SeaMoon: Prediction of molecular motions based on language models

Valentin Lombard, Dan Timsit, Sergei Grudinin, Elodie Laine

SeaMoon: Prediction of molecular motions based on language models

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

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 (∼1M trainable parameters) convolutional neural network. SeaMoon achieves a success rate of up to 40% when assessed against ∼ 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.

Keywords: protein motion, protein language models, transfer learning, PCA, deep learning

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 (Lane, 2023; Miller and Phillips, 2021; Henzler-Wildman and Kern, 2007).

Many recent approaches have concentrated on repurposing the protein structure prediction neural network AlphaFold2 (Jumper et al., 2021) to generate conformational diversity (Sala et al., 2023). Guiding the predictions with state-annotated templates proved successful for modelling the multiple functional states of a couple of protein families (Faezov and Dunbrack Jr, 2023; Heo and Feig, 2022). In addition, massive sampling strategies have shown promising results for protein complexes (Wallner, 2023) (Wallner, 2023; Johansson-Åkhe and Wallner, 2022) with notable success in the blind CASP15-CAPRI assessment (Lensink et al., 2023). While they can be deployed seamlessly with parallelized implementations (Brysbaert et al., 2024), they remain highly resource-intensive.

Other strategies have explored promoting diversity by modulating and disentangling evolutionary signals (Sfriso et al., 2016). The rationale is that amino acid co-variations in evolution reflect 3D structural constraints (Benner and Gerloff, 1991; Göbel et al., 1994; Ortiz et al., 1999; Lapedes et al., 1999; Giraud et al., 1999; Thomas et al., 2005; Weigt et al., 2009). 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 (Bepler and Berger, 2021; Elnaggar et al., 2022; Lin et al., 2023). Inputting shallow, masked, corrupted or sub-sampled alignments to AlphaFold2 allowed for modelling distinct conformations for a few protein families (Kalakoti and Wallner, 2024; Wayment-Steele et al., 2023; Del Alamo et al., 2022; Stein and Mchaourab, 2022). Nevertheless, contradictory findings have highlighted difficulties in rationalising the effectiveness of these modifications and interpreting them, particularly for metamorphic proteins (Porter et al., 2024; Chakravarty and Porter, 2022; Chakravarty et al., 2023).

More classically, physics-based molecular dynamics (MD) is a method of choice to probe protein conformational landscapes (Hollingsworth and Dror, 2018). Nonetheless, the time scales amenable to MD simulations on standard hardware remain much smaller than those spanned by slow molecular processes (Chen et al., 2023). This limitation has stimulated the development of hybrid approaches combining MD with machine learning (ML) toward accelerating or enhancing sampling (Noé et al., 2020). Deep neural networks can help to identify collective variables from MD simulations as part of importance-sampling strategies (Chen et al., 2023; Belkacemi et al., 2021; Bonati et al., 2021; Wang et al., 2020; Ribeiro et al., 2018). Or they may directly generate conformations according to a probability distribution learnt from MD trajectories or sets of experimental structures (Zheng et al., 2024; Lu et al., 2023; Ramaswamy et al., 2021; Noé et al., 2019). Diffusion-based architectures (Abramson et al., 2024; Zheng et al., 2024; Jing et al., 2023) and the more general flow-matching framework (Jing et al., 2024) 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 (Abramson et al., 2024).

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 (Hayward and Go, 1995). 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 (Grudinin et al., 2020; Hoffmann and Grudinin, 2017). 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 (Laine and Grudinin, 2021), 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 (Berman et al., 2000). X-ray crystallography, cryogenic-electron microscopy (cryo-EM), and nuclear magnetic resonance spectroscopy (NMR) have provided invaluable insights into protein diverse conformational states (Ramelot et al., 2023; Miller and Phillips, 2021), but only for a relatively small number of proteins (Bryant, 2023). 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 (Trewhella, 2022; Martel and Gabel, 2022; Flechsig and Ando, 2023).

