.

[118] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (Nov. 1, 1901). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786440109462720, pp. 559–572. ISSN: 1941-5982. DOI: 10.1080/14786440109462720.

[119] Angel E. García. "Large-amplitude nonlinear motions in proteins". In: *Physical Review Letters* 68.17 (Apr. 27, 1992). Publisher: American Physical Society, pp. 2696–2699. DOI: 10.1103/PhysRevLett.68.2696.

[120] A. Amadei, A. B. Linssen, and H. J. Berendsen. "Essential dynamics of proteins". In: *Proteins* 17.4 (Dec. 1993), pp. 412–425. ISSN: 0887-3585. DOI: 10.1002/prot.340170408.

[121] Alejandra Leo-Macias et al. "An Analysis of Core Deformations in Protein Superfamilies". In: *Biophysical Journal* 88.2 (Feb. 2005), pp. 1291–1299. ISSN: 0006-3495. DOI: 10.1529/biophysj.104.052449.

[122] Charles C. David and Donald J. Jacobs. "Characterizing Protein Motions from Structure". In: *Journal of molecular graphics & modelling* 31 (Nov. 2011), pp. 41–56. ISSN: 1093-3263. DOI: 10.1016/j.jmgm.2011.08.004.

[123] Miguel L. Teodoro, George N. Phillips, and Lydia E. Kavraki. "Understanding protein flexibility through dimensionality reduction". In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 10.3 (2003), pp. 617–634. ISSN: 1066-5277. DOI: 10.1089/10665270360688228.

[124] Lei Yang et al. "Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes". In: *Structure* 16.2 (2008). Publisher: Elsevier, pp. 321–330.

[125] Lee-Wei Yang et al. "Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics". In: *Bioinformatics* 25.5 (Mar. 1, 2009), pp. 606–614. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp023.

[126] Andrea Amadei, Marc A. Ceruso, and Alfredo Di Nola. "On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations". In: *Proteins: Structure, Function, and Bioinformatics* 36.4 (1999), pp. 419–424. ISSN: 1097-0134. DOI: 10.1002/(SICI)1097-0134(19990901)36:4<419::AID-PROT5>3.0.CO;2-U.

[127] Frank Rosenblatt. *Principles of neurodynamics; perceptrons and the theory of brain mechanisms.* Washington: Spartan Books, 1962. 616 pp.

[128] Shunichi Amari. "A Theory of Adaptive Pattern Classifiers". In: *IEEE Transactions on Electronic Computers* EC-16.3 (June 1967). Conference Name: IEEE Transactions on Electronic Computers, pp. 299–307. ISSN: 0367-7508. DOI: 10.1109/PGEC.1967.264666.

[129] Seppo Linnainmaa. “The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors”. PhD thesis. Master’s Thesis (in Finnish), Univ. Helsinki, 1970.

[130] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (Oct. 1986). Publisher: Nature Publishing Group, pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0.

[131] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[132] Kunihiko Fukushima. “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. In: *Biological Cybernetics* 36.4 (Apr. 1, 1980), pp. 193–202. ISSN: 1432-0770. DOI: 10.1007/BF00344251.

[133] Y. LeCun et al. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Computation* 1.4 (Dec. 1989). Conference Name: Neural Computation, pp. 541–551. ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.4.541.

[134] Justin Gilmer et al. “Neural Message Passing for Quantum Chemistry”. In: *Proceedings of the 34th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, July 17, 2017, pp. 1263–1272.

[135] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[136] Alexander Rives et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. In: *Proceedings of the National Academy of Sciences* 118.15 (Apr. 13, 2021). Publisher: Proceedings of the National Academy of Sciences, e2016239118. DOI: 10.1073/pnas.2016239118.

[137] Ali Madani et al. “Large language models generate functional protein sequences across diverse families”. In: *Nature Biotechnology* 41.8 (Aug. 2023). Publisher: Nature Publishing Group, pp. 1099–1106. ISSN: 1546-1696. DOI: 10.1038/s41587-022-01618-2.

[138] Zeming Lin et al. *Evolutionary-scale prediction of atomic level protein structure with a language model*. Pages: 2022.07.20.500902 Section: New Results. Dec. 21, 2022. DOI: 10.1101/2022.07.20.500902.

[139] Nadav Brandes et al. “ProteinBERT: a universal deep-learning model of protein sequence and function”. In: *Bioinformatics* 38.8 (Apr. 12, 2022). Ed. by Pier Luigi Martelli, pp. 2102–2110. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btac020.

[140] Ahmed Elnaggar et al. “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2022), pp. 7112–7127. DOI: 10.1109/TPAMI.2021.3095381.

[141] Lei Wang et al. “Deciphering the protein landscape with ProtFlash, a lightweight language model”. In: *Cell Reports Physical Science* 4.10 (Oct. 18, 2023), p. 101600. ISSN: 2666-3864. DOI: 10.1016/j.xcrp.2023.101600.

[142] Modestas Filipavicius et al. *Pre-training Protein Language Models with Label-Agnostic Binding Pairs Enhances Performance in Downstream Tasks*. Dec. 5, 2020. arXiv: 2012.03084[cs,q-bio].

[143] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. “ProtGPT2 is a deep unsupervised language model for protein design”. In: *Nature Communications* 13 (July 27, 2022), p. 4348. ISSN: 2041-1723. DOI: 10.1038/s41467-022-32007-7.

[144] Philip Gage. “A new algorithm for data compression”. In: *C Users J.* 12.2 (Feb. 1, 1994), pp. 23–38. ISSN: 0898-9788.

[145] Jeffrey Pennington, Richard Socher, and Christopher Manning. “Glove: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[146] Mor Geva et al. *Transformer Feed-Forward Layers Are Key-Value Memories*. Sept. 5, 2021. arXiv: 2012.14913[cs].

[147] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[148] Baris E Suzek et al. “UniRef: comprehensive and non-redundant UniProt reference clusters”. In: *Bioinformatics* 23.10 (2007). Publisher: Oxford University Press, pp. 1282–1288.

[149] Baris E Suzek et al. "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches". In: *Bioinformatics* 31.6 (2015). Publisher: Oxford University Press, pp. 926–932.

[150] Joshua Meier et al. *Language models enable zero-shot prediction of the effects of mutations on protein function*. Pages: 2021.07.09.450648 Section: New Results. Nov. 17, 2021. DOI: 10.1101/2021.07.09.450648.

[151] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. July 26, 2019. DOI: 10.48550/arXiv.1907.11692. arXiv: 1907.11692[cs].

[152] Roshan Rao et al. *MSA Transformer*. Pages: 2021.02.12.430858 Section: New Results. Aug. 27, 2021. DOI: 10.1101/2021.02.12.430858.

[153] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (Aug. 2021). Publisher: Nature Publishing Group, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.

[154] Zhenzhong Lan et al. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. Feb. 8, 2020. DOI: 10.48550/arXiv.1909.11942. arXiv: 1909.11942[cs].

[155] Kevin Clark et al. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*. arXiv.org. Mar. 23, 2020. URL: https://arxiv.org/abs/2003.10555v1 (visited on 09/27/2024).

[156] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140 (2020), pp. 1–67.

[157] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: ().

[158] Ali Madani et al. *ProGen: Language Modeling for Protein Generation*. arXiv.org. Mar. 8, 2020. URL: https://arxiv.org/abs/2004.03497v1 (visited on 09/27/2024).

[159] Erik Nijkamp et al. *ProGen2: Exploring the Boundaries of Protein Language Models*. arXiv.org. June 27, 2022. URL: https://arxiv.org/abs/2206.13517v1 (visited on 09/26/2024).

[160] Zihang Dai et al. *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. arXiv.org. Jan. 9, 2019. URL: https://arxiv.org/abs/1901.02860v3 (visited on 09/27/2024).

[161] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Jan. 2, 2020. DOI: 10.48550/arXiv.1906.08237. arXiv: 1906.08237[cs].

[162] Michael Heinzinger et al. "ProstT5: Bilingual language model for protein sequence and structure". In: *bioRxiv* (2023). Publisher: Cold Spring Harbor Laboratory, pp. 2023–07.

[163] Michel Van Kempen et al. "Fast and accurate protein structure search with Foldseek". In: *Nature Biotechnology* 42.2 (2024). Publisher: Nature Publishing Group US New York, pp. 243–246.

[164] Jin Su et al. *SaProt: Protein Language Modeling with Structure-aware Vocabulary*. Pages: 2023.10.01.560349 Section: New Results. Oct. 2, 2023. DOI: 10.1101/2023.10.01.560349.

[165] Zuobai Zhang et al. *A Systematic Study of Joint Representation Learning on Protein Sequences and Structures*. Oct. 18, 2023. arXiv: 2303.06275[cs,q-bio].

[166] Zuobai Zhang et al. *Protein Representation Learning by Geometric Structure Pretraining*. arXiv.org. Mar. 11, 2022. URL: https://arxiv.org/abs/2203.06125v5 (visited on 09/27/2024).

[167] Thomas Hayes et al. *Simulating 500 million years of evolution with a language model*. Pages: 2024.07.01.600583 Section: New Results. July 2, 2024. DOI: 10.1101/2024.07.01.600583.

[168] Mihaly Varadi et al. "3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources". In: *GigaScience* 11 (Jan. 1, 2022), giac118. ISSN: 2047-217X. DOI: 10.1093/gigascience/giac118.

[169] John Moult et al. "A large-scale experiment to assess protein structure prediction methods". In: (Nov. 1, 1995). DOI: 10.1002/prot.340230303.

[170] Andriy Kryshtafovych et al. "Breaking the conformational ensemble barrier: Ensemble structure modeling challenges in CASP15". In: *Proteins* 91.12 (Dec. 2023), pp. 1903–1911. ISSN: 1097-0134. DOI: 10.1002/prot.26584.

[171] Joerg Schaarschmidt et al. "Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age". In: *Proteins: Structure, Function, and Bioinformatics* 86 (S1 2018), pp. 51–66. ISSN: 1097-0134. DOI: 10.1002/prot.25407.

[172] Elodie Laine et al. “Protein sequence-to-structure learning: Is this the end(-to-end revolution)?” In: *Proteins: Structure, Function, and Bioinformatics* 89.12 (2021), pp. 1770–1786. ISSN: 1097-0134. DOI: 10.1002/prot.26235.

[173] Andrew W. Senior et al. “Improved protein structure prediction using potentials from deep learning”. In: *Nature* 577.7792 (Jan. 2020). Publisher: Nature Publishing Group, pp. 706–710. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1923-7.

[174] Thomas J. Lane. “Protein structure prediction has reached the single-structure frontier”. In: *Nature Methods* 20.2 (Feb. 2023). Number: 2 Publisher: Nature Publishing Group, pp. 170–173. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01760-4.

[175] Mihaly Varadi et al. “AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models”. In: *Nucleic Acids Research* 50 (D1 Nov. 2021), pp. D439–D444. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1061.

[176] Mihaly Varadi et al. “AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences”. In: *Nucleic Acids Research* 52 (D1 Jan. 5, 2024), pp. D368–D375. ISSN: 0305-1048. DOI: 10.1093/nar/gkad1011.

[177] Debora S. Marks et al. “Protein 3D Structure Computed from Evolutionary Sequence Variation”. In: *PLOS ONE* 6.12 (Dec. 7, 2011). Publisher: Public Library of Science, e28766. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0028766.

[178] Tadeo Saldaño et al. “Impact of protein conformational diversity on AlphaFold predictions”. In: *Bioinformatics* 38.10 (May 13, 2022), pp. 2742–2748. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac202.

[179] Carmen Al-Masri et al. “Investigating the conformational landscape of AlphaFold2-predicted protein kinase structures”. In: *Bioinformatics Advances* 3.1 (Jan. 1, 2023), vbad129. ISSN: 2635-0041. DOI: 10.1093/bioadv/vbad129.

[180] Hao-Bo Guo et al. “AlphaFold2 models indicate that protein sequence determines both structure and dynamics”. In: *Scientific Reports* 12 (June 23, 2022), p. 10696. ISSN: 2045-2322. DOI: 10.1038/s41598-022-14382-9.

[181] Carter J. Wilson, Wing-Yiu Choy, and Mikko Karttunen. "AlphaFold2: A Role for Disordered Protein/Region Prediction?" In: *International Journal of Molecular Sciences* 23.9 (Apr. 21, 2022), p. 4591. ISSN: 1422-0067. DOI: 10.3390/ijms23094591.

[182] Oliviero Carugo. "pLDDT Values in AlphaFold2 Protein Models Are Unrelated to Globular Protein Local Flexibility". In: *Crystals* 13.11 (Nov. 2023). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 1560. ISSN: 2073-4352. DOI: 10.3390/cryst13111560.

[183] Minkyung Baek et al. "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* 373.6557 (Aug. 20, 2021). Publisher: American Association for the Advancement of Science, pp. 871–876. DOI: 10.1126/science.abj8754.

[184] Ruidong Wu et al. *High-resolution de novo structure prediction from primary sequence*. Pages: 2022.07.21.500999 Section: New Results. July 22, 2022. DOI: 10.1101/2022.07.21.500999.

[185] Xiaomin Fang et al. "HelixFold-Single: MSA-free Protein Structure Prediction by Using Protein Language Model as an Alternative". In: *Nature Machine Intelligence* 5.10 (Oct. 9, 2023), pp. 1087–1096. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00721-6. arXiv: 2207.13921[cs,q-bio].

[186] Thomas D. Barrett et al. *So ManyFolds, So Little Time: Efficient Protein Structure Prediction With pLMs and MSAs*. Pages: 2022.10.15.511553 Section: New Results. Mar. 22, 2024. DOI: 10.1101/2022.10.15.511553.

[187] Josh Abramson et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* 630.8016 (June 2024). Publisher: Nature Publishing Group, pp. 493–500. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07487-w.

[188] P. A. Karplus and G. E. Schulz. "Prediction of chain flexibility in proteins: A tool for the selection of peptide antigens". In: *Naturwissenschaften* 72.4 (Apr. 1985), pp. 212–213. ISSN: 0028-1042, 1432-1904. DOI: 10.1007/BF01195768.

[189] Avner Schlessinger and Burkhard Rost. "Protein flexibility and rigidity predicted from sequence". In: *Proteins* 61.1 (Oct. 1, 2005), pp. 115–126. ISSN: 1097-0134. DOI: 10.1002/prot.20587.

[190] Avner Schlessinger, Guy Yachdav, and Burkhard Rost. "PROFbval: predict flexible and rigid residues in proteins". In: *Bioinformatics* 22.7 (Apr. 1, 2006), pp. 891–893. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl032.

[191] Alexandre G. de Brevern et al. “PredyFlexy: flexibility and local structure prediction from sequence”. In: *Nucleic Acids Research* 40 (W1 July 1, 2012), W317–W322. ISSN: 0305-1048. DOI: 10.1093/nar/gks482.

[192] Tarun J. Narwani et al. “*In silico* prediction of protein flexibility with local structure approach”. In: *Biochimie* 165 (Oct. 1, 2019), pp. 150–155. ISSN: 0300-9084. DOI: 10.1016/j.biochi.2019.07.025.

[193] Ashraf Yaseen et al. “FLEXc: protein flexibility prediction using context-based statistics, predicted structural features, and sequence information”. In: *BMC Bioinformatics* 17.8 (Aug. 31, 2016), p. 281. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1117-3.

[194] Yann Vander Meersche et al. “MEDUSA: Prediction of Protein Flexibility from Sequence”. In: *Journal of Molecular Biology*. Computation Resources for Molecular Biology 433.11 (May 28, 2021), p. 166882. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2021.166882.

[195] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Comput.* 9.8 (Nov. 1, 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

[196] Akash Pandey et al. “B-factor prediction in proteins using a sequence-based deep learning model”. In: *Patterns* 4.9 (Sept. 8, 2023), p. 100805. ISSN: 2666-3899. DOI: 10.1016/j.patter.2023.100805.

[197] Qianqian Wang et al. *Prediction of Protein B-factor Profiles based on Bidirectional Long Short-Term Memory Network*. Sept. 14, 2023. DOI: 10.26434/chemrxiv-2023-59cp5.

[198] Gang Xu et al. *OPUS-BFactor: Predicting protein B-factor with sequence and structure information*. Pages: 2024.07.17.604018 Section: New Results. July 19, 2024. DOI: 10.1101/2024.07.17.604018.

[199] David Bramer and Guo-Wei Wei. “Multiscale weighted colored graphs for protein flexibility and rigidity analysis”. In: *The Journal of Chemical Physics* 148.5 (Feb. 7, 2018), p. 054103. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.5016562.

[200] Domenico Scaramozzino et al. “Structural Compliance – A New Metric for Protein Flexibility”. In: *Proteins* 88.11 (Nov. 2020), pp. 1482–1492. ISSN: 0887-3585. DOI: 10.1002/prot.25968.

[201] Sjors H. W. Scheres. "RELION: Implementation of a Bayesian approach to cryo-EM structure determination". In: *Journal of Structural Biology* 180.3 (Dec. 1, 2012), pp. 519–530. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2012.09.006.

[202] Ali Punjani et al. "cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination". In: *Nature Methods* 14.3 (Mar. 2017). Publisher: Nature Publishing Group, pp. 290–296. ISSN: 1548-7105. DOI: 10.1038/nmeth.4169.

[203] N. Grigorieff. "Frealign: An Exploratory Tool for Single-Particle Cryo-EM". In: *Methods in Enzymology* 579 (2016), pp. 191–226. ISSN: 1557-7988. DOI: 10.1016/bs.mie.2016.04.013.

[204] Weiping Liu and Joachim Frank. "Estimation of variance distribution in three-dimensional reconstruction. I. Theory". In: *JOSA A* 12.12 (Dec. 1, 1995). Publisher: Optica Publishing Group, pp. 2615–2627. ISSN: 1520-8532. DOI: 10.1364/JOSAA.12.002615.

[205] Pawel A. Penczek, Marek Kimmel, and Christian M.T. Spahn. "Identifying Conformational States of Macromolecules by Eigen-Analysis of Resampled Cryo-EM Images". In: *Structure* 19.11 (Nov. 2011), pp. 1582–1590. ISSN: 09692126. DOI: 10.1016/j.str.2011.10.003.

[206] Ali Punjani and David J. Fleet. "3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM". In: *Journal of Structural Biology* 213.2 (June 1, 2021), p. 107702. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2021.107702.

[207] Florence Tama, Osamu Miyashita, and Charles L. Brooks III. "Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM". In: *Journal of Structural Biology*. Time-Resolved Imaging of Macromolecular Processes and Interactions 147.3 (Sept. 1, 2004), pp. 315–326. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2004.03.002.

[208] Qiyu Jin et al. "Iterative Elastic 3D-to-2D Alignment Method Using Normal Modes for Studying Structural Dynamics of Large Macromolecular Complexes". In: *Structure* 22.3 (Mar. 4, 2014), pp. 496–506. ISSN: 0969-2126. DOI: 10.1016/j.str.2014.01.004.

[209] Rémi Vuillemot et al. “NMMD: Efficient Cryo-EM Flexible Fitting Based on Simultaneous Normal Mode and Molecular Dynamics atomic displacements”. In: *Journal of Molecular Biology* 434.7 (Apr. 15, 2022), p. 167483. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2022.167483.

[210] Rémi Vuillemot et al. “MDSPACE: Extracting Continuous Conformational Landscapes from Cryo-EM Single Particle Datasets Using 3D-to-2D Flexible Fitting based on Molecular Dynamics Simulation”. In: *Journal of Molecular Biology*. New Frontier of Cryo-Electron Microscopy Technology 435.9 (May 1, 2023), p. 167951. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2023.167951.

[211] Diederik P. Kingma and Max Welling. *Auto-Encoding Variational Bayes*. version: 1. Dec. 20, 2013. DOI: 10.48550/arXiv.1312.6114. arXiv: 1312.6114.

[212] Ellen D. Zhong et al. “CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks”. In: *Nature Methods* 18.2 (Feb. 2021). Publisher: Nature Publishing Group, pp. 176–185. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01049-4.

[213] Ellen D. Zhong et al. “CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, Oct. 2021, pp. 4046–4055. ISBN: 978-1-66542-812-5. DOI: 10.1109/ICCV48922.2021.00403.

[214] Muyuan Chen and Steven J. Ludtke. “Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM”. In: *Nature Methods* 18.8 (Aug. 2021). Publisher: Nature Publishing Group, pp. 930–936. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01220-5.

[215] Dari Kimanius, Kiarash Jamali, and Sjors Scheres. “Sparse Fourier Backpropagation in Cryo-EM Reconstruction”. In: *Advances in Neural Information Processing Systems* 35 (Dec. 6, 2022), pp. 12395–12408.

[216] Johannes Schwab et al. “DynaMight: estimating molecular motions with improved reconstruction from cryo-EM images”. In: *Nature Methods* 21.10 (Oct. 2024). Publisher: Nature Publishing Group, pp. 1855–1862. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02377-5.