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 (Wankowicz and Fraser, 2023; Ramelot et al., 2023; Ellaway et al., 2023; Varadi et al., 2022; Modi and Dunbrack Jr, 2022; Parker et al., 2022; Tordai et al., 2022; Pándy-Szekeres et al., 2023). 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 (Lombard et al., 2024). DANCE application to the Protein Data Bank (PDB) (Berman et al., 2000) 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 (Lombard et al., 2024).

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. 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 (Lombard et al., 2024). 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 (Lin et al., 2023) 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 (Hayes et al., 2024; Heinzinger et al., 2023) 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.

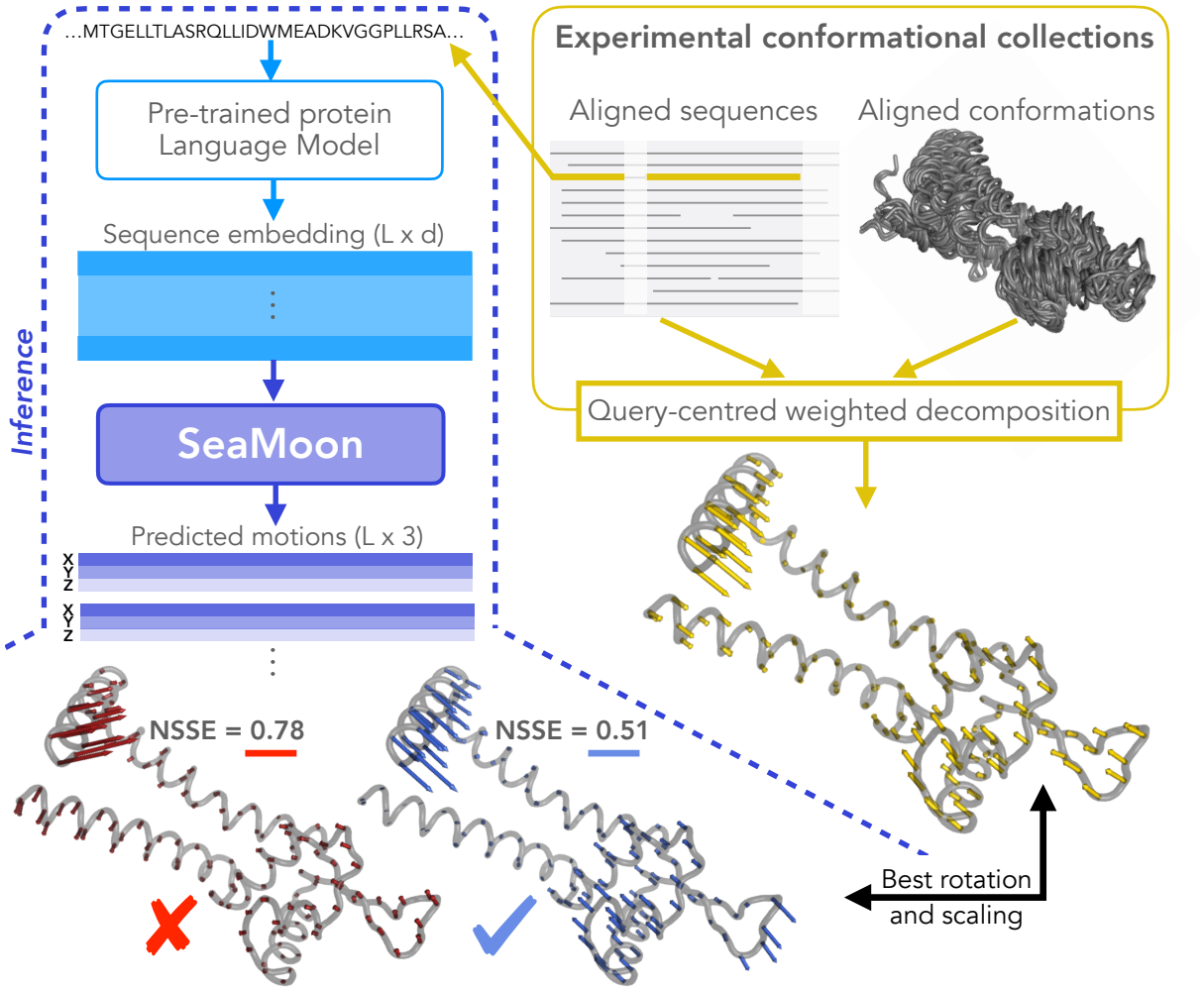

Figure 1: **Outline of SeaMoon's approach.** SeaMoon takes as input a high-dimensional $L \times d$ matrix representation of a protein sequence of length L computed by a pre-trained pLM. It outputs a set of 3D vectors of length L representing linear motions. The training procedure regresses these output 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..

Results and Discussion

The approach introduced in this work, SeaMoon, predicts continuous representations of protein motions with a convolutional neural network inputting pLM sequence embeddings (**Fig. 1**). We considered the purely sequence-based pLM ESM2 (Lin et al., 2023) and two structure-aware pLMs, namely ESM3 (Hayes et al., 2024) and ProstT5 (Heinzinger et al., 2023). ESM3 is the largest model (**Table S1**), and it can condition on and reconstruct several protein sequence and structural properties. ProstT5, the smallest model (**Table S1**), 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 (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. 1**). See **Fig. S1** for illustrative examples of different error levels. By comparison, random predictions typically display errors above 0.9 (**Fig. S2**). 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.

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 I** and **Fig. 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 I** and **Fig. 2A**). ProstT5, with the smallest number of parameters and embedding dimensions (**Table S1**), yielded the best overall performance (**Fig. 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 I** and **Fig. 2A**). More specifically, we augmented the training data by using multiple (up to 5) reference conformations per experimental collection (**Table S2**). While the pLM embeddings within a collection should be highly similar, the extracted motions may differ substantially from one reference to another (Lombard et al., 2024). The positive impact of this data augmentation strategy was most visible for the ESM-based version of SeaMoon (**Table I** and **Fig. 2A**).

SeaMoon effectively generalised to unseen proteins across diverse families (**Table I**, **Fig. 2B**, and **Fig. S4-5**). 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. S5A**), 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

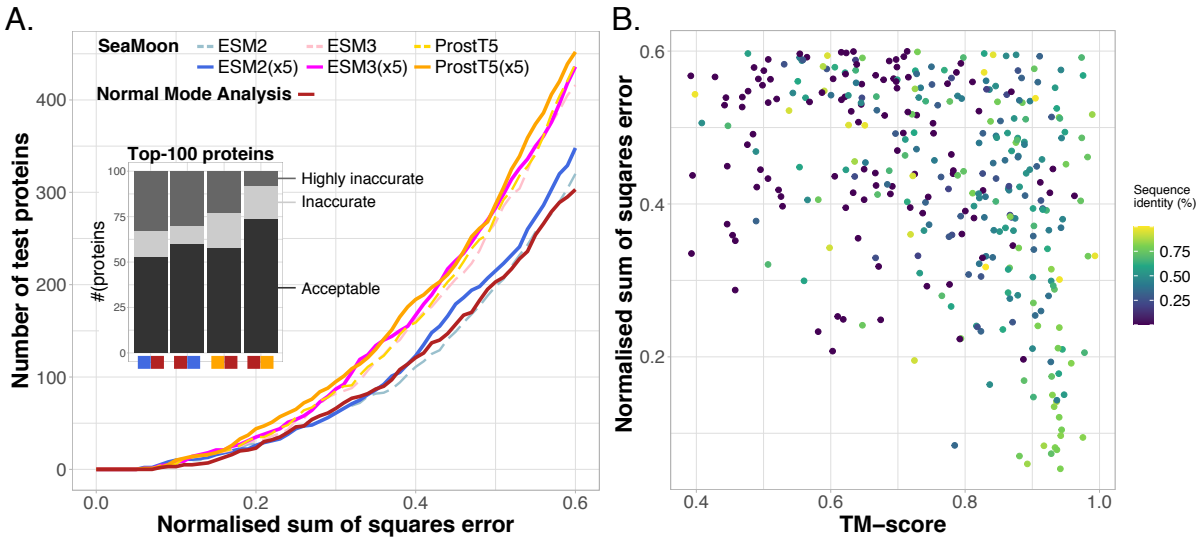

Figure 2: **SeaMoon performance and generalisation capability.** We report the NSSE of the best match between 3 predictions and 3 ground-truth motions for each of the 1 121 test proteins. **A.** Cumulative NSSE for six different versions of SeaMoon and for the NMA. We tested three pLMs, namely ESM2, 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. S3 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.

Table I: **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.

3D folds but highly divergent sequences. The ATP-binding cassette (ABC) transporter superfamily provides an illustrative example of this intermediate difficulty (**Fig. S5B**). SeaMoon-ESM2(x5) accurately predicted the opening-closing motion of a putative ABC transporter from *Campylobacter jejuni* (**Fig. S5B**, 5T1PE, $NSSE = 0.33$) that does not have any detectable sequence similarity with the training set. This motion is character-

istic of the "Venus Fly-trap" mechanism for transporting sugars (Chandravanshi et al., 2020) and is shared with a structurally similar ABC transporter from the training set (**Fig. S5B**, 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. S5C**, 4Z3ZF, $NSSE = 0.37$).

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. 2A**, inset, and **Fig. S6**). Among the top-100 proteins best-predicted by SeaMoon-ESM2(x5), about half exhibit motions accessible to the NMA (**Fig. 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. S7** 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. 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. 3**). SeaMoon-ESM2(x5) captured the PemK's loop L12 motion with high precision (**Fig. 3A**, $NSSE = 0.24$) whereas the NMA failed to delineate the mobile region in the protein and to infer its direction of movement (**Fig. 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 (Kim et al., 2022). 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. 3B**). By contrast, the NMA predicted a breathing motion poorly approximating the ground-truth one (**Fig. 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 (Machen et al., 2021). 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 (Anderson et al., 2020; Dhanasingh et al., 2021).

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. 2A**, inset, and **Fig. S6**). Using implicit structural knowledge allowed recovering elastic motions such as that exhibited by the mammalian plexin A4 ectodomain (**Fig. S8**, $NSSE = 0.28$). Taken together, SeaMoon-ProstT5(x5) and the NMA approximated the motions of 554 test proteins (out of 1121, 49%) with reasonable accuracy (**Table. I**). This result suggests that combining SeaMoon transfer learning approach with the physics- and geometry-based NMA could be a valuable strategy.