[217] Ali Punjani and David J. Fleet. "3DFlex: determining structure and motion of flexible proteins from cryo-EM". In: *Nature Methods* 20.6 (June 2023). Publisher: Nature Publishing Group, pp. 860–870. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01853-8.

[218] Devlina Chakravarty and Lauren L. Porter. "AlphaFold2 fails to predict protein fold switching". In: *Protein Science: A Publication of the Protein Society* 31.6 (June 2022), e4353. ISSN: 1469-896X. DOI: 10.1002/pro.4353.

[219] Diego del Alamo et al. "Sampling alternative conformational states of transporters and receptors with AlphaFold2". In: *eLife* 11 (Mar. 3, 2022). Ed. by Janice L Robertson, Kenton J Swartz, and Janice L Robertson. Publisher: eLife Sciences Publications, Ltd, e75751. ISSN: 2050-084X. DOI: 10.7554/eLife.75751.

[220] Richard A. Stein and Hassane S. Mchaourab. "SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2". In: *PLOS Computational Biology* 18.8 (Aug. 22, 2022). Publisher: Public Library of Science, e1010483. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010483.

[221] Hannah K. Wayment-Steele et al. "Predicting multiple conformations via sequence clustering and AlphaFold2". In: *Nature* 625.7996 (Jan. 2024). Publisher: Nature Publishing Group, pp. 832–839. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06832-9.

[222] Gabriel Monteiro da Silva et al. "High-throughput prediction of protein conformational distributions with subsampled AlphaFold2". In: *Nature Communications* 15.1 (Mar. 27, 2024). Publisher: Nature Publishing Group, p. 2464. ISSN: 2041-1723. DOI: 10.1038/s41467-024-46715-9.

[223] Björn Wallner. "AFsample: improving multimer prediction with AlphaFold using massive sampling". In: *Bioinformatics* 39.9 (Sept. 2, 2023). Ed. by Janet Kelso, btad573. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad573.

[224] Björn Wallner. "Improved multimer prediction using massive sampling with AlphaFold in CASP15". In: *Proteins: Structure, Function, and Bioinformatics* 91.12 (2023). Publisher: Wiley Online Library, pp. 1734–1746.

[225] Patrick Bryant and Frank Noé. "Structure prediction of alternative protein conformations". In: *Nature Communications* 15.1 (Aug. 26, 2024). Publisher: Nature Publishing Group, p. 7328. ISSN: 2041-1723. DOI: 10.1038/s41467-024-51507-2.

[226] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. Dec. 16, 2020. DOI: 10.48550/arXiv.2006.11239. arXiv: 2006.11239[cs,stat].

[227] John Ingraham et al. *Illuminating protein space with a programmable generative model*. Pages: 2022.12.01.518682 Section: New Results. Dec. 2, 2022. DOI: 10.1101/2022.12.01.518682.

[228] Joseph L. Watson et al. *Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models*. Pages: 2022.12.09.519842 Section: New Results. Dec. 14, 2022. DOI: 10.1101/2022.12.09.519842.

[229] Yeqing Lin and Mohammed AlQuraishi. *Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds*. June 6, 2023. arXiv: 2301.12485[cs,q-bio].

[230] Yeqing Lin et al. *Out of Many, One: Designing and Scaffolding Proteins at the Scale of the Structural Universe with Genie 2*. May 24, 2024. arXiv: 2405.15489[cs,q-bio].

[231] Urszula Julia Komorowska et al. "DYNAMICS-INFORMED PROTEIN DESIGN WITH STRUCTURE CONDITIONING". In: (2024).

[232] Simon V Mathis et al. "Normal Mode Diffusion: Towards Dynamics-Informed Protein Design". In: ().

[233] Bowen Jing et al. *EigenFold: Generative Protein Structure Prediction with Diffusion Models*. Apr. 4, 2023. DOI: 10.48550/arXiv.2304.02198. arXiv: 2304.02198[physics,q-bio].

[234] Shuxin Zheng et al. "Predicting equilibrium distributions for molecular systems with deep learning". In: *Nature Machine Intelligence* 6.5 (May 2024). Publisher: Nature Publishing Group, pp. 558–567. ISSN: 2522-5839. DOI: 10.1038/s42256-024-00837-3.

[235] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. "AlphaFold meets flow matching for generating protein ensembles". In: *arXiv preprint:2402.04845* (2024).

[236] The UniProt Consortium. "UniProt: the Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Research* 51 (D1 Nov. 2022), pp. D523–D531. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1052.

[237] Cathy H. Wu et al. “The Universal Protein Resource (UniProt): an expanding universe of protein information”. In: *Nucleic Acids Research* 34 (suppl_1 Jan. 2006), pp. D187–D191. ISSN: 0305-1048. DOI: 10.1093/nar/gkj161.

[238] Mitchell D Miller and George N Phillips. “Moving beyond static snapshots: Protein dynamics and the ”Protein Data Bank””. In: *Journal of Biological Chemistry* 296 (2021). Publisher: ASBMB.

[239] Isak Johansson-Åkhe and Björn Wallner. “Improving peptide-protein docking with AlphaFold-Multimer using forced sampling”. In: *Frontiers in Bioinformatics* 2 (Sept. 26, 2022). Publisher: Frontiers. ISSN: 2673-7647. DOI: 10.3389/fbinf.2022.959160.

[240] Lim Heo and Michael Feig. “Multi-state modeling of G-protein coupled receptors at experimental accuracy”. In: *Proteins: Structure, Function, and Bioinformatics* 90.11 (2022). Publisher: Wiley Online Library, pp. 1873–1885.

[241] Bulat Faezov and Roland L Dunbrack Jr. “AlphaFold2 models of the active form of all 437 catalytically-competent typical human kinase domains”. In: *bioRxiv* (2023). Publisher: Cold Spring Harbor Laboratory, pp. 2023–07.

[242] Devlina Chakravarty et al. “AlphaFold2 has more to learn about protein energy landscapes”. In: *bioRxiv* (2023). Publisher: Cold Spring Harbor Laboratory, pp. 2023–12.

[243] Venkata K Ramaswamy et al. “Deep learning protein conformational space with convolutions and latent interpolations”. In: *Physical Review X* 11.1 (2021). Publisher: APS, p. 011052.

[244] Theresa A Ramelot, Roberto Tejero, and Gaetano T Montelione. “Representing structures of the multiple conformational states of proteins”. In: *Current Opinion in Structural Biology* 83 (2023). Publisher: Elsevier, p. 102703.

[245] Stephanie Wankowicz and James Fraser. “Comprehensive Encoding of Conformational and Compositional Protein Structural Ensembles through mmCIF Data Structure”. In: *ChemRxiv* (2023). DOI: 10.26434/chemrxiv-2023-ggd1w-v2.

[246] Joseph IJ Ellaway et al. “Identifying Protein Conformational States in the PDB and Comparison to AlphaFold2 Predictions”. In: *bioRxiv* (2023). Publisher: Cold Spring Harbor Laboratory, pp. 2023–07.

[247] Mihaly Varadi et al. “PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education”. In: *Protein Science* 31.10 (Sept. 2022). Publisher: Wiley. ISSN: 1469-896X. DOI: 10.1002/pro.4439.

[248] Vivek Modi and Roland L Dunbrack Jr. “Kincore: a web resource for structural classification of protein kinases and their inhibitors”. In: *Nucleic Acids Research* 50 (D1 2022). Publisher: Oxford University Press, pp. D654–D664.

[249] Mitchell I Parker et al. “Delineating the RAS conformational landscape”. In: *Cancer research* 82.13 (2022). Publisher: AACR, pp. 2485–2498.

[250] Hedvig Tordai et al. “Comprehensive collection and prediction of ABC transmembrane protein structures in the AI era of structural biology”. In: *International Journal of Molecular Sciences* 23.16 (2022). Publisher: MDPI, p. 8877.

[251] Gáspár Pándy-Szekeres et al. “GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources”. In: *Nucleic Acids Research* 51 (D1 2023). Publisher: Oxford University Press, pp. D395–D402.

[252] Ian T Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016). Publisher: The Royal Society Publishing, p. 20150202.

[253] Atanu Maity, Sarmistha Majumdar, and Shubhra Ghosh Dastidar. “Flexibility enables to discriminate between ligands: Lessons from structural ensembles of Bcl-xl and Mcl-1”. In: *Computational Biology and Chemistry* 77 (2018). Publisher: Elsevier, pp. 17–27.

[254] Xin-Qiu Yao et al. “Navigating the conformational landscape of G protein–coupled receptor kinases during allosteric activation”. In: *Journal of Biological Chemistry* 292.39 (2017). Publisher: ASBMB, pp. 16032–16043.

[255] Jordi Mestres. “Structure conservation in cytochromes P450”. In: *Proteins: Structure, Function, and Bioinformatics* 58.3 (2005). Publisher: Wiley Online Library, pp. 596–609.

[256] DM Van Aalten et al. “Protein dynamics derived from clusters of crystal structures”. In: *Biophysical Journal* 73.6 (1997). Publisher: Elsevier, pp. 2891–2896.

[257] Elodie Laine, Valentin Lombard, and Sergei Grudinin. *Explaining Conformational Diversity in Protein Families through Molecular Motions*. July 2024. DOI: 10.6084/m9.figshare.c.7050008.v1.

[258] Martin Steinegger and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". In: *Nature Biotechnology* 35.11 (Nov. 1, 2017), pp. 1026–1028. ISSN: 1546-1696. DOI: 10.1038/nbt.3988.

[259] Kazutaka Katoh and Daron M Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". In: *Molecular biology and evolution* 30.4 (2013). Publisher: Society for Molecular Biology and Evolution, pp. 772–780.

[260] Steven Henikoff and Jorja G Henikoff. "Amino acid substitution matrices from protein blocks." In: *Proceedings of the National Academy of Sciences* 89.22 (1992). Publisher: National Acad Sciences, pp. 10915–10919.

[261] Douglas L Theobald. "Rapid calculation of RMSDs using a quaternion-based characteristic polynomial". In: *Acta Crystallographica Section A: Foundations of Crystallography* 61.4 (2005). Publisher: International Union of Crystallography, pp. 478–480.

[262] Pu Liu, Dimitris K Agrafiotis, and Douglas L Theobald. "Fast determination of the optimal rotational matrix for macromolecular superpositions". In: *Journal of Computational Chemistry* 31.7 (2010). Publisher: Wiley Online Library, pp. 1561–1563.

[263] Rafael Brüschweiler. "Collective protein dynamics and nuclear spin relaxation". In: *The Journal of Chemical Physics* 102.8 (1995). Publisher: AIP, pp. 3396–3403.

[264] F. Tama and Y. H. Sanejouand. "Conformational change of proteins arising from normal mode calculations". In: *Protein Engineering* 14.1 (Jan. 2001), pp. 1–6.

[265] W. Kabsch. "A solution for the best rotation to relate two sets of vectors". In: *Acta Crystallographica Section A* 32.5 (Sept. 1976), pp. 922–923. DOI: 10.1107/S0567739476001873.

[266] Marcin Wojdyr. "GEMMI: A library for structural biology". In: *Journal of Open Source Software* 7.73 (2022). Publisher: The Open Journal, p. 4200. DOI: 10.21105/joss.04200.

[267] Charles R. Harris et al. “Array programming with NumPy”. In: *Nature* 585.7825 (2020). ISBN: 1476-4687, pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[268] Stephen K Burley et al. “RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences”. In: *Nucleic Acids Research* 49 (D1 Nov. 2020), pp. D437–D451. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1038.

[269] Robbie P Joosten et al. “The PDB_REDO server for macromolecular structure model optimization”. In: *IUCrJ* 1.4 (2014). Publisher: International Union of Crystallography, pp. 213–220.

[270] Lars Skjærven et al. “Integrating protein structural dynamics and evolutionary analysis with Bio3D”. In: *BMC bioinformatics* 15.1 (2014). Publisher: BioMed Central, pp. 1–11.

[271] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. “Kernel principal component analysis”. In: *International conference on artificial neural networks*. Springer, 1997, pp. 583–588.

[272] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”. In: *Neural Computation* 10.5 (1998), pp. 1299–1319. DOI: 10.1162/089976698300017467.

[273] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[274] Jason Weston et al. “Kernel Dependency Estimation”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press, 2002.

[275] Jason Weston, Bernhard Schölkopf, and Gökhan Bakir. “Learning to Find Pre-Images”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Thrun, L. Saul, and B. Schölkopf. Vol. 16. MIT Press, 2003.

[276] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[277] J. A. Hartigan and M. A. Wong. “Algorithm AS 136: A K-Means Clustering Algorithm”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979). Publisher: [Wiley, Royal Statistical Society], pp. 100–108. ISSN: 00359254, 14679876.

[278] Bernard Chazelle. "An optimal convex hull algorithm in any fixed dimension". In: *Discrete & Computational Geometry* 10.4 (Dec. 1, 1993), pp. 377–409. ISSN: 1432-0444. DOI: 10.1007/BF02573985.

[279] Bernd Gärtner and Sven Schönherr. "An Efficient, Exact, and Generic Quadratic Programming Solver for Geometric Optimization". In: *Proceedings of the Sixteenth Annual Symposium on Computational Geometry*. SCG '00. event-place: Clear Water Bay, Kowloon, Hong Kong. New York, NY, USA: Association for Computing Machinery, 2000, pp. 110–118. ISBN: 1-58113-224-7. DOI: 10.1145/336154.336191.

[280] The CGAL Project. *CGAL User and Reference Manual*. 5.6. CGAL Editorial Board, 2023.

[281] Sergei Grudinin, Elodie Laine, and Alexandre Hoffmann. "Predicting protein functional motions: an old recipe with a new twist". In: *Biophysical Journal* 118.10 (2020). Publisher: Elsevier, pp. 2513–2525.

[282] Stephen G Aller et al. "Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding". In: *Science* 323.5922 (2009). Publisher: American Association for the Advancement of Science, pp. 1718–1722.

[283] Peter E. Czabotar et al. "Bax Crystal Structures Reveal How BH3 Domains Activate Bax and Nucleate Its Oligomerization to Induce Apoptosis". In: *Cell* 152.3 (Jan. 31, 2013). ISBN: 0092-8674 Publisher: Elsevier, pp. 519–531. DOI: 10.1016/j.cell.2012.12.031.

[284] Michael Zahn et al. "Mechanistic details of the actinobacterial lyase-catalyzed degradation reaction of 2-hydroxyisobutyryl-CoA". In: *Journal of Biological Chemistry* 298.1 (2022). Publisher: ASBMB.

[285] CW Müller et al. "Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding". In: *Structure* 4.2 (1996). Publisher: Elsevier, pp. 147–156.

[286] Paul C Whitford et al. "Conformational transitions of adenylate kinase: switching by cracking". In: *Journal of Molecular Biology* 366.5 (2007). Publisher: Elsevier, pp. 1661–1671.

[287] Andrej Perdih et al. "Targeted molecular dynamics simulation studies of binding and conformational changes in E. coli MurD". In: *PROTEINS: Structure, Function, and Bioinformatics* 68.1 (2007). Publisher: Wiley Online Library, pp. 243–254.

[288] David L Stokes and N Michael Green. “Structure and function of the calcium pump”. In: *Annual Review of Biophysics and Biomolecular Structure* 32.1 (2003). Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, pp. 445–468.

[289] Yoshiki Kabashima et al. “What ATP binding does to the Ca2+ pump and how nonproductive phosphoryl transfer is prevented in the absence of Ca2+”. In: *Proceedings of the National Academy of Sciences* 117.31 (2020). Publisher: National Acad Sciences, pp. 18448–18458.

[290] Karl-Peter Hopfner. “Invited review: Architectures and mechanisms of ATP binding cassette proteins”. In: *Biopolymers* 105.8 (2016). Publisher: Wiley Online Library, pp. 492–504.

[291] Wilfried W De Jong, Jack A Leunissen, and CE Voorter. “Evolution of the alpha-crystallin/small heat-shock protein family.” In: *Molecular biology and evolution* 10.1 (1993), pp. 103–126.

[292] Eman Basha, Heather O’Neill, and Elizabeth Vierling. “Small heat shock proteins and alpha-crystallins: dynamic proteins with flexible functions”. In: *Trends in biochemical sciences* 37.3 (2012). Publisher: Elsevier, pp. 106–117.

[293] Kristin A Krukenberg et al. “Conformational dynamics of the molecular chaperone Hsp90”. In: *Quarterly reviews of biophysics* 44.2 (2011). Publisher: Cambridge University Press, pp. 229–255.

[294] Jing Li, Joanna Soroka, and Johannes Buchner. “The Hsp90 chaperone machinery: conformational dynamics and regulation by co-chaperones”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1823.3 (2012). Publisher: Elsevier, pp. 624–635.

[295] David Chin and Anthony R Means. “Calmodulin: a prototypical calcium sensor”. In: *Trends in cell biology* 10.8 (2000). Publisher: Elsevier, pp. 322–328.

[296] Mingjie Zhang, Toshiyuki Tanaka, and Mitsuhiko Ikura. “Calcium-induced conformational transition revealed by the solution structure of apo calmodulin”. In: *Nature structural biology* 2.9 (1995). Publisher: Nature Publishing Group UK London, pp. 758–767.

[297] Alexandr P Kornev and Susan S Taylor. “Dynamics-driven allostery in protein kinases”. In: *Trends in biochemical sciences* 40.11 (2015). Publisher: Elsevier, pp. 628–647.

[298] Vivek Modi and Roland L Dunbrack Jr. "Defining a new nomenclature for the structures of active and inactive kinases". In: *Proceedings of the National Academy of Sciences* 116.14 (2019). Publisher: National Acad Sciences, pp. 6818–6827.

[299] Dhirendra K Simanshu, Dwight V Nissley, and Frank McCormick. "RAS proteins and their regulators in human disease". In: *Cell* 170.1 (2017). Publisher: Elsevier, pp. 17–33.

[300] Wesley I Sundquist and Hans-Georg Kräusslich. "HIV-1 assembly, budding, and maturation". In: *Cold Spring Harbor perspectives in medicine* (2012). Publisher: Cold Spring Harbor Laboratory Press, a006924.

[301] Gongpu Zhao et al. "Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics". In: *Nature* 497.7451 (2013). ISBN: 1476-4687, pp. 643–646. DOI: 10.1038/nature12162.

[302] Joshua B Tenenbaum, Vin de Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *Science* 290.5500 (2000). Publisher: American Association for the Advancement of Science, pp. 2319–2323.

[303] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.11 (2008).

[304] Randall Balestriero, Jerome Pesenti, and Yann LeCun. "Learning in high dimension always amounts to extrapolation". In: *arXiv preprint:2110.09485* (2021).

[305] Roman A Laskowski et al. "PROCHECK: a program to check the stereochemical quality of protein structures". In: *Journal of Applied Crystallography* 26.2 (1993). Publisher: International Union of Crystallography, pp. 283–291.

[306] Steven Hayward and Nobuhiro Go. "Collective variable description of native protein dynamics". In: *Annual Review of Physical Chemistry* 46.1 (1995). Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, pp. 223–250.

[307] Rohith Krishna et al. "Generalized biomolecular modeling and design with RoseTTA Fold All-Atom". In: *Science* 384.6693 (2024). Publisher: American Association for the Advancement of Science, eadl2528.

[308] Subhroshekhar Ghosh and Philippe Rigollet. “Sparse multi-reference alignment: Phase retrieval, uniform uncertainty principles and the beltway problem”. In: *Foundations of Computational Mathematics* (2022). Publisher: Springer, pp. 1–48.

[309] Afonso S Bandeira et al. “Estimation under group actions: recovering orbits from invariants”. In: *Applied and Computational Harmonic Analysis* (2023). Publisher: Elsevier.

[310] Asaf Abas, Tamir Bendory, and Nir Sharon. “The generalized method of moments for multi-reference alignment”. In: *IEEE Transactions on Signal Processing* 70 (2022). Publisher: IEEE, pp. 1377–1388.

[311] Douglas L Theobald and Phillip A Steindel. “Optimal simultaneous superpositioning of multiple structures with missing data”. In: *Bioinformatics* 28.15 (2012). Publisher: Oxford University Press, pp. 1972–1979.

[312] Afonso S Bandeira, Jonathan Niles-Weed, and Philippe Rigollet. “Optimal rates of estimation for multi-reference alignment”. In: *Mathematical Statistics and Learning* 2.1 (2020), pp. 25–75.

[313] D. Sala et al. “Modeling conformational states of proteins with AlphaFold”. In: *Current Opinion in Structural Biology* 81 (Aug. 1, 2023), p. 102645. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2023.102645.

[314] Marc F Lensink et al. “Impact of AlphaFold on structure prediction of protein complexes: The CASP15-CAPRI experiment”. In: *Proteins: Structure, Function, and Bioinformatics* 91.12 (2023). Publisher: Wiley Online Library, pp. 1658–1683.

[315] Guillaume Brysbaert et al. “MassiveFold: unveiling AlphaFold’s hidden potential with optimized and parallelized massive sampling”. In: (2024).