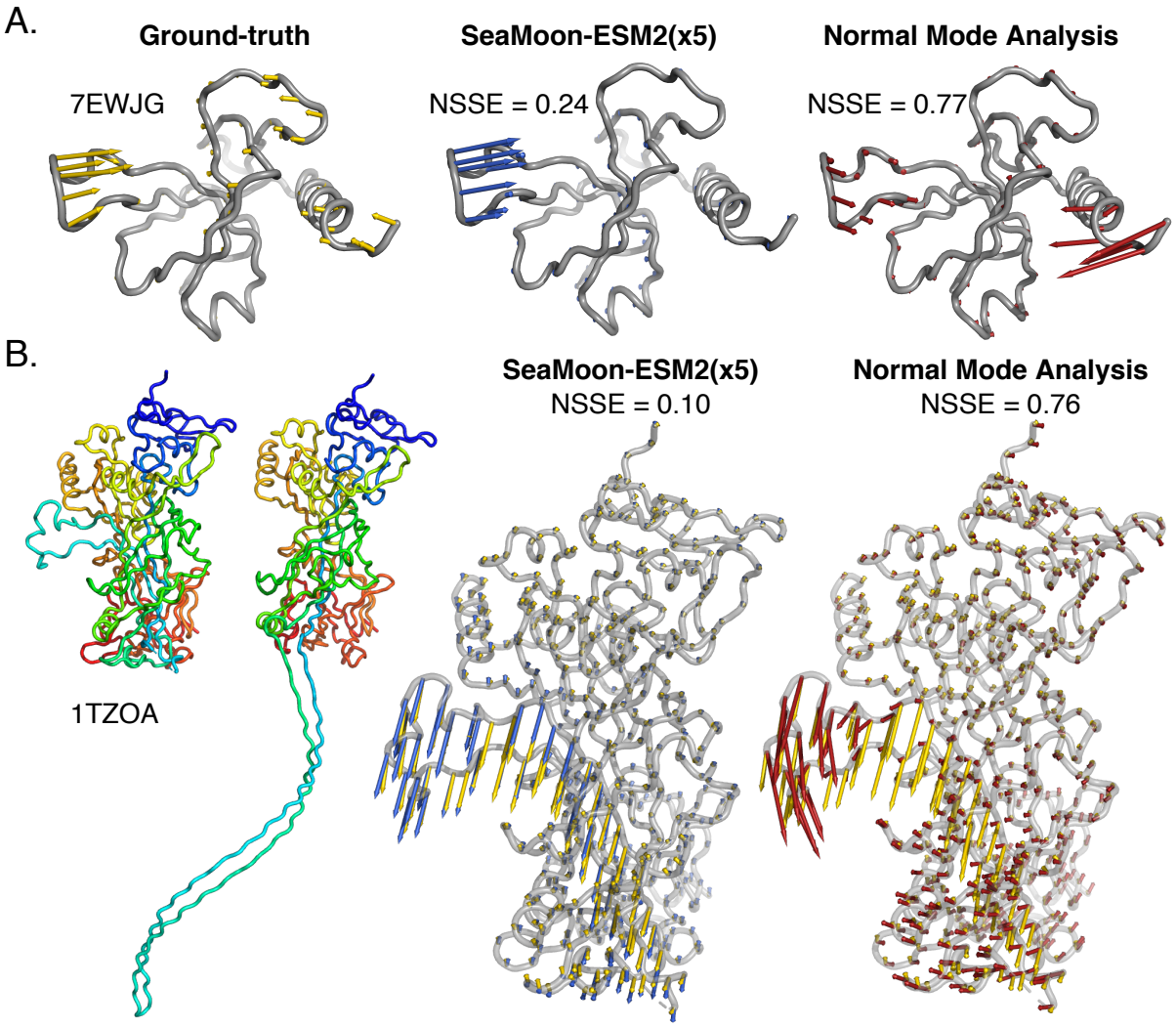

Figure 3: **Examples of motions well predicted by SeaMoon and not by the NMA.** The arrows depicted in yellow, blue and red on the grey 3D structures represent the ground-truth motions and the best-matching predictions from SeaMoon-ESM2(x5) and the NMA, respectively. **A.** Bacterial toxin PemK (PDB 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.

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 I**). We found that SeaMoon motion subspaces were fairly similar to the ground-truth ones, with a Root Mean Square Inner Product (RMSIP) (Amadei et al., 1999; Leo-Macias et al., 2005; David and Jacobs, 2011) 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** and **Fig. S9**). 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. 4A**). The associated NSSE were all smaller than 0.5 (**Fig. 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 (Podgorski et al., 2023).

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-ProstT5 baseline model (**Table S3** and **Fig. S10**). 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. S10**, 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 S3** and **Fig. S10**). Moreover, a 7-layer transformer architecture (see *Methods*) underperformed SeaMoon's convolutional neural network, despite having roughly the same number of free parameters (**Fig. S10**, 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. S10**, in light green). This result underlines the utility of implementing a permissive and flexible comparison of the predicted and ground-truth motions during training.

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. 4C-E**). Overall, the procedure allowed for correctly orienting

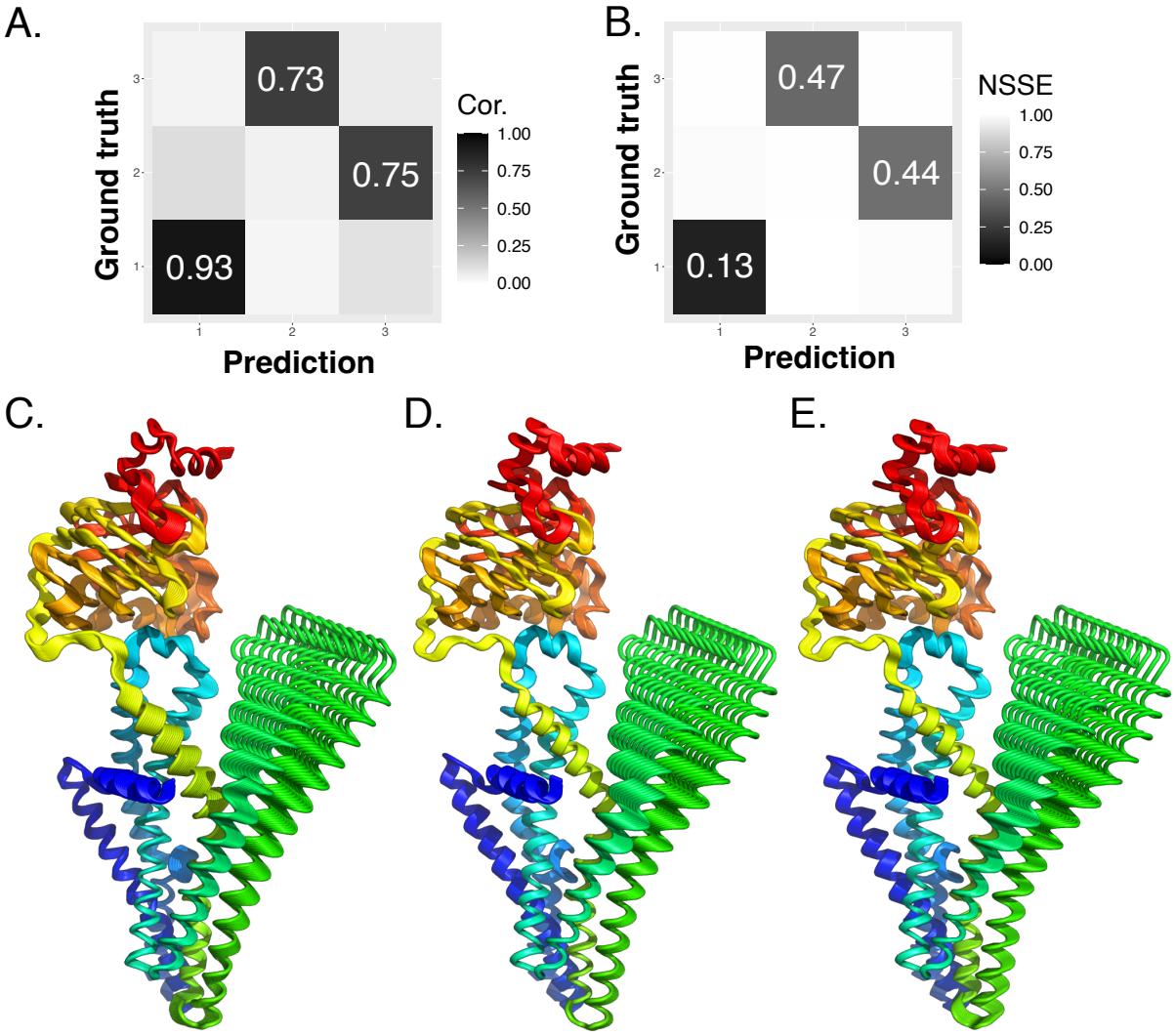

Figure 4: **Motion subspace comparison and deformation trajectories. A-B.** Ogopogo major capsid protein motion subspace. PDB code: 8ECN, chain B. **A.** Pairwise similarities measured as Pearson correlations between the ground-truth motions and SeaMoon-ProstT5(x5) predictions. **B.** Pairwise discrepancies measured as NSSE. **C-E.** Trajectories of a human ABC transporter (PDB code: 7D7R, chain A) deformed along its first ground-truth principal component (A) and the best-matching SeaMoon-ProstT5(x5) prediction (B-C). **B.** The prediction is optimally aligned with the ground truth. **C.** 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.

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.

Methods

Datasets

To generate training data, we constructed a non-redundant set of conformational collections representing the whole PDB (as of June 2023) using DANCE (Lombard et al., 2024). To ensure high quality of the data, we replaced the raw PDB coordinates with their updated and optimised versions from PDB-REDO whenever possible (Joosten et al., 2014). 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 (Lombard et al., 2024). 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 (Lombard et al., 2024). 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.

Model Specifications

Input features

SeaMoon takes as input embeddings computed from pre-trained pLMs, namely Evolutionary Scale Models ESM2-T33-650M-UR50 (Lin et al., 2023) and ESM3-small (1.4B) (Hayes et al., 2024), as well as Protein sequence-structure T5 (Heinzinger et al., 2023). ESM2-T33-650M-UR50 is a BERT (Devlin et al., 2018) style 650-million-parameter encoder-only transformer architecture trained on all clusters from Uniref50 (Suzek et al., 2015, 2007), a version of UniProt (Consortium, 2022) 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 (Vaswani et al., 2017) 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 (Richardson et al., 2023), OAS (Olsen et al., 2022) and the PDB (Berman et al., 2000), augmented with synthetic sequences generated by an inverse folding model (Hayes et al., 2024). Protein sequence-structure