[316] Pedro Sfriso et al. “Residues coevolution guides the systematic identification of alternative functional conformations in proteins”. In: *Structure* 24.1 (2016). Publisher: Elsevier, pp. 116–126.

[317] Steven A Benner and Dietlinde Gerloff. “Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases”. In: *Advances in Enzyme Regulation* 31 (1991). Publisher: Elsevier, pp. 121–181.

[318] Ulrike Göbel et al. “Correlated mutations and residue contacts in proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 18.4 (1994). Publisher: Wiley Online Library, pp. 309–317.

[319] Angel R Ortiz et al. “Ab initio folding of proteins using restraints derived from evolutionary information”. In: *Proteins: Structure, Function, and Bioinformatics* 37 (S3 1999). Publisher: Wiley Online Library, pp. 177–185.

[320] Alan S Lapedes et al. “Correlated mutations in models of protein sequences: phylogenetic and structural effects”. In: *Lecture Notes-Monograph Series* (1999). Publisher: JSTOR, pp. 236–256.

[321] BG Giraud, John M Heumann, and Alan S Lapedes. “Superadditive correlation”. In: *Physical Review E* 59.5 (1999). Publisher: APS, p. 4983.

[322] John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg. “Graphical models of residue coupling in protein families”. In: *Proceedings of the 5th international workshop on Bioinformatics*. 2005, pp. 12–20.

[323] Martin Weigt et al. “Identification of direct residue contacts in protein–protein interaction by message passing”. In: *Proceedings of the National Academy of Sciences* 106.1 (2009). Publisher: National Acad Sciences, pp. 67–72.

[324] Tristan Bepler and Bonnie Berger. “Learning the protein language: Evolution, structure, and function”. In: *Cell systems* 12.6 (2021). Publisher: Elsevier, pp. 654–669.

[325] Yogesh Kalakoti and Björn Wallner. “AFsample2: Predicting multiple conformations and ensembles with AlphaFold2”. In: *bioRxiv* (2024). Publisher: Cold Spring Harbor Laboratory, pp. 2024–05.

[326] Lauren L Porter, Irina Artsimovitch, and César A Ramírez-Sarmiento. “Metamorphic proteins and how to find them”. In: *Current opinion in structural biology* 86 (2024). Publisher: Elsevier, p. 102807.

[327] Scott A. Hollingsworth and Ron O. Dror. “Molecular dynamics simulation for all”. In: *Neuron* 99.6 (Sept. 19, 2018), pp. 1129–1143. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2018.08.011.

[328] Haochuan Chen, Benoît Roux, and Christophe Chipot. “Discovering reaction pathways, slow variables, and committor probabilities with machine learning”. In: *Journal of Chemical Theory and Computation* 19.14 (2023). Publisher: ACS Publications, pp. 4414–4426.

[329] Frank Noé et al. “Machine learning for molecular simulation”. In: *Annual review of physical chemistry* 71.1 (2020). Publisher: Annual Reviews, pp. 361–390.

[330] Zineb Belkacemi et al. “Chasing collective variables using autoencoders and biased trajectories”. In: *Journal of chemical theory and computation* 18.1 (2021). Publisher: ACS Publications, pp. 59–78.

[331] Luigi Bonati, GiovanniMaria Piccini, and Michele Parrinello. “Deep learning the slow modes for rare events sampling”. In: *Proceedings of the National Academy of Sciences* 118.44 (2021). Publisher: National Acad Sciences, e2113533118.

[332] Yihang Wang, Joao Marcelo Lamim Ribeiro, and Pratyush Tiwary. “Machine learning approaches for analyzing and enhancing molecular dynamics simulations”. In: *Current opinion in structural biology* 61 (2020). Publisher: Elsevier, pp. 139–145.

[333] João Marcelo Lamim Ribeiro et al. “Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)”. In: *The Journal of chemical physics* 149.7 (2018). Publisher: AIP Publishing.

[334] Jiarui Lu, Bozitao Zhong, and Jian Tang. “Score-based enhanced sampling for protein molecular dynamics”. In: *ICML 2023 Workshop on Structured Probabilistic Inference {\backslash&} Generative Modeling*. 2023.

[335] Frank Noé et al. “Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning”. In: *Science* 365.6457 (2019). Publisher: American Association for the Advancement of Science, eaaw1147.

[336] Jill Trewhella. “Recent advances in small-angle scattering and its expanding impact in structural biology”. In: *Structure* 30.1 (2022). Publisher: Elsevier, pp. 15–23.

[337] Anne Martel and Frank Gabel. “Time-resolved small-angle neutron scattering (TR-SANS) for structural biology of dynamic systems: Principles, recent developments, and practical guidelines”. In: *Methods in enzymology* 677 (2022). Publisher: Elsevier, pp. 263–290.

[338] Holger Flechsig and Toshio Ando. “Protein dynamics by the combination of high-speed AFM and computational modeling”. In: *Current Opinion in Structural Biology* 80 (2023). Publisher: Elsevier, p. 102591.

[339] Valentin Lombard, Sergei Grudinin, and Elodie Laine. "Explaining Conformational Diversity in Protein Families through Molecular Motions". In: *Scientific Data* 11.1 (July 10, 2024), p. 752. ISSN: 2052-4463. DOI: 10.1038/s41597-024-03524-5.

[340] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (Mar. 17, 2023). Publisher: American Association for the Advancement of Science, pp. 1123–1130. DOI: 10.1126/science.ade2574.

[341] Monika Chandravanshi, Reshama Samanta, and Shankar Prasad Kanaujia. "Conformational trapping of a beta-glucosides-binding protein unveils the selective two-step ligand-binding mechanism of ABC importers". In: *Journal of molecular biology* 432.20 (2020). Publisher: Elsevier, pp. 5711–5734.

[342] Do-Hee Kim et al. "Role of PemI in the Staphylococcus aureus PemIK toxin–antitoxin complex: PemI controls PemK by acting as a PemK loop mimic". In: *Nucleic Acids Research* 50.4 (2022). Publisher: Oxford University Press, pp. 2319–2333.

[343] Alexandra J Machen, Mark T Fisher, and Bret D Freudenthal. "Anthrax toxin translocation complex reveals insight into the lethal factor unfolding and refolding mechanism". In: *Scientific Reports* 11.1 (2021). Publisher: Nature Publishing Group UK London, p. 13038.

[344] David M Anderson et al. "Structural insights into the transition of Clostridioides difficile binary toxin from prepore to pore". In: *Nature microbiology* 5.1 (2020). Publisher: Nature Publishing Group UK London, pp. 102–107.

[345] Immanuel Dhanasingh et al. "Functional and structural characterization of Deinococcus radiodurans R1 MazEF toxin-antitoxin system, Dr0416-Dr0417". In: *Journal of Microbiology* 59 (2021). Publisher: Springer, pp. 186–201.

[346] Jennifer M Podgorski et al. "A structural dendrogram of the actinobacteriophage major capsid proteins provides important structural insights into the evolution of capsid stability". In: *Structure* 31.3 (2023). Publisher: Elsevier, pp. 282–294.

[347] Lorna Richardson et al. "MGnify: the microbiome sequence data analysis resource in 2023". In: *Nucleic Acids Research* 51 (D1 2023). Publisher: Oxford University Press, pp. D753–D759.

[348] Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. "Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences". In: *Protein Science* 31.1 (2022). Publisher: Wiley Online Library, pp. 141–146.

[349] Inigo Barrio-Hernandez et al. "Clustering predicted structures at the scale of the known protein universe". In: *Nature* 622.7983 (2023). Publisher: Nature Publishing Group UK London, pp. 637–645.

[350] Martin Steinegger, Milot Mirdita, and Johannes Söding. "Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold". In: *Nature methods* 16.7 (2019). Publisher: Nature Publishing Group US New York, pp. 603–606.

[351] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015). Publisher: Nature Publishing Group UK London, pp. 436–444.

[352] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30. Issue: 1. Atlanta, GA, 2013, p. 3.

[353] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014). Publisher: JMLR. org, pp. 1929–1958.

[354] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[355] John C Gower and Garmt B Dijksterhuis. *Procrustes problems*. Vol. 30. OUP Oxford, 2004.

[356] Peter H Schönemann. "A generalized solution of the orthogonal procrustes problem". In: *Psychometrika* 31.1 (1966). Publisher: Springer, pp. 1–10.

[357] Yang Zhang and Jeffrey Skolnick. "TM-align: a protein structure alignment algorithm based on the TM-score". In: *Nucleic acids research* 33.7 (2005). Publisher: Oxford University Press, pp. 2302–2309.

[358] Valentin Lombard, Sergei Grudinin, and Elodie Laine. *Data for "SeaMoon: from protein language models to continuous structural heterogeneity"*. 2024. DOI: 10.1101/2024.09.23.614585.

[359] Simon K Kearsley. “On the orthogonal transformation used for structural comparisons”. In: *Acta Crystallographica Section A: Foundations of Crystallography* 45.2 (1989). Publisher: International Union of Crystallography, pp. 208–210.

Appendix A

Additional information for DANCE

Valentin Lombard[1], Sergei Grudinin[*2], Elodie Laine[*1,3]

[1] Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France.
[2] Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.
[3] Institut universitaire de France (IUF).

[*] corresponding authors: sergei.grudinin@univ-grenoble-alpes.fr, elodie.laine@sorbonne-universite.fr

Supplemental tables and figures

Table A.1: **Execution time on the PDB (748 297 protein chains)**

Step	**Library or tool**	**(CPU)**	**Time#**
a- Sequence extraction	GEMMI	16	9min 30s
b- Sequence clustering	MMseqs2	16	15s
c- Sequence alignment	MAFFT	16	23min 54s
de- Structure extraction and alignment	GEMMI	16	88min 2s
f- Linear motion extraction	NumPy	16	11min 46s
Total			2h 14min

Table A.2: **Properties of the ensembles in the most conservative and the most relaxed set ups.**

Property	Id., Cov.	Min.	1st Quart.	Median	3rd Quart.	Max.	Mean ± Sd
Ensemble size	80,80	2.00	2.00	4.00	8.00	3 334	12.00 ± 45.80
(in conformation)	30,50	2.00	3.00	5.00	14.00	12694	24.64 ± 137.35
Reference length	80,80	5.00	112.00	207.00	335.00	4579.00	252.03 ± 223.35
(in residue)	30,50	5.00	81.00	165.00	304.00	4516.00	224.07 ± 231.74
Max deviation	80,80	0.00	0.52	1.10	2.45	72.93	2.39 ± 3.95
(in Å)	30,50	0.00	0.78	2.15	5.11	114.50	4.25 ± 5.88
Motion complexity[a]	80,80	1.00	2.00	3.00	4.00	118.00	3.89 ± 4.43
(in mode)	30,50	1.00	2.00	3.00	4.00	105.00	3.98 ± 4.45
1st mode contribution[a]	80,80	7.30	50.00	64.90	80.90	100	64.85 ± 19.99
(in percentage)	30,50	7.30	49.90	65.40	82.20	100	65.32 ± 20.31
1st mode collectivity[a]	80,80	0.30	13.30	30.20	50.90	98.20	33.07 ± 21.92
(in percentage)	30,50	0.30	15.90	29.40	48.90	98.20	33.26 ± 20.90

[a] To compute motion properties, we focused on the subset of ensembles with at least three members. Indeed, pairs of conformations trivially exhibit single-mode motions and are thus disregarded variability exhibited by pairs of conformations can be trivially explained by only one mode

Table A.3: **Properties of the ensembles chosen for benchmarking manifold learning techniques.**

Reference PDB Id	**Protein name**	**Size** (# conf.)	**Length** (# res.)	**Mean deviation** (Å)	**Max deviation** (Å)	**Motion complexity** (# mode)	**1st mode contrib.** (%)	**1st mode coll.** (%)
1AKEA	ADK	42	214	2.48 ± 2.78	7.30	1	0.963	0.462
2WJPA	MurD	22	435	2.58 ± 4.09	16.07	2	0.816	0.616
1IWOA	ATPase	104	994	7.12 ± 4.19	15.84	3	0.597	0.481
5KOYB	ABC	61	1181	5.95 ± 3.20	14.94	4	0.786	0.681
2KLRA	Crys	23	82	2.08 ± 1.24	5.37	6	0.502	0.311
1AH6A	HSP90	52	213	1.13 ± 1.10	4.56	7	0.410	0.105
1NIWA	CALM	388	136	10.38 ± 4.23	23.68	8	0.415	0.580
2G1TA	KIN	122	271	2.85 ± 1.88	8.89	9	0.669	0.086
6YXWC	RAS	744	167	1.56 ± 1.02	8.58	24	0.347	0.173
3J345	CAP	3334	231	4.06 ± 1.68	17.55	30	0.232	0.070

Table A.4: **Proportion of conformations reconstructed with high accuracy.**

Reference	Protein	PCA (in %)	Poly-kPCA (in %)	RBF-kPCA (in %)	Sigmoid-kPCA (in %)	UMAP (in %)
1AKEA	ADK	83	83	83	83	-
2WJPA	MurD	5	82	82	82	0
1IWOA	ATPase	0	0	0	0	0
5KOYB	ABC	18	28	30	18	0
2KLRA	Crys	96	96	96	91	83
1AH6A	HSP90	92	90	90	92	90
1NIWA	CALM	0	1	2	0	-
2G1TA	KIN	93	93	93	93	57
6YXWC	RAS	99	99	98	99	-
3J345	CAP	99	99	99	99	-

We consider reconstructions with RMSD errors lower than 2 Å as highly accurate. For the kPCA, we set the hyperparameters to the values leading to the lowest average error over each conformational collection (see **Supplementary Table S5**).

Table A.5: **Average reconstruction errors for unseen conformations and hyperparameter values.**

Protein name	Method	Kernel type	sigma	alpha	n_{neigh}	d_{min}	Mean RMSD (Å)
ADK	pca						1.635
ADK	kpca	rbf	5.96e+03	1.00E-14			1.633
ADK	kpca	poly	13.3	2.81e+03			1.598
ADK	kpca	sigmoid	0.309	32.4			1.525
MurD	pca						3.115
MurD	umap				2	0.223	3.928
MurD	kpca	rbf	0.494	1.00E+05			1.774
MurD	kpca	poly	33.9	2.81e+03			1.755
MurD	kpca	sigmoid	1.00E+05	1.00E+05			1.774
ATPase	pca						3.607
ATPase	umap				9	0.112	5.161
ATPase	kpca	rbf	569	1.46e-13			3.567
ATPase	kpca	poly	910	5.96e-14			3.591
ATPase	kpca	sigmoid	2.33e+03	5.96e-14			3.606
ABC	pca						2.599
ABC	umap				2	1	3.911
ABC	kpca	rbf	569	5.96e-14			2.554
ABC	kpca	poly	569	3.56e-13			2.533
ABC	kpca	sigmoid	3.73e+03	5.96e-14			2.600
Crys	pca						1.360
Crys	umap				2	0.001	1.676
Crys	kpca	rbf	86.9	1.26e-11			1.356
Crys	kpca	poly	112	5.18e-12			1.357
Crys	kpca	sigmoid	0.193	5.43			1.302
HSP90	pca						0.719
HSP90	umap				4	0.889	0.865
HSP90	kpca	rbf	54.3	7.54e-11			0.704
HSP90	kpca	poly	222	2.44e-14			0.711
HSP90	kpca	sigmoid	1.00E+03	1.00E-14			0.718
CALM	pca						4.075
CALM	kpca	rbf	356	1.46e-13			3.525
CALM	kpca	poly	356	8.69e-13			3.531
CALM	kpca	sigmoid	356	1.6e-08			4.060
KIN	pca						1.228
KIN	umap				11	0.001	1.871
KIN	kpca	rbf	222	7.54e-11			1.225
KIN	kpca	poly	222	4.5e-10			1.226
KIN	kpca	sigmoid	910	5.96e-14			1.228
RAS	pca						0.612
RAS	kpca	rbf	139	2.12e-12			0.605
RAS	kpca	poly	139	3.09e-11			0.606
RAS	kpca	sigmoid	910	2.12e-12			0.612
CAP	pca						1.014
CAP	kpca	rbf	222	8.69e-13			0.986
CAP	kpca	poly	222	1.26e-11			0.989
CAP	kpca	sigmoid	356	4.5e-10			1.014

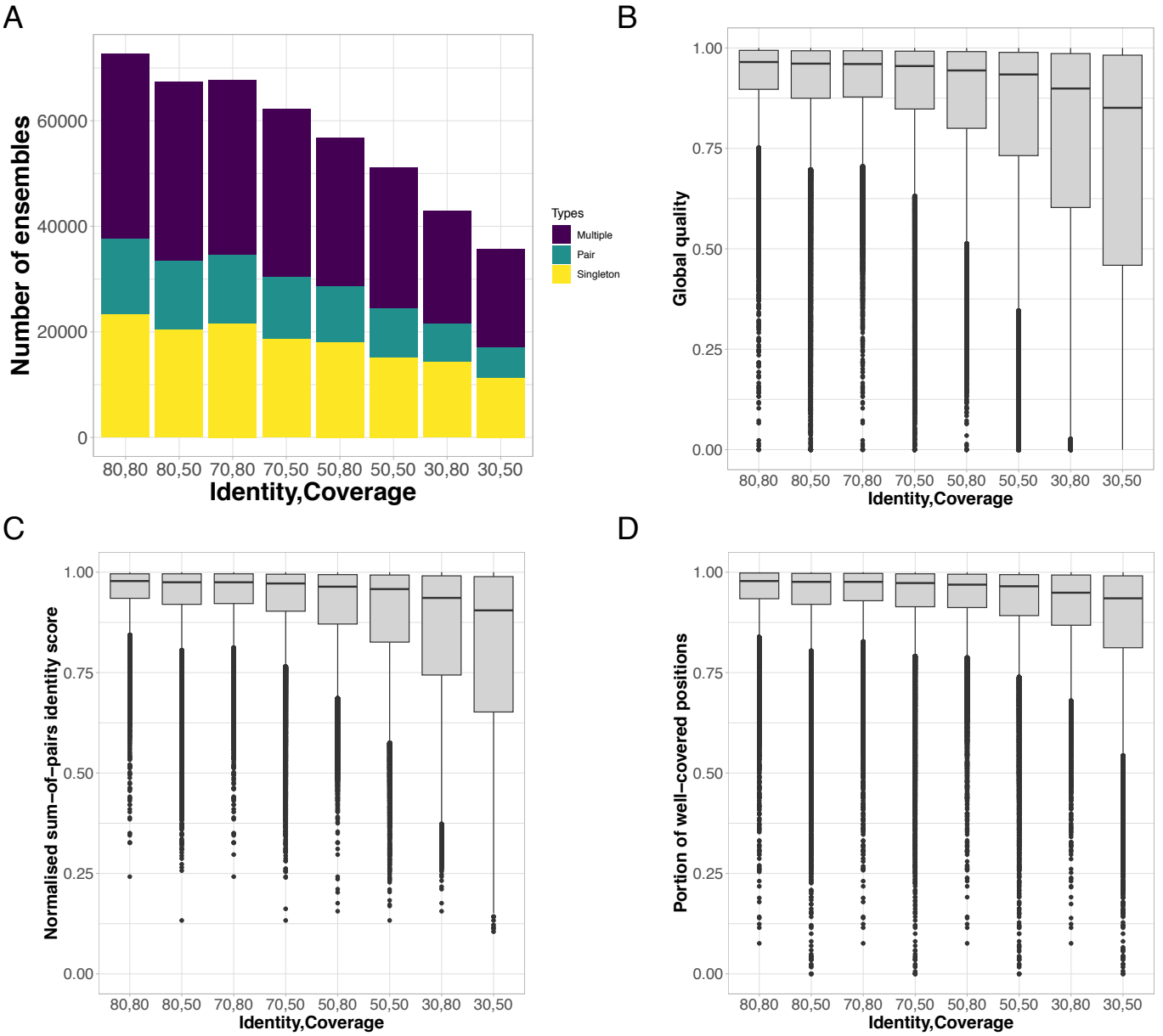


Figure A.1: **Global properties of the ensembles and their sequence alignments.** We report values computed across eight versions of the database, corresponding to eight combinations of sequence similarity and coverage thresholds. These combinations are given in x-axis. **A.** Number of singletons, pairs, and ensembles with at least 3 members. **B.** Distributions of sequence identity measured as a normalised sum-of-pairs scores with null mismatch and gap penalties. **C.** Distribution of coverage expressed as the fraction of positions with less than 80% gaps. **D.** Distribution of global alignment quality computed as a normalised sum-of-pairs scores with the following parameters: $\sigma_{match} = 1$, $\sigma_{mismatch} = \sigma_{gap} =$ -0.5 (see *Methods*).

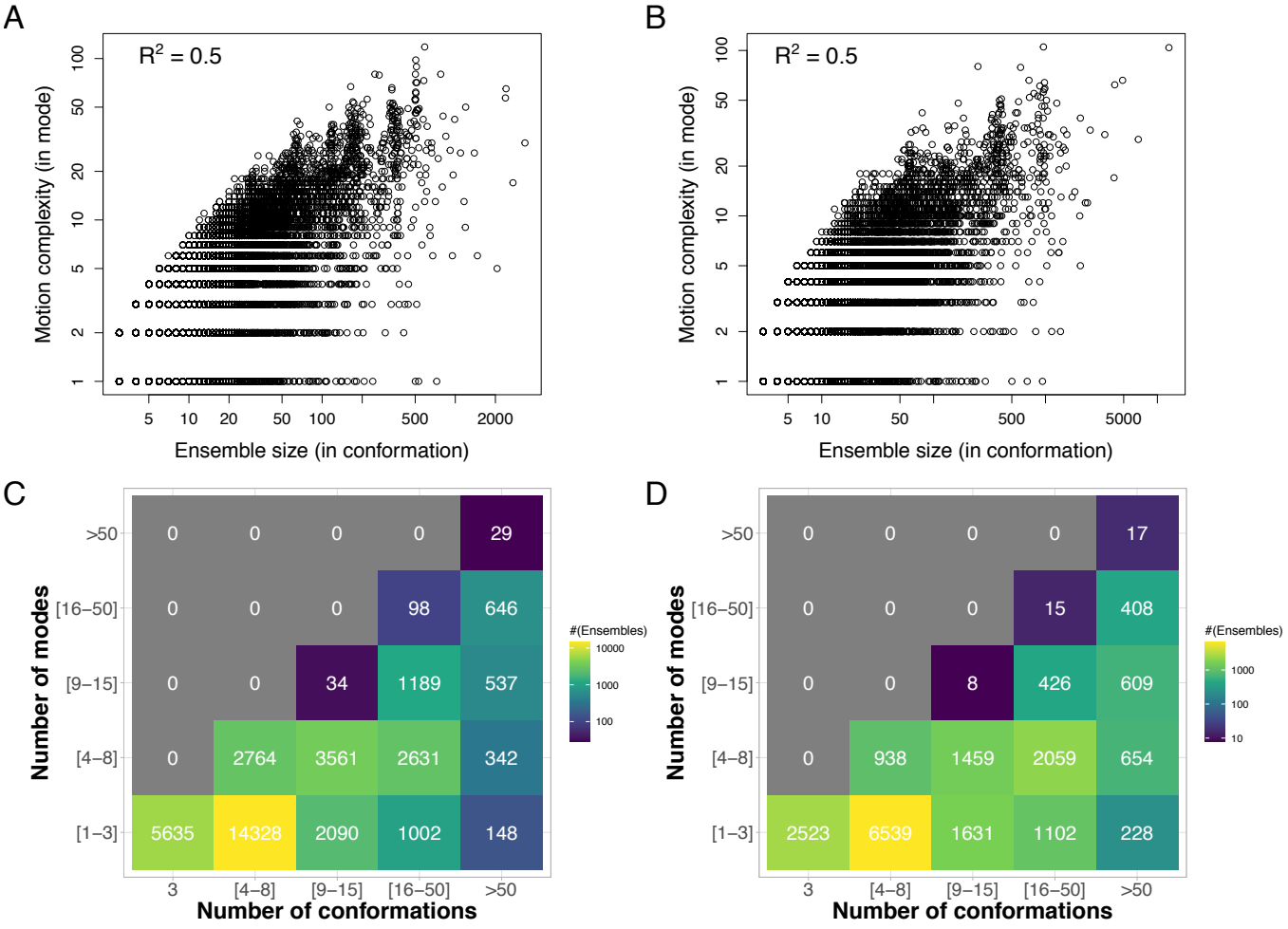


Figure A.2: **Influence of ensemble size on motion complexity** We report motion complexity, measured as the number of principal components or modes required to explain 80% of the positional variance, in function of the ensemble size, *i.e.* number of conformations. **A-B.** Scatterplots in log scale. **C-D.** Discretized heatmaps. We consider the most stringent set up, namely l_{80}^{80} (A,C), and the most relaxed one, namely l_{50}^{30} (B,D).

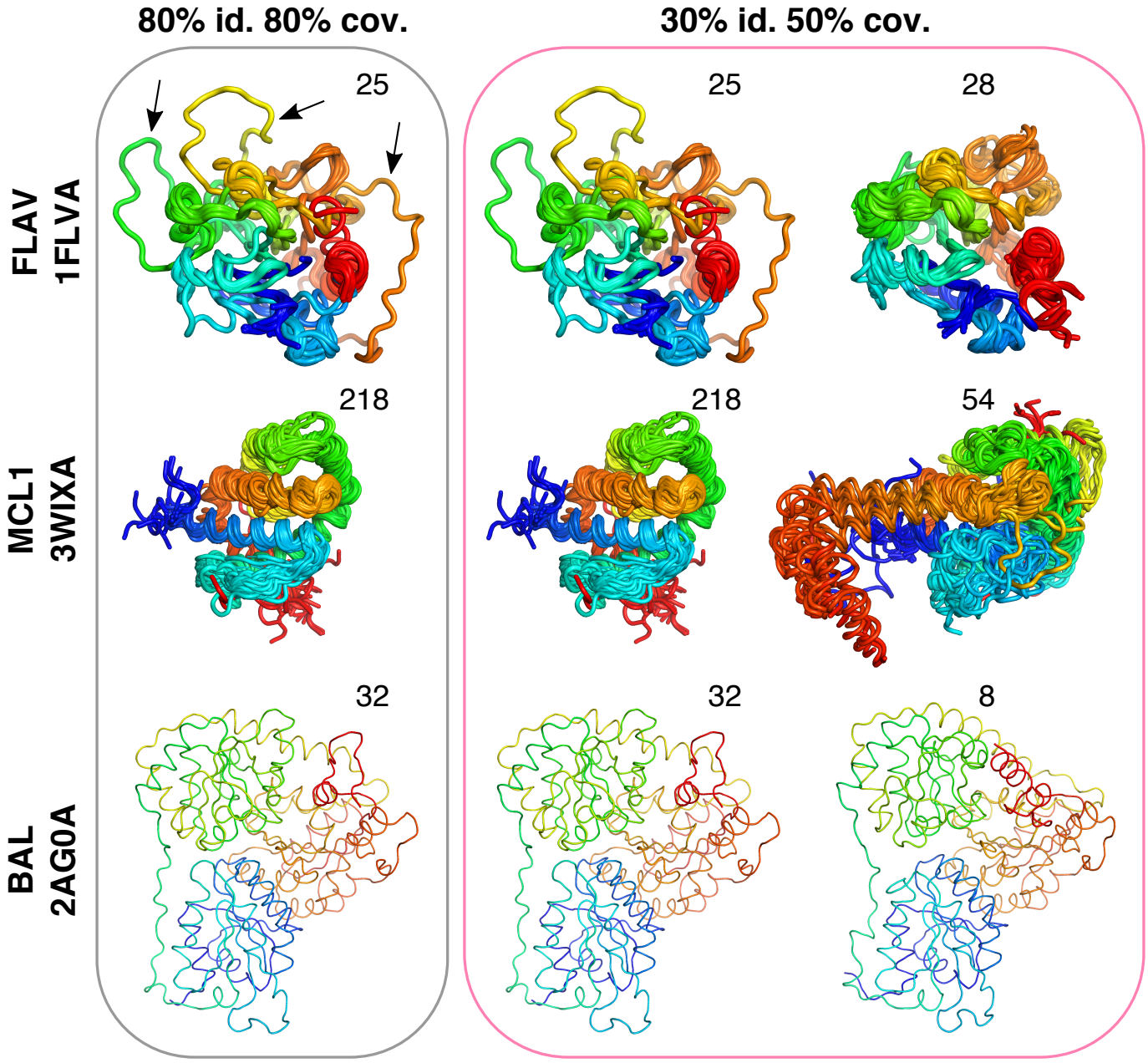


Figure A.3: **Expansion of three conformational ensembles upon relaxing sequence selection criteria.** We compare the set of conformations detected at two different levels of sequence similarity and coverage, namely l_{80}^{80} (on the left) and l_{50}^{30} (on the right). For the latter, we show separately the conformations already included in the ensemble at l_{80}^{80} (on the left) and the new additional conformations (on the right). The number of conformations in each (sub)ensemble is given on top. The color code indicates the position in the sequence, from the N-terminus in blue to the C-terminus in red. The flavodoxin (FLAV) ensemble contains one partially unfolded conformation, highlighted with the arrows. Some properties of these three examples are reported in Figure 2.

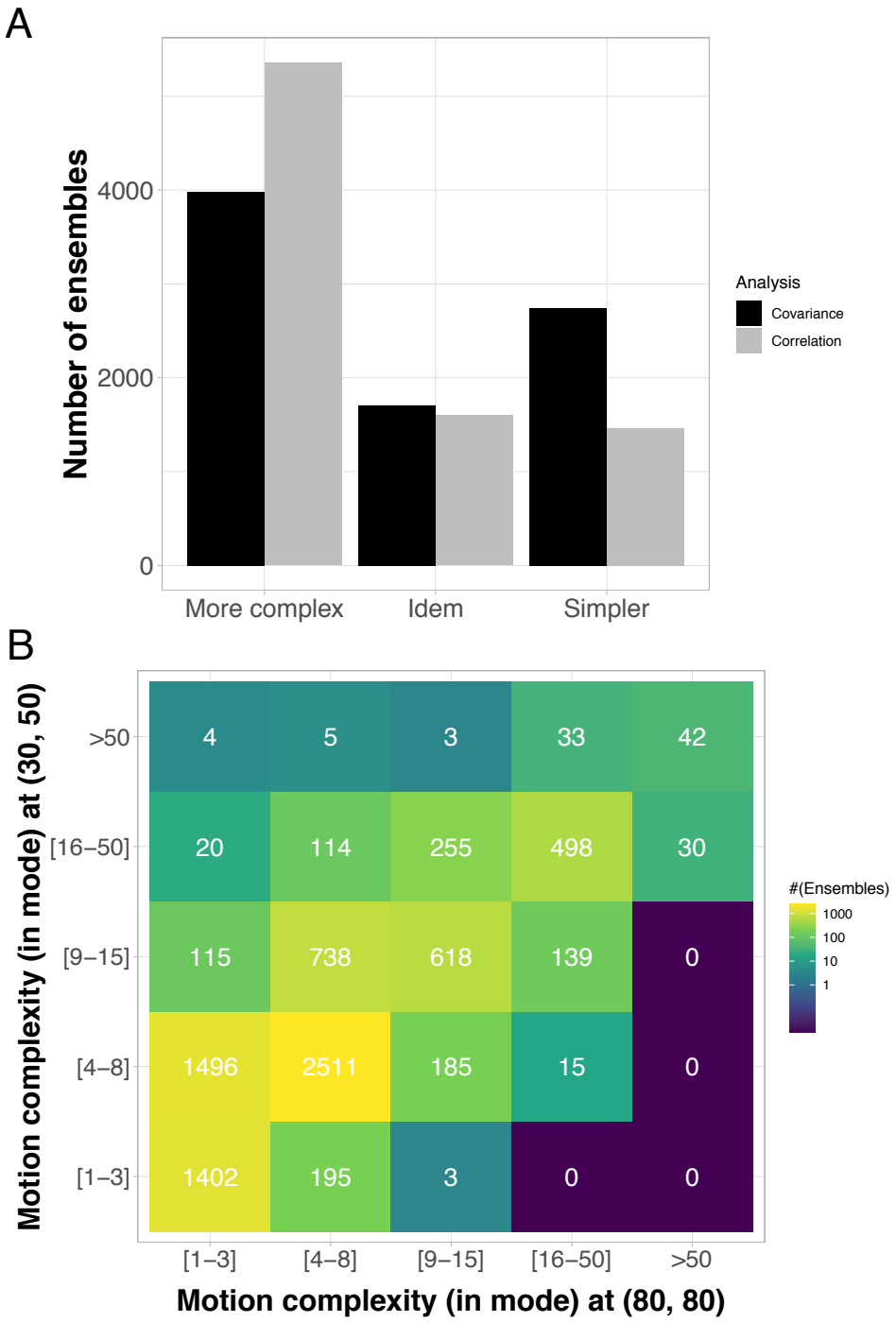


Figure A.4: **Evolution of motion complexity upon protein family expansion. A.** Number of ensembles where motion complexity increases, remains the same, or decreases between the most stringent and the most relaxed set ups. We extracted the motions from either the covariance (in black) or the correlation (in grey) matrix. **B.** Comparison of motion complexity estimated from the correlation matrix in the most stringent set up (x-axis) versus the most relaxed one (y-axis).

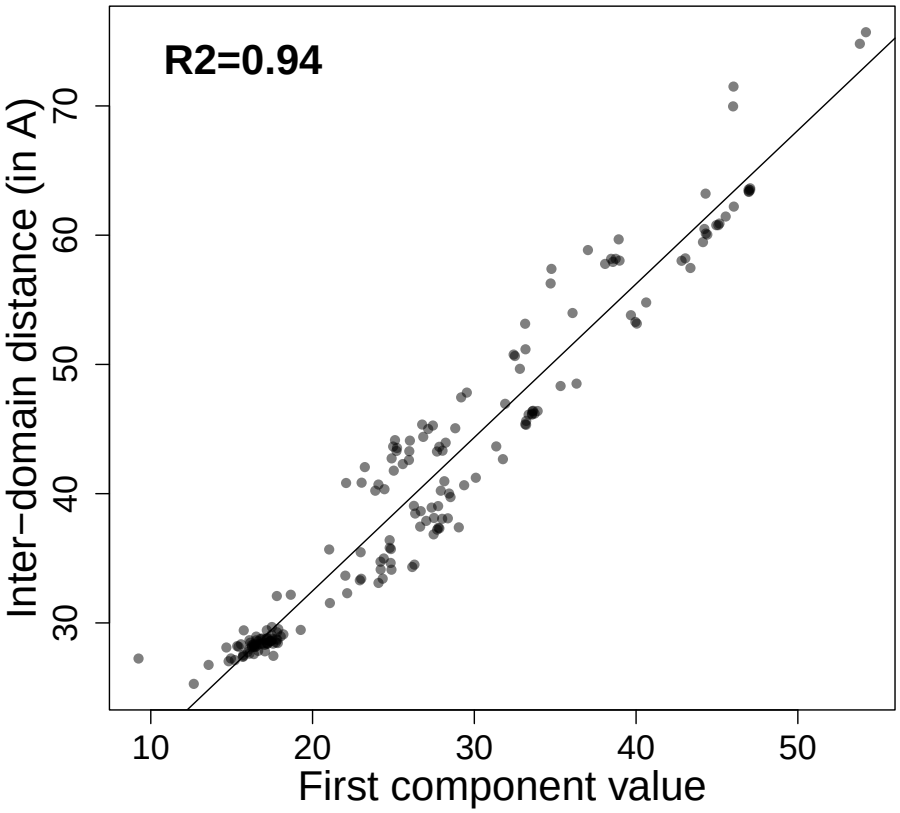


Figure A.5: **ABC protein opening in function of the first PCA component values.** The degree of opening of the ABC transporters is measured as the distance between the geometric centres of the two NBDs (in Å). The analysis is performed on the 188 conformations from the ABC structure similarity-based ensemble.

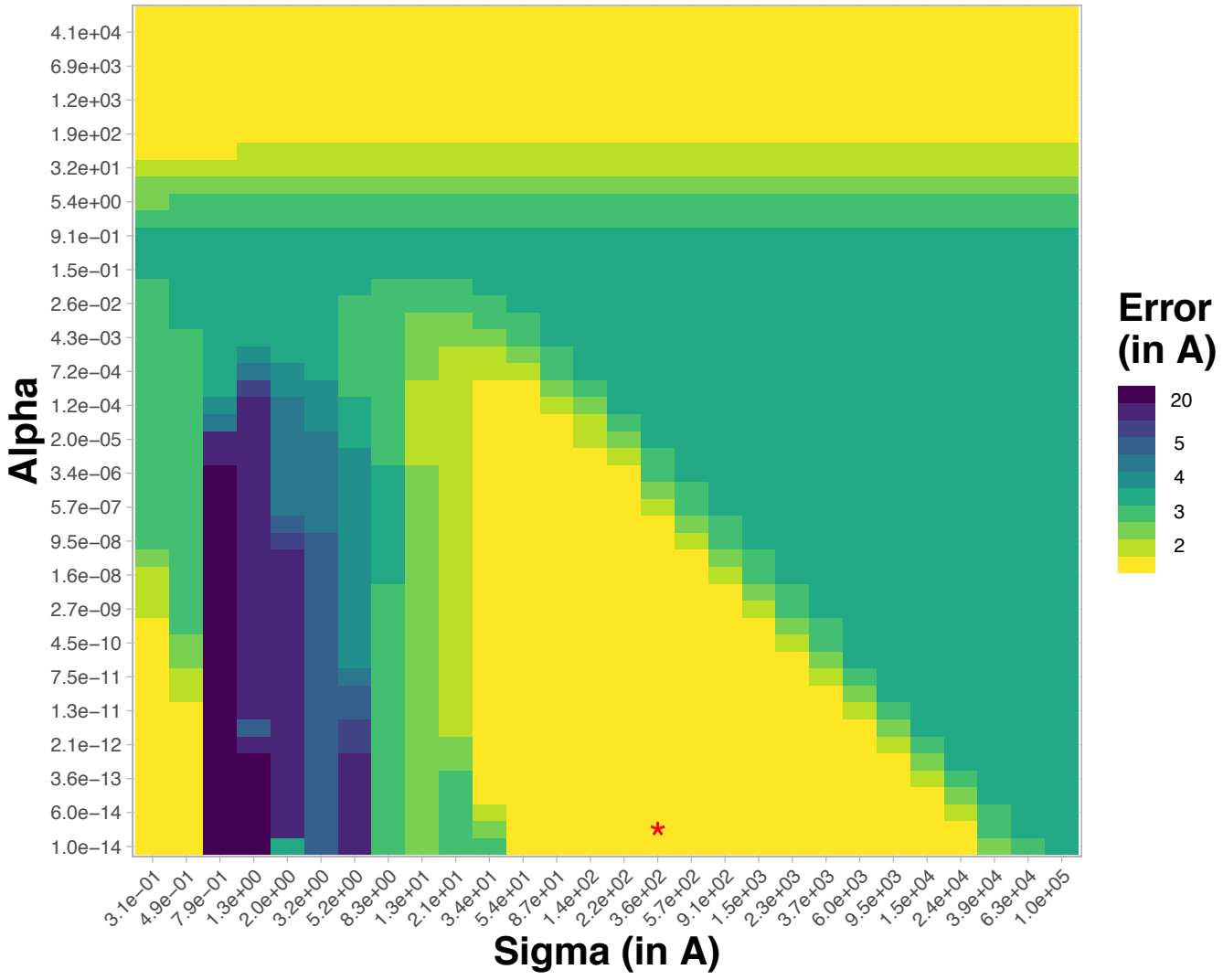


Figure A.6: **Systematic exploration of the two hyperparameters for kPCA-based conformation reconstruction.** We illustrate the influence of the hyper parameters σ and α on the reconstruction error (in Å) for a randomly picked up conformation (4th one) from the ADK protein ensemble. The red star highlights the optimal parameter values. We used the RBF kernel.

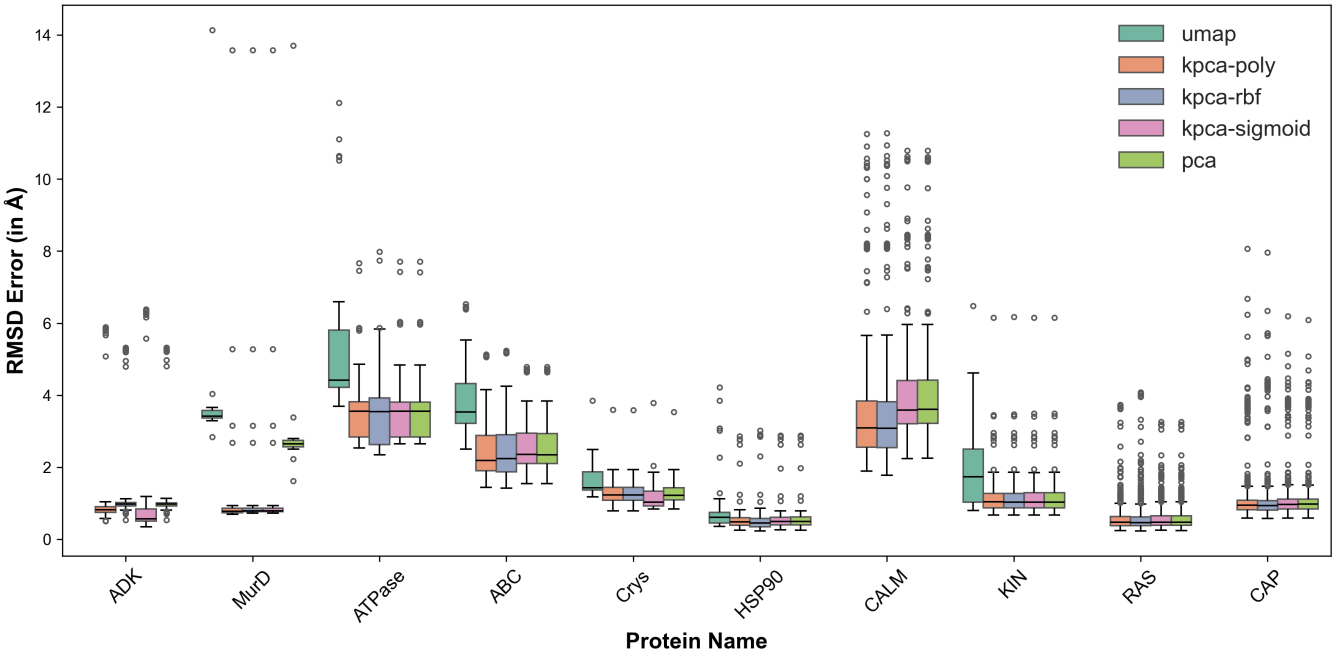


Figure A.7: **Distributions of the RMSD reconstruction errors (in Å) for each ensemble in the benchmark set.**We systematically reconstructed each conformation through a leave-one-cluster-out cross-validation procedure (see *Methods*). We set the hyperparameters of the kPCA and UMAP to the values yielding the best reconstruction, for each ensemble. The protein names in the x-axis are ordered according to motion complexity.

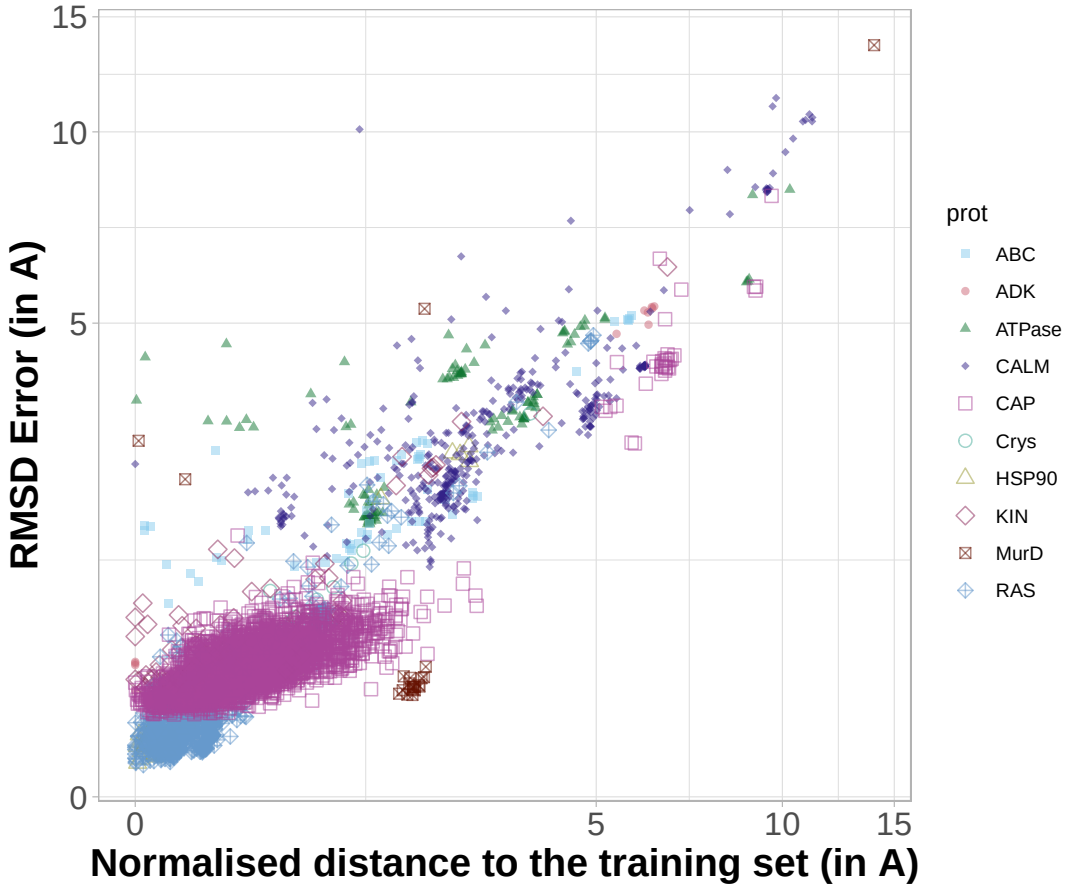


Figure A.8: **Reconstruction error in function of the distance to the training set for kPCA with RBF kernel.** The distance is computed between the test conformation and the convex hull defined by the training conformations in the low-dimensional representation space. It is normalised by the number of residues.

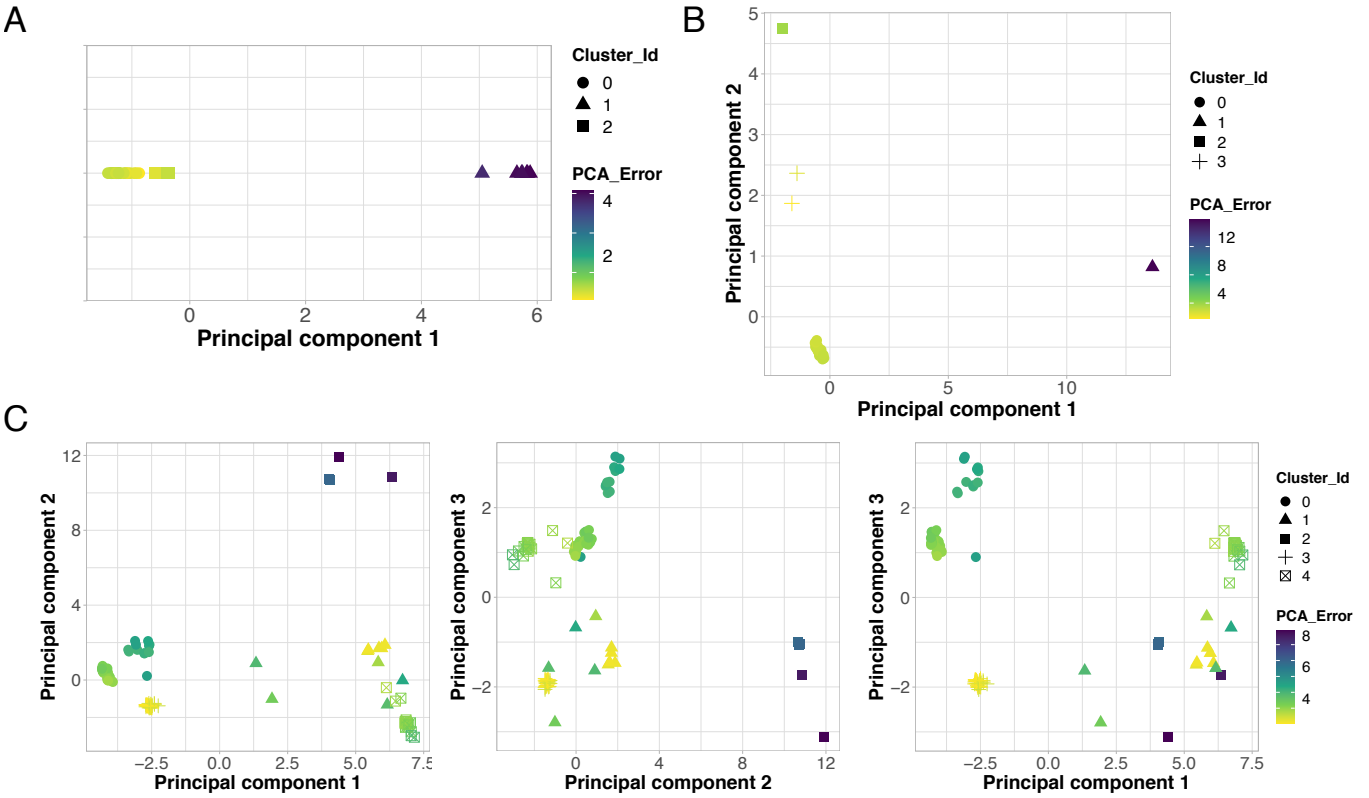


Figure A.9: **PCA feature spaces for three proteins from the benchmark.** We show the projections of the conformations in the l-dimensional PCA feature space, where l is the number of principal components needed to explain 90% of the total positional variance, for ADK (A), MurD (B) and ATPase (C). The point shapes indicate the clusters to which the conformations belong as determined by k-means clustering where $k = l + 2$. The colors reflect the RMSD reconstruction error (in Å). We reconstructed each conformation using the principal components computed from the set of conformations not belonging to the same cluster.

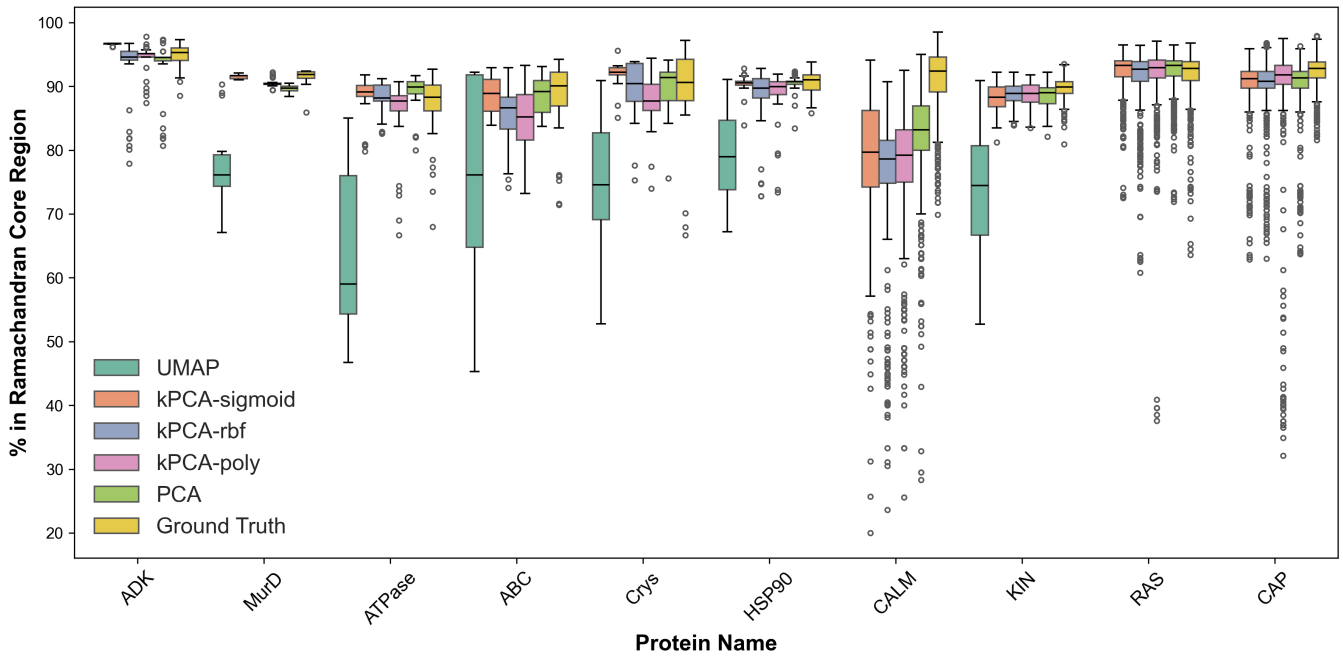


Figure A.10: **Distributions of the percentage of residues in the core region of the Ramachandran plot for each ensemble in the benchmark set.** We systematically reconstructed each conformation through a leave-one-cluster-out cross-validation procedure (see *Methods*). We set the hyperparameters of the kPCA and UMAP to the values yielding the best reconstruction, for each ensemble. The protein names on the x-axis are ordered according to motion complexity. The Ramachandran core region indicates the most favoured phi-psi angle combinations.

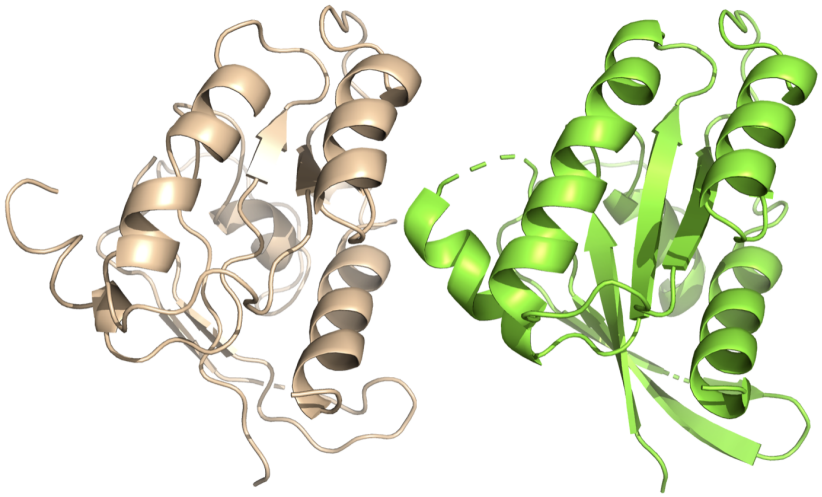


Figure A.11: **X-ray crystallographic structure of RAS (PDB code: 1PPL, chain A, in beige) and its PCA reconstruction (in green).** The PCA reconstruction displays a better secondary structure. The original structure has 63.9% of its residues in Ramachandran core region, while the PCA reconstruction has 96.4%.

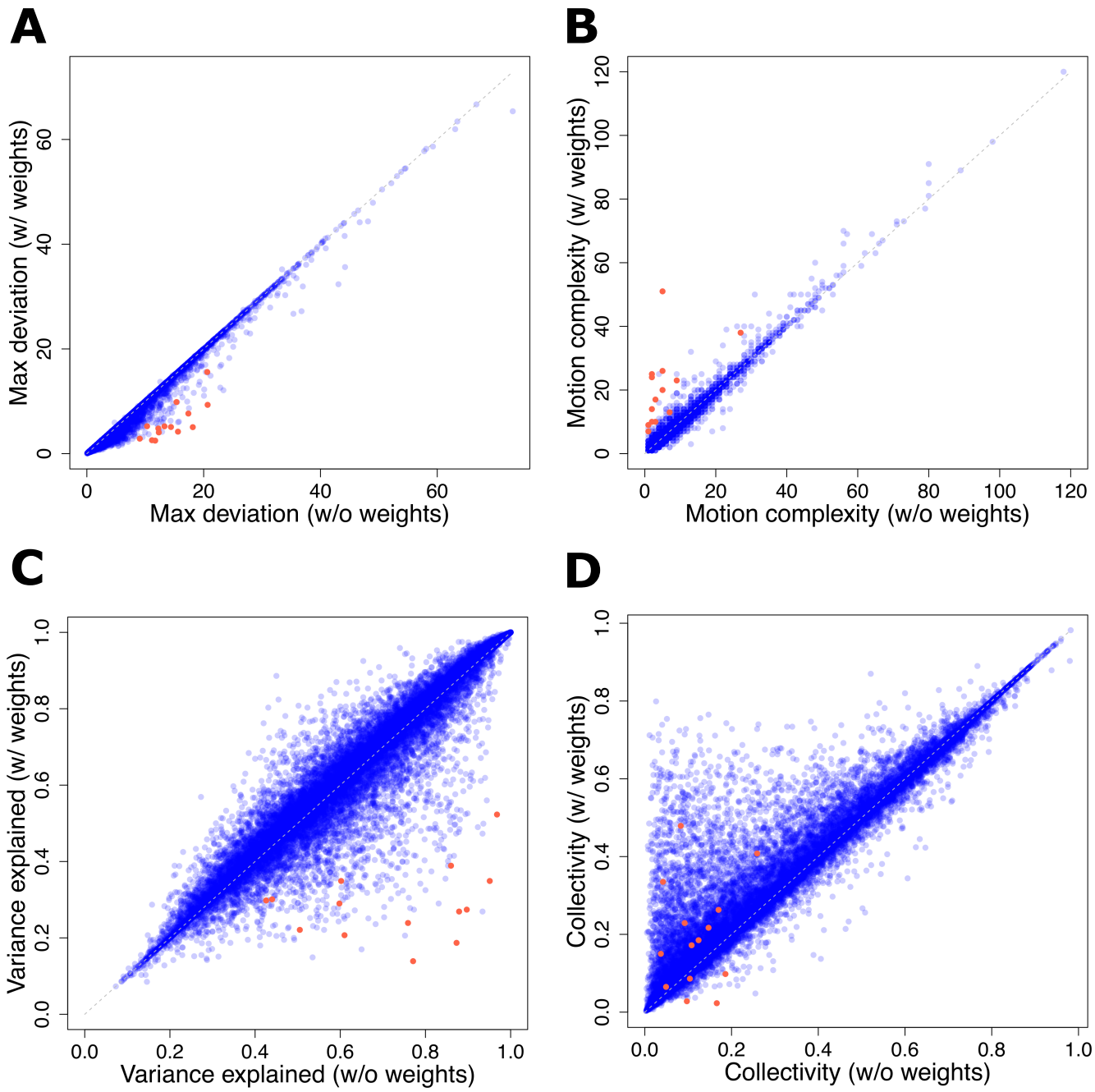


Figure A.12: **Influence of data uncertainty handling.** We compare some properties of the ensembles obtained at l_{80}^{80} when applying the weighting scheme accounting for uncertainty versus without weights. We consider only the ensembles with at least 3 members. **A.** Largest deviation between any two conformations (in Å). **B.** Motion complexity (in mode). **C.** Percentage of the variance explained by the most contributing linear motion. **D.** Collectivity of the most contributing linear motion. We highlight the ensembles for which applying the weighting scheme leads to a maximum deviation decrease of more than 5Å and an increased motion complexity by more than 5 modes in red.

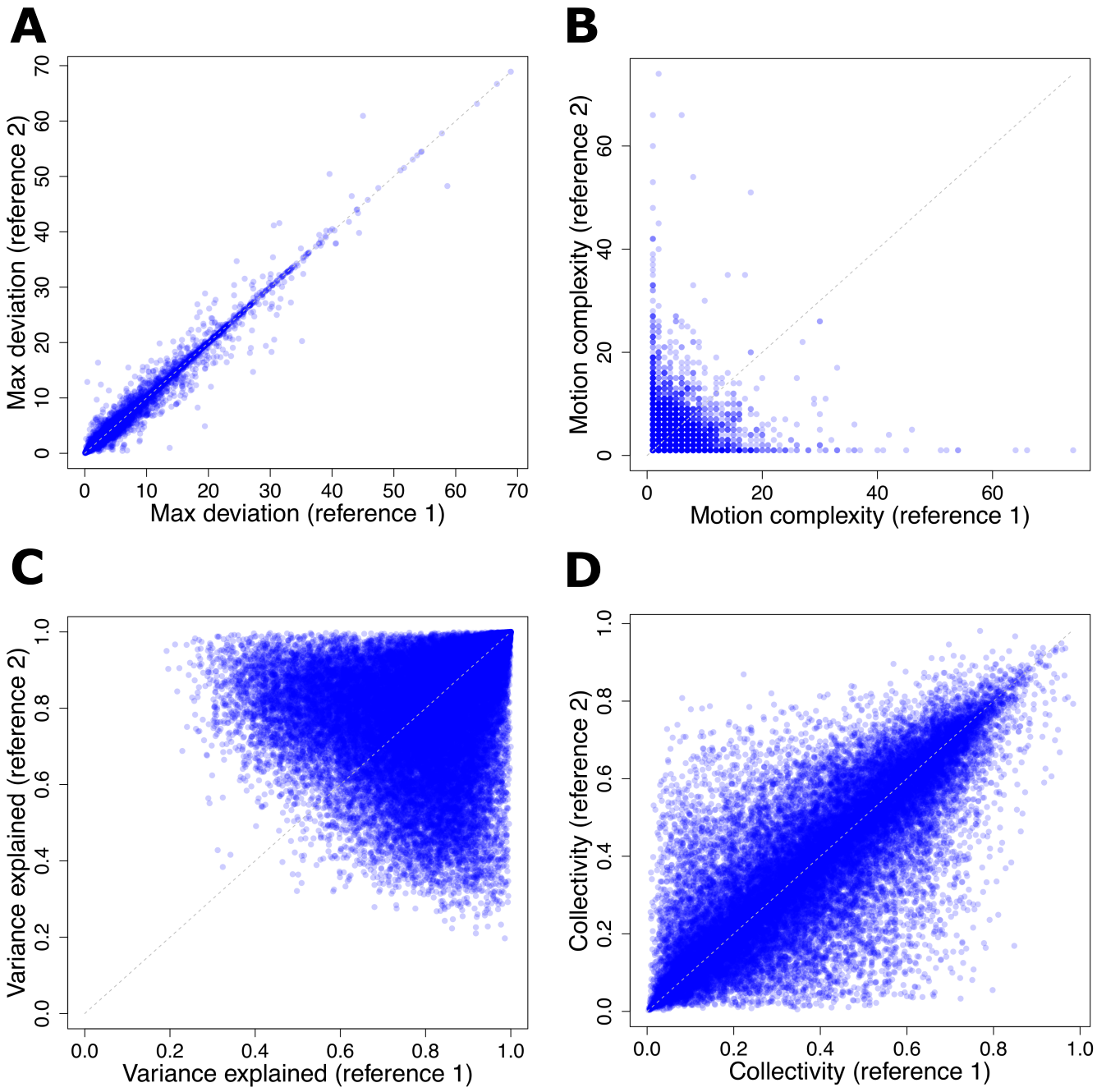


Figure A.13: **Influence of data conformation-specific centring.** We compare some properties of the ensembles obtained at l_{80}^{80} when superimposing and centring the conformations with respect to two different references. The first one is the closest to the multiple sequence alignment consensus (see *Materials and methods*). The second one has the highest RMS deviation from the first one. We consider only the ensembles with at least 3 members. **A.** Largest deviation between any two conformations (in Å). **B.** Motion complexity (in mode). **C.** Percentage of the variance explained by the most contributing linear motion. **D.** Collectivity of the most contributing linear motion.

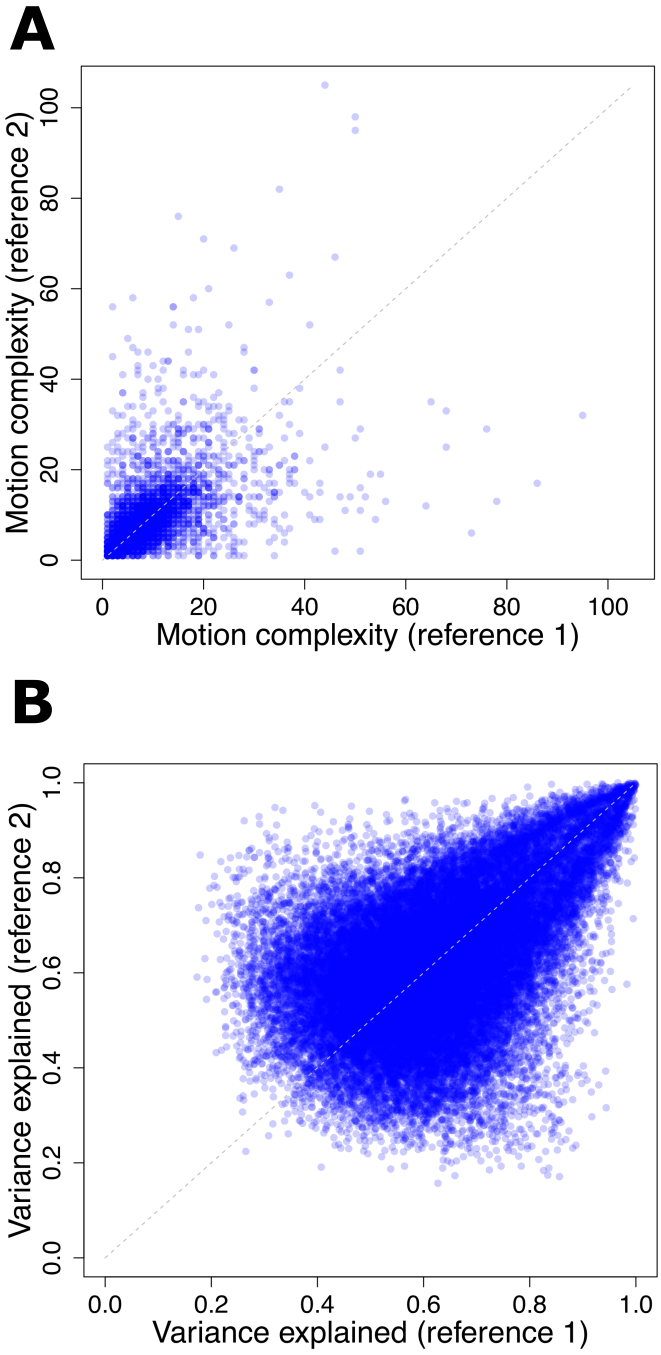


Figure A.14: **Influence of data conformation-specific centring when extracting motions from the correlation matrix.** We consider the ensembles with at least 3 members produced at l_{80}^{80} when superimposing and centring the conformations with respect to two different references. The first one is the closest to the multiple sequence alignment consensus (see *Materials and methods*). The second one has the highest RMS deviation from the first one. **A.** Motion complexity (in mode). **B.** Percentage of the variance explained by the most contributing linear motion. These plots can be compared with panels B and C from Figure S9.

Appendix B

Additional information for SeaMoon

SeaMoon: from protein language models to continuous structural heterogeneity

Valentin Lombard[1], **Dan Timsit**[1], **Sergei Grudinin**[*2], **Elodie Laine**[*1,3]

[1] Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France.
[2] Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.
[3] Institut universitaire de France (IUF).

[*] corresponding authors: sergei.grudinin@univ-grenoble-alpes.fr, elodie.laine@sorbonne-universite.fr

We compute the ground-truth motions with DANCE [339] by applying Principal Component Analysis on the 3D coordinates of a collection of protein conformations superimposed onto one another. The superimposition puts the protein conformations' centers of mass to zero and then aims at determining the optimal least-squares rotation matrix minimizing the Root Mean Square Deviation (RMSD) between any conformation and a reference conformation. Let us first derive some preliminary properties of this optimal rigid superposition.

Implications of optimal superimposition

Let us consider two sets of 3-vectors $\mathcal{A} = \{\vec{a}_i, i = 1, ..., m\}$ and $\mathcal{B} = \{\vec{b}_i, i = 1, ..., m\}$, such that a rigid transformation defined by the translation vector $\vec{T}$ and the rotation matrix R minimizes their squared geometric mismatch:

$$\sum_i (\vec{a}_i - R\vec{b}_i - \vec{T})^2 \rightarrow \min . \tag{B.1}$$

By taking the derivative of the above equation with respect to the rigid translation $\vec{T}$, we obtain

$$\sum_i (\vec{a}_i - R\vec{b}_i - \vec{T}) = \vec{0}. \tag{B.2}$$

Similarly, by taking the derivative of the above equation with respect to the rotation matrix R about x, y, and z axes, we obtain

$$\sum_i \vec{b}_i \times (\vec{a}_i - R\vec{b}_i - \vec{T}) = \vec{0}, \tag{B.3}$$

where $\vec{b}_i \times (\vec{a}_i - R\vec{b}_i - \vec{T})$ is the angular velocity of the particle i from $\mathcal{B}$. Therefore, the total translational force and the total rotational force or torque applied to a set of points in a rigid body must be zero at the optimal superposition conformation.

Ground-truth motions' translational and rotational constraints

The Cartesian coordinates of the 3D protein conformations in a collection are stored in a matrix X of dimension $3m \times n$. Their positional covariance matrix C of dimensions $3m \times 3m$ is expressed as,

$$C = \frac{1}{n-1} XX^T. \tag{B.4}$$

The eigen decomposition of the covariance matrix $C = VDV^T$ leads to a set of eigenvectors $\{V_{\cdot j}, j = 1, ..., 3m\}$ which are the columns of the matrix V, associated with the set of eigenvalues $\{\lambda_j, j = 1, ..., 3m\}$ stored in the diagonal matrix D. We show below that these eigenvectors comply with specific translational and rotational constraints.

Translational constraint. Since the conformations are centred at the origin, we have $\sum_{l=1}^{m} \vec{x}_{lj} = \vec{0}$ for every conformation $j = 1, ..., n$, with $\vec{x}_{lj}$ the 3-component vector from X corresponding to atom l. It follows that the elements of each x,y,z component in each ow of the covariance matrix C sum to zero,

$$\sum_{x_j, y_j, z_j = 1}^{m} C_{ij} = \frac{1}{n-1} \sum_{x_j, y_j, z_j = 1}^{m} \sum_{k=1}^{n} X_{ik} X_{jk} = \frac{1}{n-1} \sum_{k=1}^{n} X_{ik} \sum_{x_j, y_j, z_j = 1}^{m} X_{jk} = 0. \tag{B.5}$$

Without loss of generality, we can define coordinate indexes as $j_x = 3j$, $j_y = 3j + 1$, and $j_z = 3j + 2$. According to the eigen decomposition, we can then write, $\forall i$, $i = 1, ..., 3m$,

$$\sum_{x_j,y_j,z_j=1}^{m} C_{ij} = \sum_{x_j,y_j,z_j=1}^{m} \sum_{k=1}^{3m} \lambda_k V_{ik} V_{jk} = 0. \tag{B.6}$$

Summing up over the x,y, and z components of the rows of C, we obtain,

$$\sum_{x_i,y_i,z_i=1}^{m} \sum_{x_j,y_j,z_j=1}^{m} C_{ij} = \sum_{x_i,y_i,z_i=1}^{m} \sum_{x_j,y_j,z_j=1}^{m} \sum_{k=1}^{3m} \lambda_k V_{ik} V_{jk} = 0, \tag{B.7}$$

which can be rewritten as,

$$\lambda_1 \left(\sum_i V_{i1}^x \right)^2 + \lambda_2 \left(\sum_i V_{i1}^y \right)^2 + \lambda_3 \left(\sum_i V_{i1}^z \right)^2 + \lambda_4 \left(\sum_i V_{i2}^x \right)^2 \lambda_5 \left(\sum_i V_{i2}^y \right)^2 + \lambda_6 \left(\sum_i V_{i2}^z \right)^2 \tag{B.8}$$

$$+ ... + \lambda_{3m-2} \left(\sum_i V_{i(m)}^x \right)^2 + \lambda_{3m-1} \left(\sum_i V_{i(m)}^y \right)^2 + \lambda_{3m} \left(\sum_i V_{i(m)}^z \right)^2 = 0. \tag{B.9}$$

Since the covariance matrix C is symmetric and positive semi-definite, all the eigenvalues λ_k are real and non-negative and it follows that the sum of components of each eigenvector $V_{.k}$ corresponding to a positive eigenvalue λ_k must be zero. And the number of such eigenvectors equals the rank of the covariance matrix.

Rotational constraint. We can express the angular velocity of an atom l from conformation j from our set as $\vec{r}_{lj} \times \vec{x}_{lj}$, where $\vec{r}_{lj}$ is its 3D position relative to the conformation's center of mass, *i.e.*, the rotation center, and $\vec{x}_{lj}$ is its 3D displacement vector from the reference conformation. This angular velocity can be re-written as $(\vec{r}_l + \vec{x}_{lj}) \times \vec{x}_{lj} = \vec{r}_l \times \vec{x}_{lj}$, where $\vec{r}_l$ is the 3D position of the atom l in the *reference* conformation relative its center of mass. We can further rewrite this vector product in a matrix form as $[r_l]_\times \vec{x}_{lj}$, where $[r_l]_\times$ is a 3×3 skew-symmetric matrix that corresponds to the cross product operation. Let us also define a $3m \times 3m$ block-diagonal matrix R formed of m matrices $[r_i]_\times$, $R \equiv \text{diag}([r_1]_\times, [r_2]_\times, ..., [r_m]_\times)$.

Since the global rotations between protein conformations, with respect to the reference, have been removed during superimposition, the angular velocities for any conformation j result in a null vector, $\sum_{l=1}^{m} [r_l]_\times \vec{x}_{lj} = \vec{0}$. Or, in the matrix form, for each column (conformation) j, $\sum_{i=1}^{m} (RX)_{i_x j} = 0$, $\sum_{i=1}^{m} (RX)_{i_y j} = 0$, $\sum_{i=1}^{m} (RX)_{i_z j} = 0$, where, without loss of generality, we can define coordinate indexes as $i_x = 3i$, $i_y = 3i + 1$, and $i_z = 3i + 2$. It follows that,

$$\forall a \in \{x, y, z\} : \quad \sum_{j=1}^{m} (RX(RX)^T)_{ij_a} = \sum_{j=1}^{m} \sum_{k=1}^{n} (RX)_{ik} (RX)_{j_a k} = \sum_{k=1}^{n} (RX)_{ik} \sum_{j=1}^{m} (RX)_{j_a k} = 0. \tag{B.10}$$

The decomposition of $RX(RX)^T$ leads to, $\forall i$, $i = 1, ..., 3m$,

$$\forall a \in \{x, y, z\} : \quad \sum_{j=1}^{m} (RX(RX)^T)_{ij_a} = \sum_{j=1}^{m} (R[\sum_k \lambda_k V_k V_k^T] R^T)_{ij_a} = \sum_k \lambda_k \sum_{j=1}^{m} (RV_k [RV_k]^T)_{ij_a} = 0 \tag{B.11}$$

Summing up over the x,y, and z components of the rows of the previous system, we obtain,

$$\lambda_1 \left(\sum_i (RV_1)_{i_x} \right)^2 + \lambda_1 \left(\sum_i (RV_1)_{i_y} \right)^2 + \lambda_1 \left(\sum_i (RV_1)_{i_z} \right)^2 + \tag{B.12}$$

$$\lambda_2 \left(\sum_i (RV_2)_{i_x} \right)^2 + \lambda_2 \left(\sum_i (RV_2)_{i_y} \right)^2 + \lambda_2 \left(\sum_i (RV_2)_{i_z} \right)^2 + \tag{B.13}$$

$$... + \lambda_{3m} \left(\sum_i (RV_{3m})_{i_x} \right)^2 + \lambda_{3m} \left(\sum_i (RV_{3m})_{i_y} \right)^2 + \lambda_{3m} \left(\sum_i (RV_{3m})_{i_z} \right)^2 = 0. \tag{B.14}$$

Thus, the positional eigenvectors must have additional rotational constraints, $\forall k,\ k = 1, ..., 3m$, $\sum_i (RV_k)_{i_x} = 0$,$\sum_i (RV_k)_{i_y} = 0$, and $\sum_i (RV_k)_{i_z} = 0$ for all eigenvectors with the corresponding non-zero eigenvalues.

Orienting a predicted motion with respect to a 3D conformation

We exploit the rotational constraints of the ground-truth motions, *i.e.*, the eigenvectors of the positional covariance matrix, to align the motion vectors predicted by SeaMoon on a given protein 3D conformation. More specifically, we aim at determining the rotation $R \in \mathrm{SO}(3)$ that minimizes the overall angular velocity of the conformation subjected to the predicted motion. If we denote $(v^i)_{i \leq m}$ the 3D displacement vectors predicted by SeaMoon for m protein atoms, the problem we solve is $\sum_i (Rv^i) \times r^i = 0$ for the rotation R, where r^i is the 3D positional vector of atom i.

For coordinate $d \in \{1, 2, 3\}$, we can then write, using Einstein's notation:

$$\begin{aligned} 0 &= e_d.(Rv^i) \times r^i \\ &= \varepsilon_{pqd} (Rv^i)_p r^i_q \\ &= \varepsilon_{pqd} R_{pk} v^i_k r^i_q \\ &= R_{pk} \varepsilon_{pqd} v^i_k r^i_q \\ &= \mathrm{Tr}(RA^T_d) \end{aligned}$$

where $(A_d)_{pk} = \varepsilon_{pqd} v^i_k r^i_q$ and where $\varepsilon_{i,j,k}$ is 0 if i, j, k are not different and elsewise is equal to the signature of the permutation

$$\begin{pmatrix} 1 & 2 & 3 \\ i & j & k \end{pmatrix}$$

To solve these equations, we use 4-dimensional quaternions which have fewer dimensions than 3×3 matrices, associating to a unitary quaternion the rotation matrix [359]:

$$R_q = \begin{pmatrix} q_1^2 + q_2^2 - q_3^2 + q_4^2 & 2(q_2 q_3 + q_1 q_4) & 2(q_2 q_4 - q_1 q_3) \\ 2(q_2 q_3 - q_1 q_4) & q_1^2 + q_3^2 - q_2^2 + q_4^2 & 2(q_3 q_4 + q_1 q_2) \\ 2(q_2 q_4 + q_1 q_3) & 2(q_3 q_4 - q_1 q_2) & q_1^2 + q_4^2 - q_2^2 + q_3^2 \end{pmatrix} \tag{B.15}$$

We can develop the expression $\mathrm{Tr}(RA^T)$ to give:

$$\begin{aligned}\mathrm{Tr}(RA^T) = &(-2q_1q_2 + 2q_3q_4)A_{3,2} + (2q_1q_2 + 2q_3q_4)A_{2,3} \\ &+ (-2q_1q_3 + 2q_2q_4)A_{1,3} + (2q_1q_3 + 2q_2q_4)A_{3,1} \\ &+ (-2q_1q_4 + 2q_2q_3)A_{2,1} + (2q_1q_4 + 2q_2q_3)A_{1,2} \\ &+ (q_1^2 - q_2^2 - q_3^2 + q_4^2)A_{3,3} + (q_1^2 - q_2^2 + q_3^2 - q_4^2)A_{2,2} + (q_1^2 + q_2^2 - q_3^2 - q_4^2)A_{1,1}\end{aligned}$$

Writing the same equations changing A for A_m for $m = 1, 2, 3$, we get a system of 3 quadratic equations for 4 unknowns by setting the previous terms to 0, to which we add the unitary constraint $q_0^2 + q_1^2 + q_3^2 + q_4^2 = 1$. We solve this system using the symbolic package *wolframclient* in Python, yielding at most 4 quaternions as solutions which we then transform in rotation matrices.

Supplemental tables and figures

Layer type	Output shape	Filter size	#(Parameters)
Input	ESM2:(L, 1 280) ESM3:(L, 1 536) ProstT5:(L, 1 024)	-	0
Conv1D	(256, L)	1	ESM2: 327 936 ESM3: 393 216 ProstT5: 262 400
Conv1D	(128, L)	15	491 648
Conv1D	(64, L)	31	254 016
K Conv1D	(K, 3, L)	1	$K \times 195$

Table B.1: **Description of SeaMoon neural network architecture.** We report the layer types, output shapes, kernel sizes, and parameters for ESM2-, ESM3-, and ProstT5-based SeaMoon models. The length (in amino acid) of the input sequence is denoted as L and the number of predicted vectors as K. The 1D convolutional layers with filter size of 1 are equivalent to linear layers.

Method	Input	pLM	Supervised	#(Train samples)
SeaMoon-ESM2	sequence	ESM2	✓	5 119
SeaMoon-ESM2(x5)	sequence	ESM2	✓	14 921
SeaMoon-ESM3	sequence	ESM3	✓	5 119
SeaMoon-ESM3(x5)	sequence	ESM3	✓	14 921
SeaMoon-ProstT5	sequence	ProstT5	✓	5 119
SeaMoon-ProstT5(x5)	sequence	ProstT5	✓	14 921
NMA	3D structure	×	×	0

Table B.2: **Description of the tested models and methods.** For SeaMoon-ESM2(x5), SeaMoon-ESM3(x5) and SeaMoon-T5(x5), we increased the number of training samples by defining up to 5 reference conformations per experimental collection.

Ablation applied	**# of proteins w. acceptable prediction**
Base model, SeaMoon-T5	439 (39.2%)
Network architecture:	
Size 1 kernel	271 (24.2%)
7-layer Transformer architecture	411 (36.7%)
Training loss:	
Without sign flip	413 (36.8%)
Without permutation	375 (33.4%)
Without reflection	402 (35.9%)
Input data:	
Random embeddings	119 (10.6%)
Positional encoding only	177 (15.8%)
Random baseline:	
Random neural network weights	0 (0.0%)

Table B.3: **Success rate in ablation study.** The transformer architecture comprises one linear layer from dimension 1024 to 128, seven Transformer layers of size 128, with 4 heads, and three 1D CNNs with kernel size 1 (same as the base model) going from dimension 128 to 3 for each predicted mode. It has a similar number of free parameters compared to the base model.

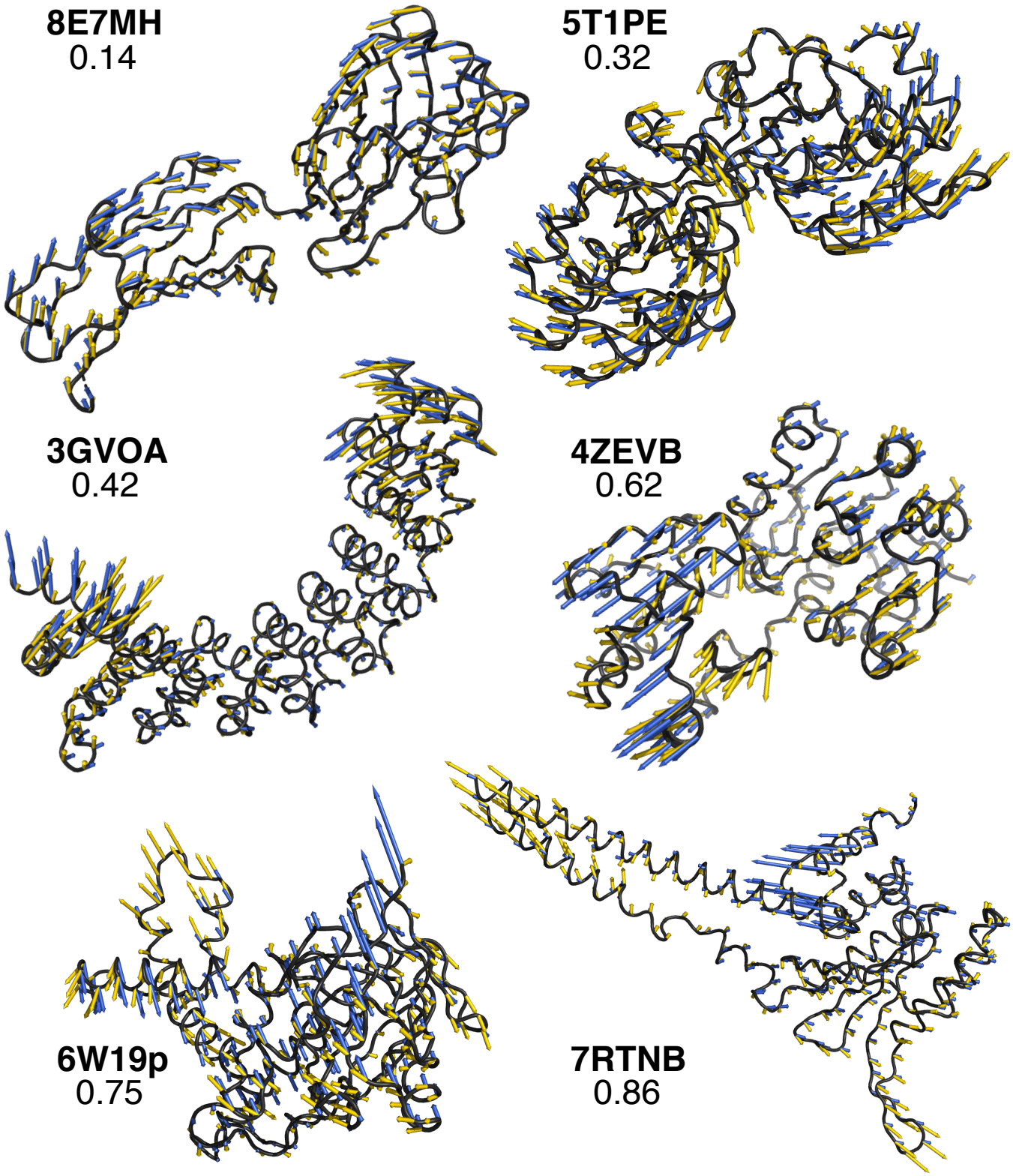


Figure B.1: **Examples of predictions.** They allow for a visual assessment of how well the predicted vectors (in blue) approximate the ground-truth motions (in yellow) at different levels of NSSE (indicated on each panel). For each example, the query conformation is shown in black cartoons and labelled with its PDB chain identifier (in bold). We obtained the predicted vectors with SeaMoon-ESM2(x5).

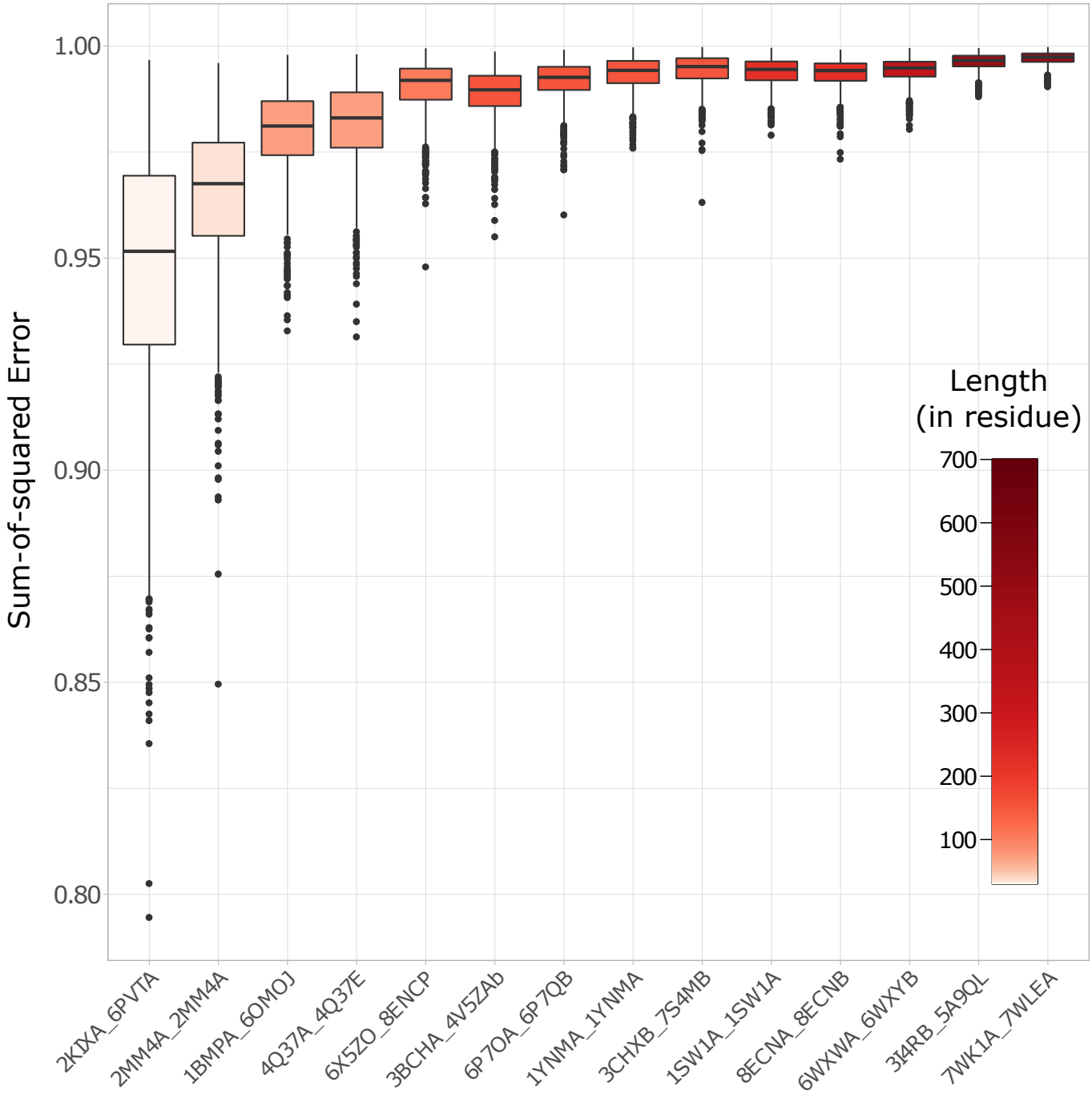


Figure B.2: **Normalised sum-of-squares errors for random predictions.** Distributions of normalised sum-of-squares errors computed after optimal rotation and scaling of 1000 random vectors against 14 ground-truth motions from the test set. The PDB chain identifiers of the corresponding proteins are given in x-axis. The boxes are colored according to the protein length.

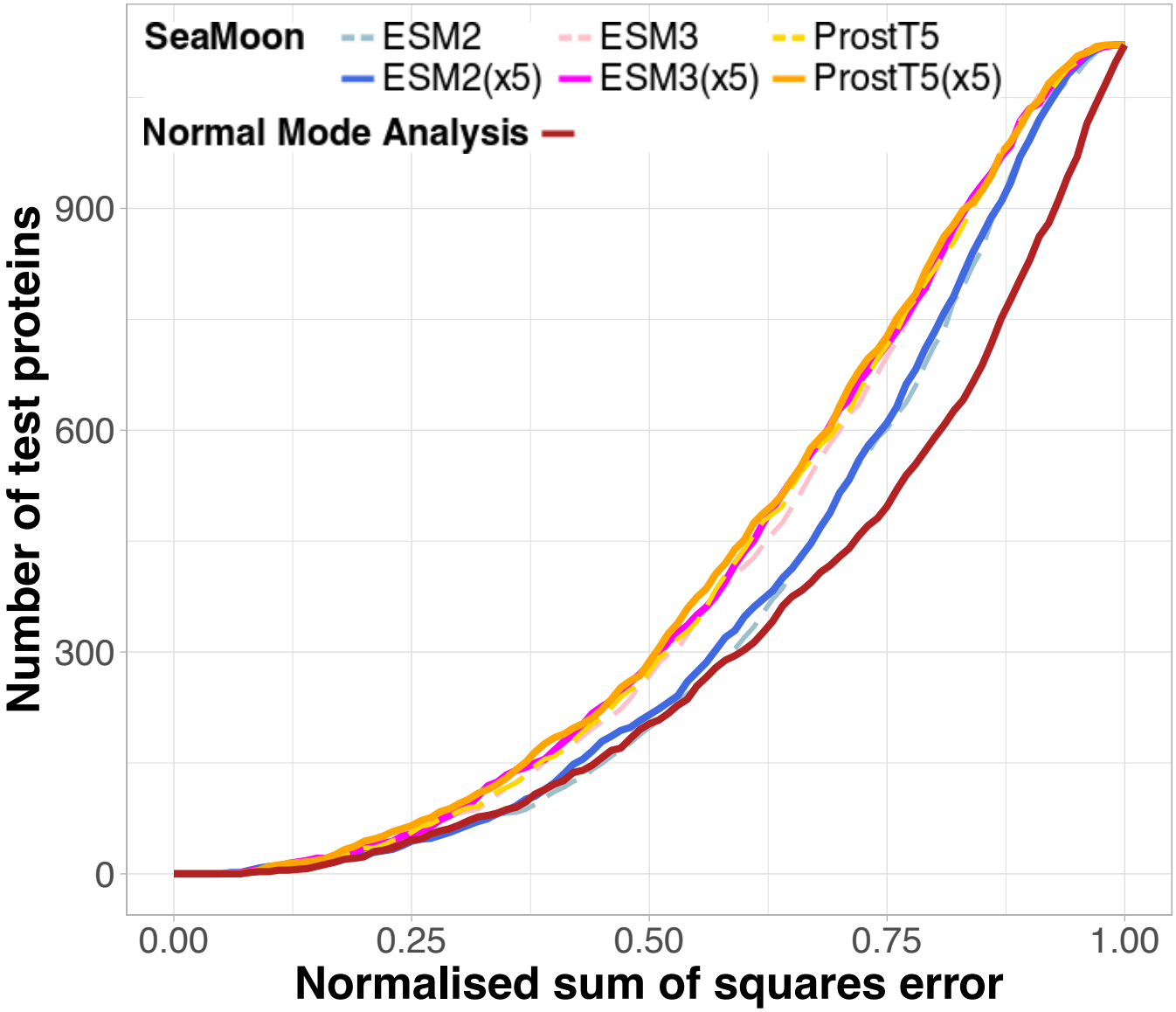


Figure B.3: **Performance on a test set of 1 121 proteins. A.** Comparison of the cumulative normalised sum-of-squares error (SSE) curves computed for different versions of SeaMoon and for the Normal Mode Analysis (NMA, performed with NOLB). For SeaMoon, we tested three pLMs, namely ESM2, ESM3, and ProstT5. During training, we gave only the reference conformation of each collection to SeaMoon-ESM2, SeaMoon-ESM3, and SeaMoon-ProstT5, while we gave 5 conformations per collection to SeaMoon-ESM2(x5), SeaMoon-ESM3(x5), and SeaMoon-ProstT5(x5).

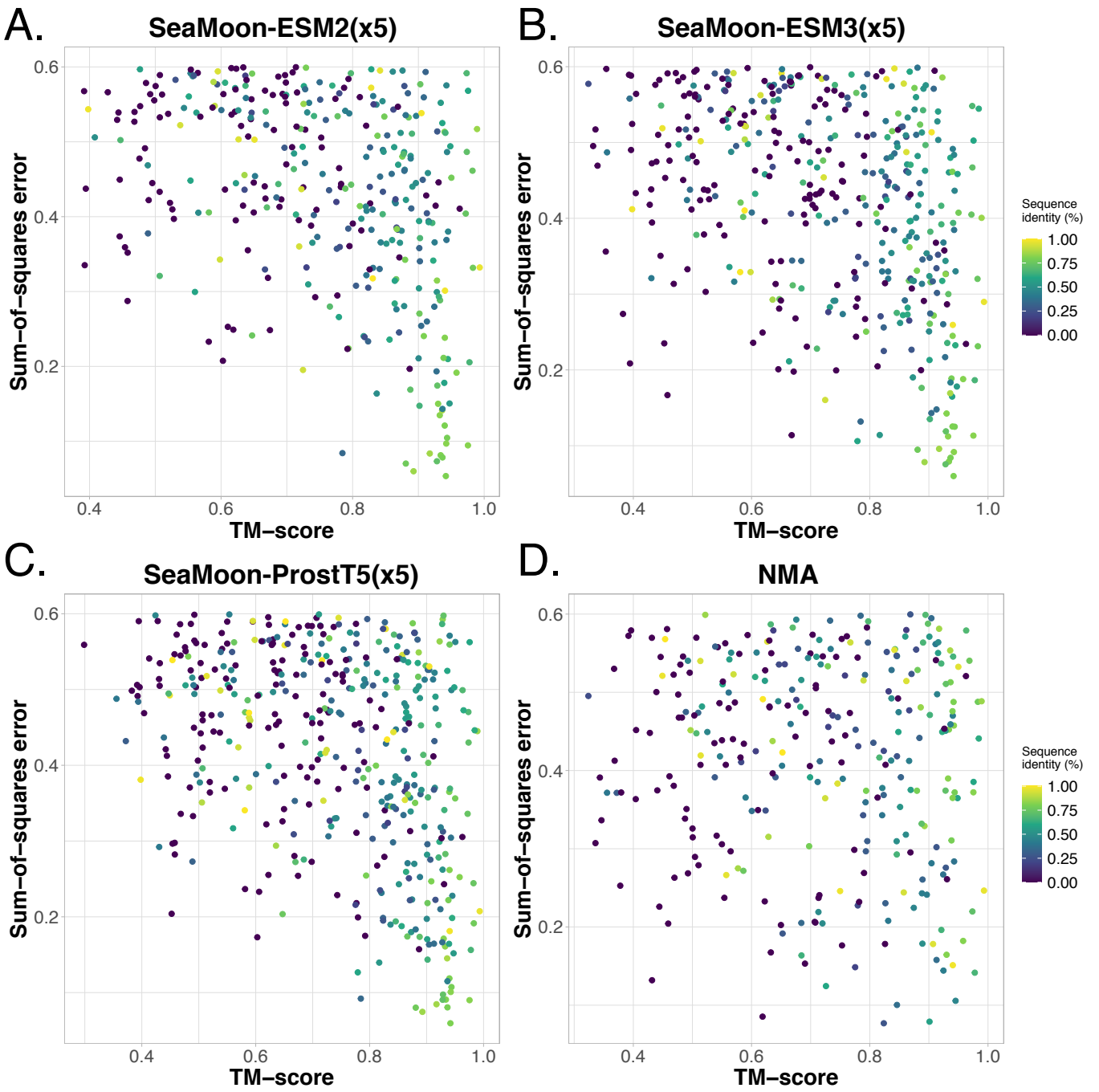


Figure B.4: **Influence of sequence and structure similarity.** Normalised sum-of-squares errors (y-axis) in function of the maximum TM-score (x-axis) and maximum sequence identity (color) of the test conformations computed over the whole training set. **A.** SeaMoon-ESM2(x5). **B.** SeaMoon-ESM3(x5). **C.** SeaMoon-ProstT5(x5). **D.** NMA.

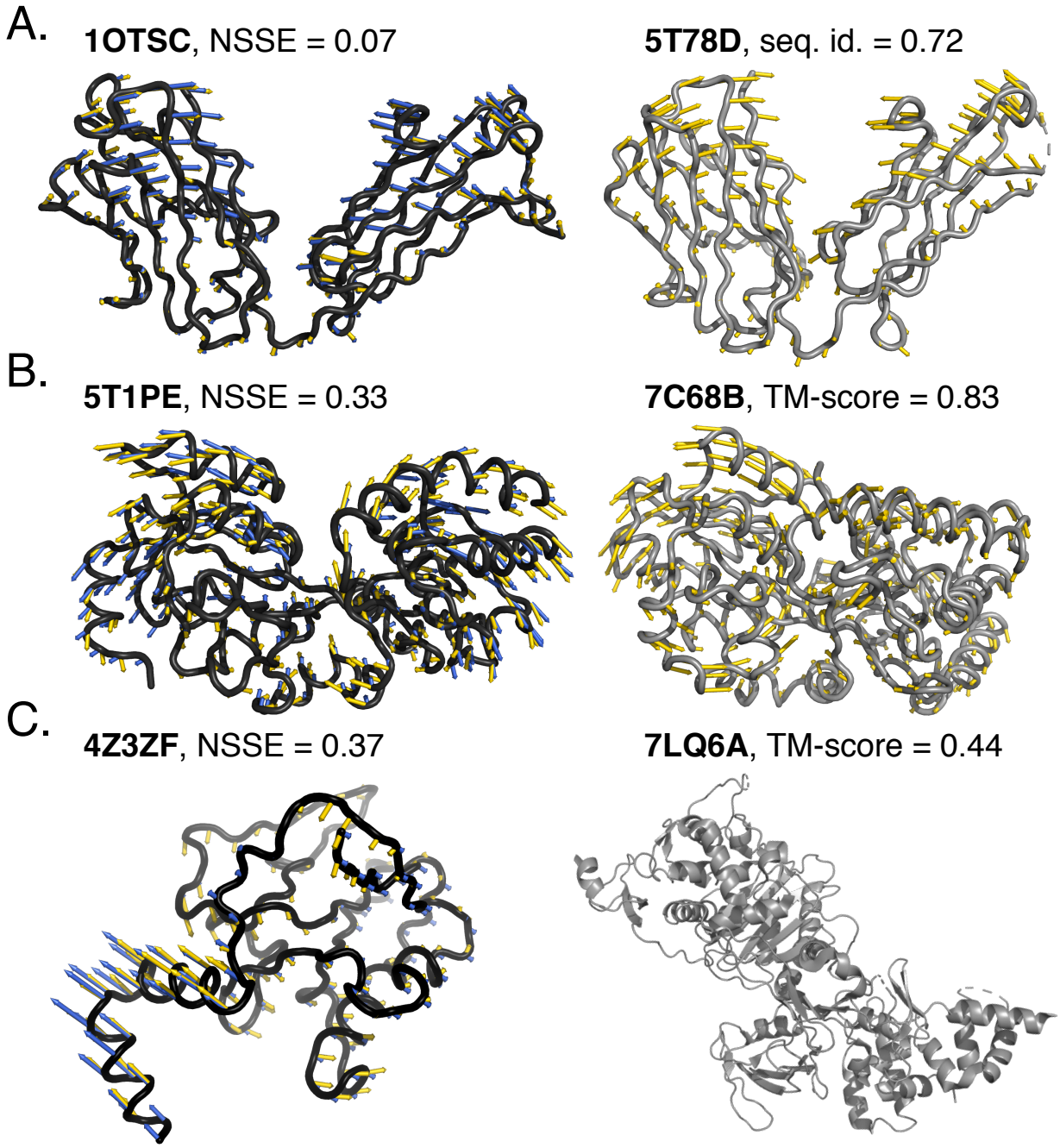


Figure B.5: **Examples of predictions for test proteins with decreasing similarity to the training set.** The conformations are shown in cartoons and labelled with their PDB chain identifier. The ground-truth and SeaMoon-ESM2(x5) predicted motions are depicted with yellow and blue arrows, respectively. Left, in black: test proteins. Right, in grey: closest proteins from the training set. **A.** Fab fragment (heavy chain), 221 residues, 107 conformations in the collection, collectivity *kappa* = 0.74 for the ground-truth motion. Its sequence, structure and main motion are highly similar to the Fab fragment displayed on the right **B.** Putative ABC transporter from *Campylobacter jejuni*, 326 residues, 8 conformations, *kappa* = 0.74. It does not have any detectable sequence similarity to the training set. Its structure and main motion bear some resemblance with the ABC transporter from *Thermus thermophilus* shown on the right. **C.** Iron-sulfur cluster-binding oxidoreductase, 170 residues, 20 conformations, *kappa* = 0.52. It does not have any detectable sequence similarity to the training set and the structurally closest training protein, the bacterial penicillin-binding protein 1B, exhibits a different 3D fold and different motions.

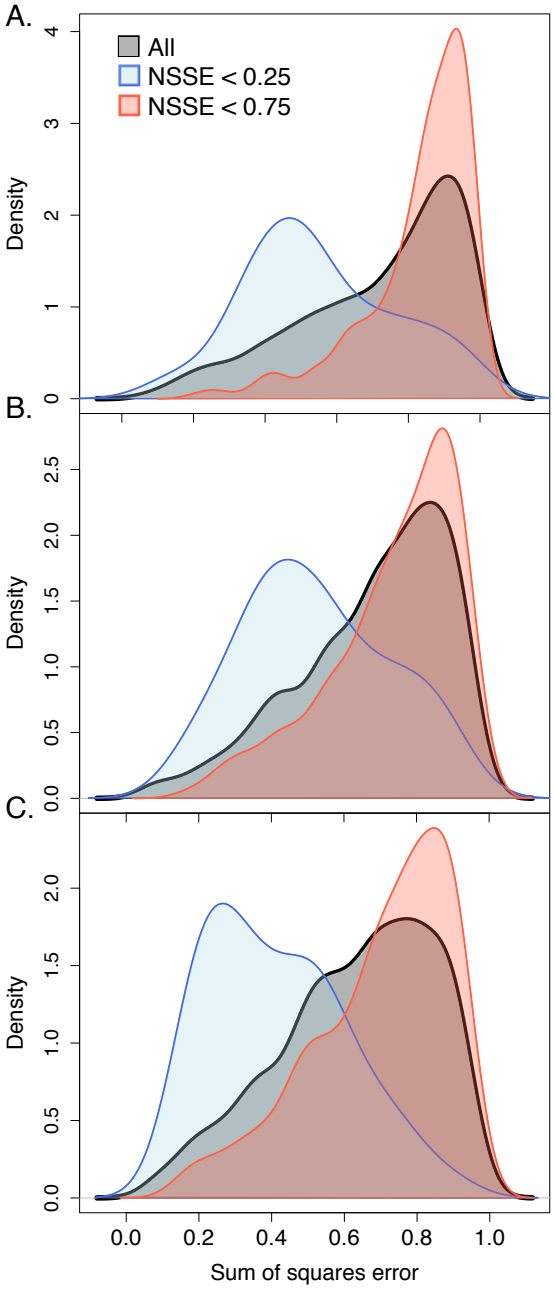


Figure B.6: **Agreement between a selection of methods. A.** Distribution densities of the NSSE computed for the NMA over the full test set (in black, 1 121 proteins) and over the subsets best-predicted ($NSSE < 0.25$, in blue, 36 proteins) and worst-predicted ($NSSE > 0.75$, in red, 326 proteins) by SeaMoon-ESM2(x5) and SeaMoon-ProstT5(x5). **B-C.** Distribution densities of the NSSE computed for SeaMoon-ESM2(x5) (B) and SeaMoon-ProstT5(x5) (C) over the full test set (in black) and over the subsets best-predicted ($NSSE < 0.25$, in blue, 46 proteins) and worst-predicted ($NSSE > 0.75$, in red, 624 proteins) by the NMA.

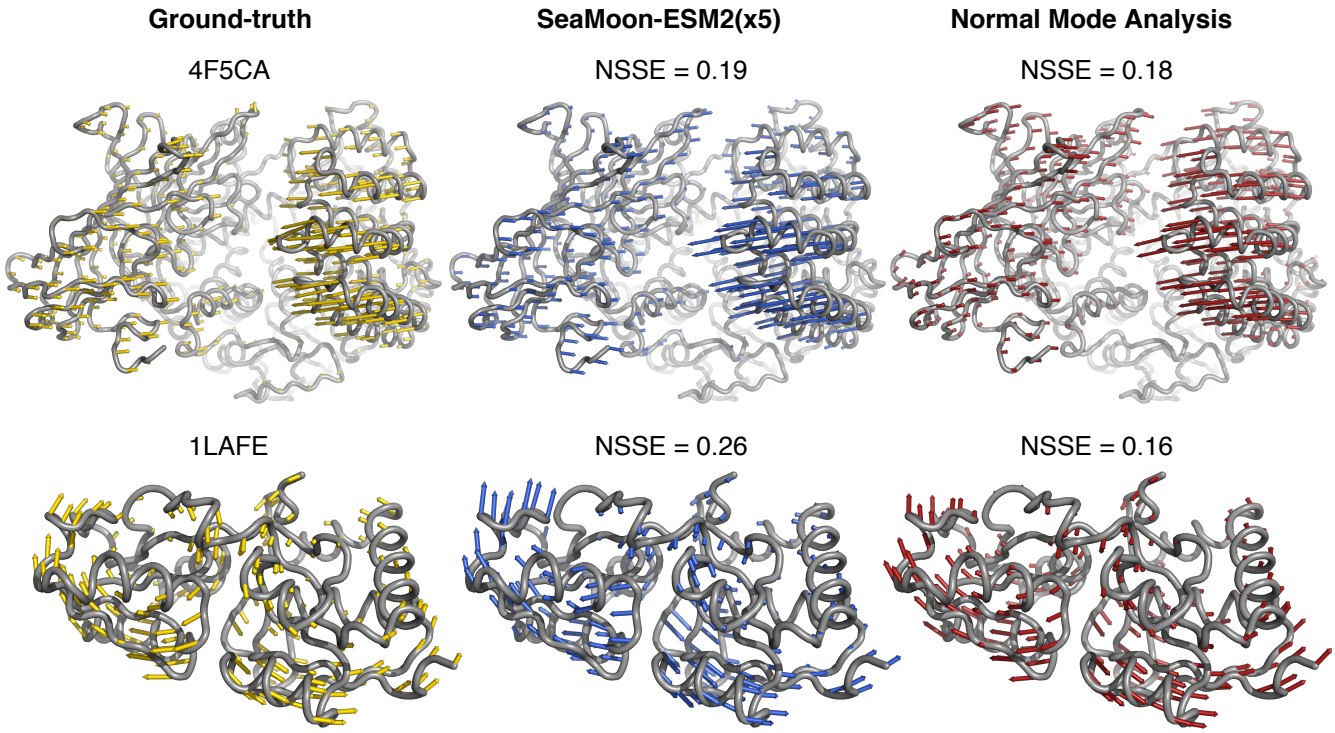


Figure B.7: **Examples of motions well predicted by SeaMoon-ESM2(x5) and the NMA.** The arrows depicted in yellow (left), blue (middle) and red (right) onto the 3D structure represent the ground-truth motion, the best-matching prediction from SeaMoon-ESM2(x5), and the best-matching prediction from the NMA. Top: Mammalian aminopeptidase N (PDB code: 4F5C, chain A). It shares 81% sequence similarity with a human aminopeptidase from the train set (TM-score = 0.96). Bottom: Bacterial periplasmic lysine-, arginine-, ornithine-binding protein (PDB code: 1LAF, chain E). It shares only 35% sequence similarity with its closest homolog from the train set, a nopaline-binding periplasmic protein from bacteria. Their structures are highly similar, with a TM-score of 0.91.

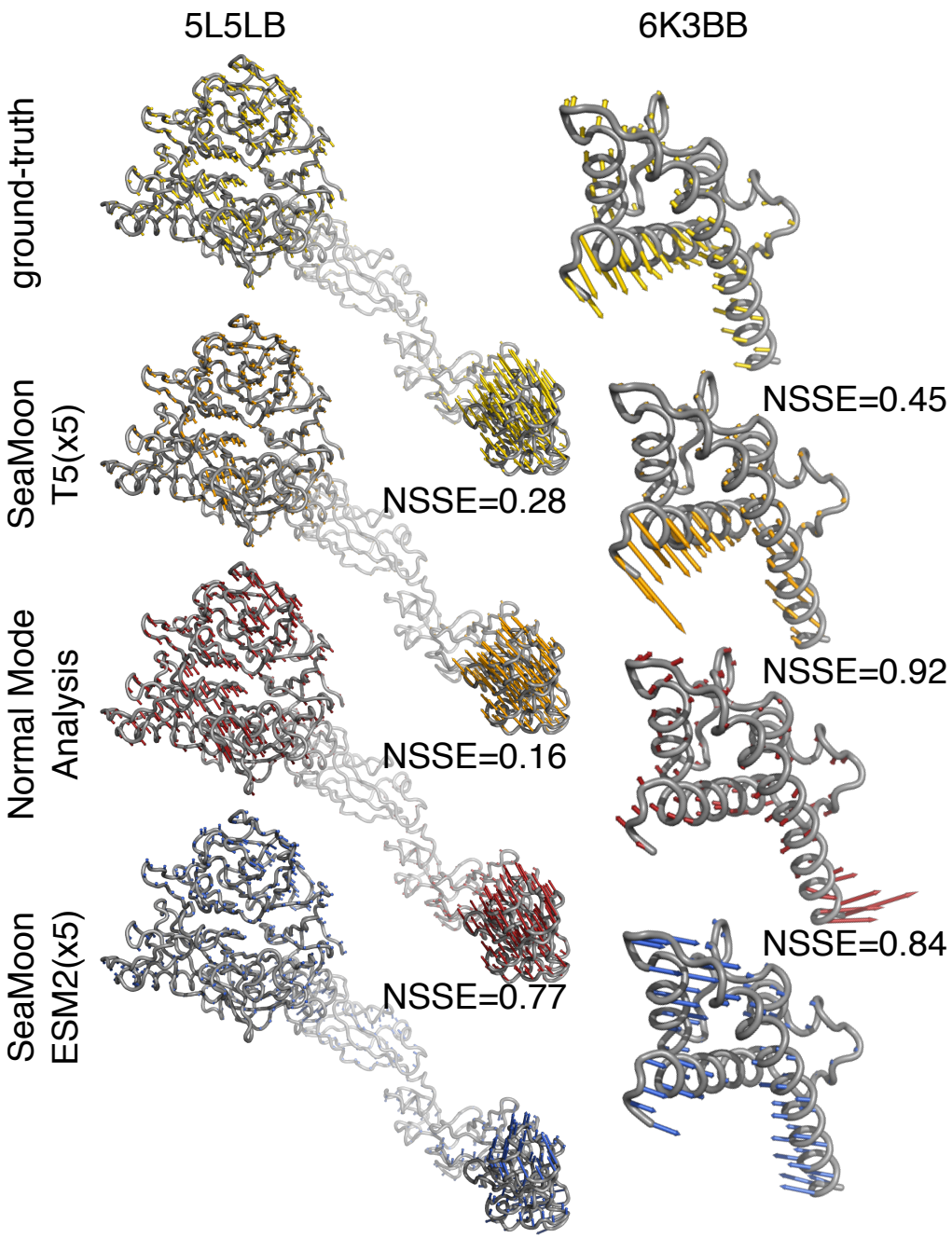


Figure B.8: **Examples of motions better captured by SeaMoon-ProstT5(x5) than SeaMoon-ESM2(x5).** The arrows depicted in yellow, orange, red, and blue onto the 3D structures represent the ground-truth motions and the best-matching predictions from SeaMoon-ProstT5(x5), the NMA, and SeaMoon-ESM2(x5), respectively. Left: Mammalian plexin A4 ectodomain (PDB code: 5L5L, chain B). It shares 64% sequence similarity with a plexin A2 ectodomain from the train set (TM-score = 0.68). Right: Legionella effector MavC (PDB code: 6K3B, chain B). It does not have any detectable sequence similarity with the training set and shares only a weak structural similarity (TM-score = 0.59) with the mammalian cytochrome P450 2B4.

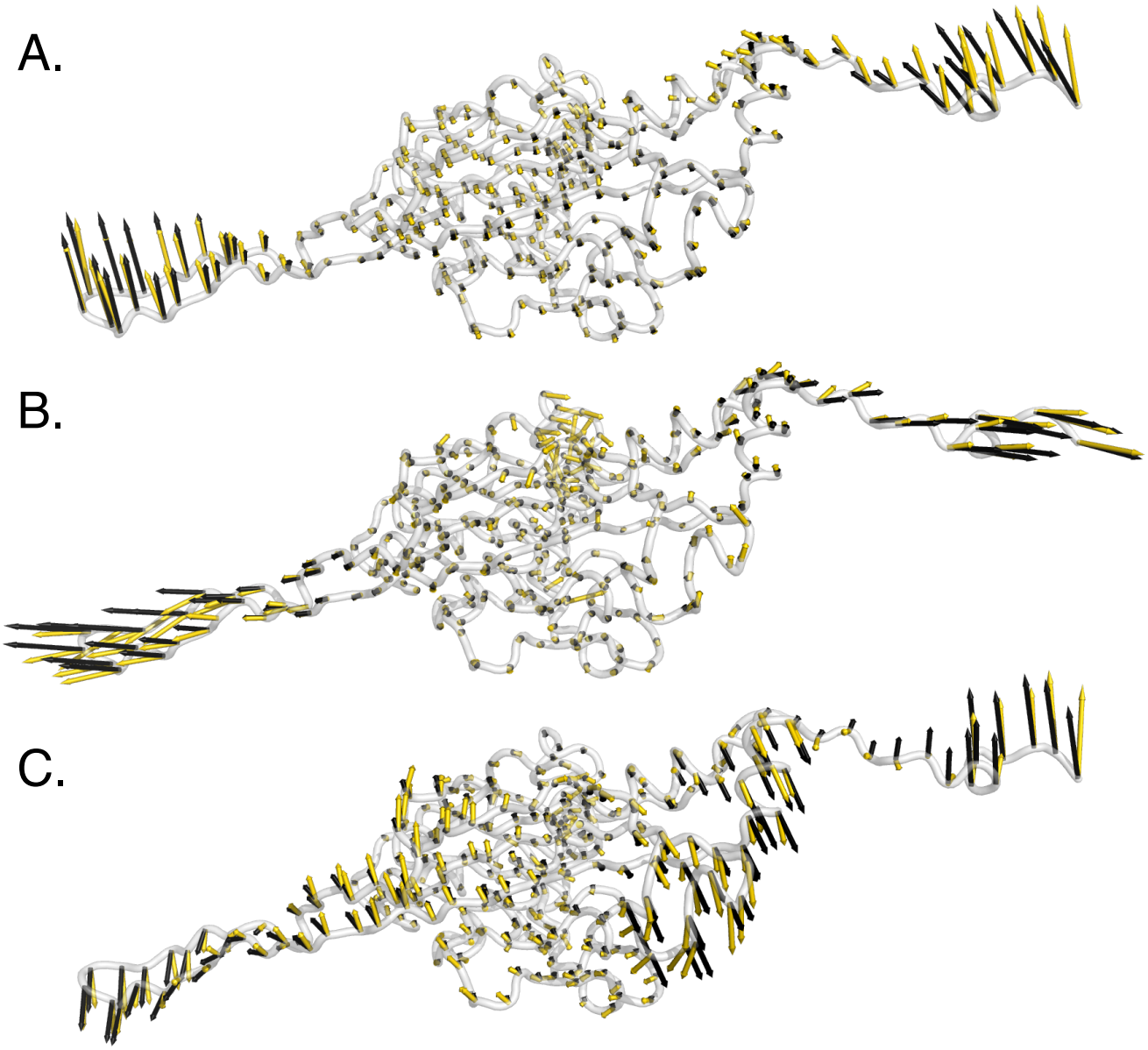


Figure B.9: **Ogopogo major capsid protein motion subspace.** Visualisation of the ground-truth (yellow) and predicted (black) motion relative directions and amplitudes for the best-matching pairs (1,1) (A), (3,2) (B) and (2,3) (C) on the reference 3D conformation, PDB code: 8ECN, chain B.

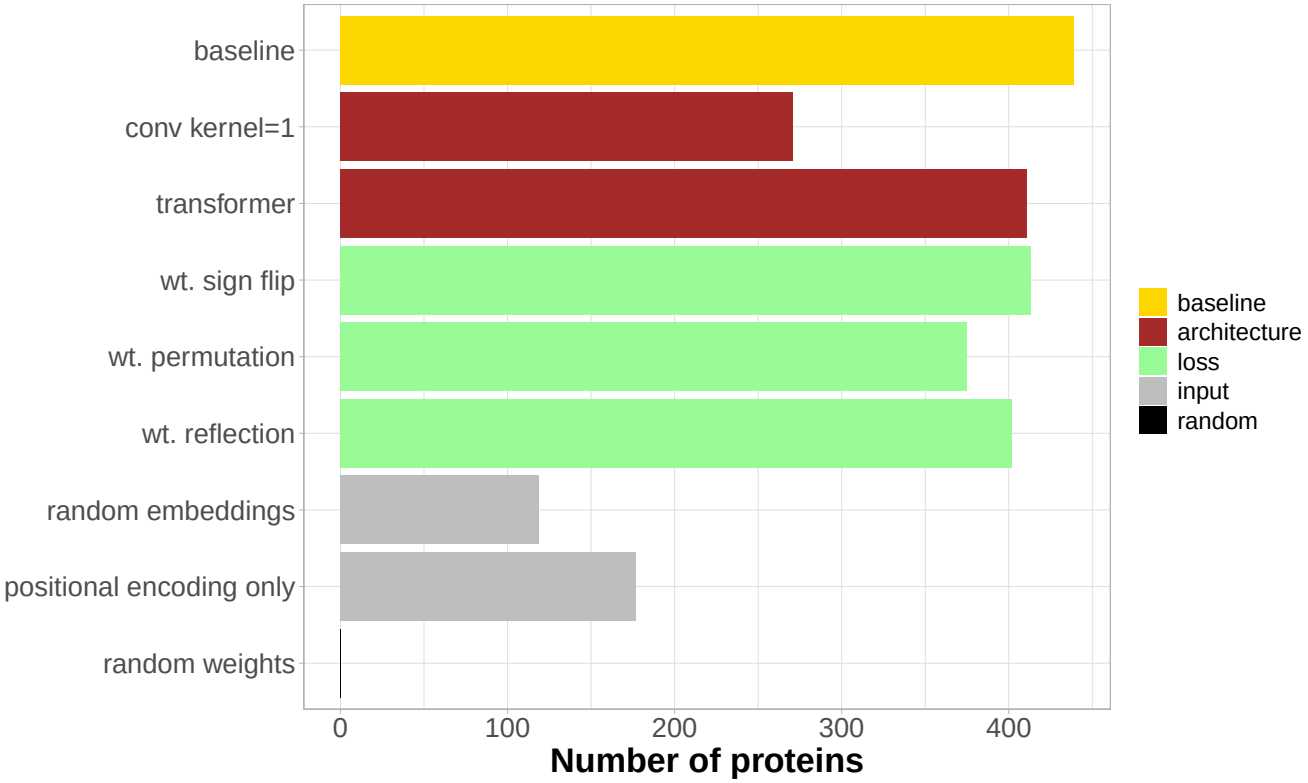


Figure B.10: **Ablation study** We report the number of test proteins with at least one acceptable prediction. The baseline model is SeaMoon-ProstT5.

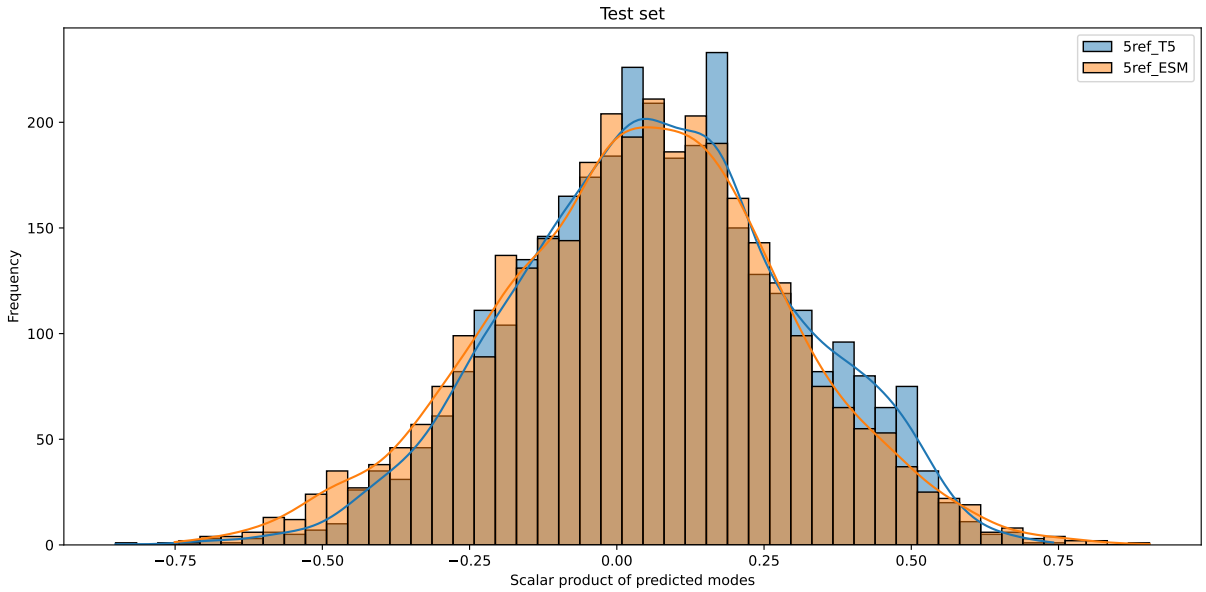


Figure B.11: **Pearson correlation computed between motions predicted by SeaMoon.** We performed an all-to-all pairwise comparison for each protein from the test set. About 95 percent of the pairs have an absolute Pearson correlation below 0.5.