T5 is a bilingual pLM trained on a high-quality clustered version of the AlphaFold Protein Structure Database (Barrio-Hernandez et al., 2023; Varadi et al., 2021) to translate 1D sequences of amino acids into 1D sequences of 3Di tokens representing 3D structural states (Van Kempen et al., 2024) and vice versa. The 3Di alphabet, introduced by the 3D-alignment method Foldseek (Van Kempen et al., 2024), 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 (Elnaggar et al., 2022), an encoder-decoder transformer architecture (Raffel et al., 2020) trained on reconstructing corrupted amino acids from the Big Fantastic Database (Steinegger et al., 2019) 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 (LeCun et al., 2015) 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 S1**). 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 (Maas et al., 2013), as well as 80% dropout (Srivastava et al., 2014). 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 S3** and **Fig. S10** for more details. We implemented the models in PyTorch (Paszke et al., 2019) 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,

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

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 (Lombard et al., 2024). 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 (Gower and Dijksterhuis, 2004; Schönemann, 1966). 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,

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

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,

$$S_{kk} = \frac{\sum_{i=1}^{L} w_i (\vec{x}_{ik}^{\text{GT-trans}})^T \vec{x}_{ik}}{\sum_{i=1}^{L} w_i \| \vec{x}_{ik} \|^2}. \tag{3}$$

Training

We trained six models (**Table S2**) to predict $K = 3$ motions using the Adam optimizer (Kingma and Ba, 2014) 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,

$$\sum_{i=1}^{L} \vec{C}_i^{\,0} \times \vec{x}_{ik}^{\,\mathrm{GT}} = \vec{0}. \tag{4}$$

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

$$\sum_{i=1}^{L} \vec{C}_i \times R\vec{x}_{ik} = \vec{0}. \tag{5}$$

This problem has at most four solutions and we solve it exactly using the symbolic *wolframclient* package 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.

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},

$$SSE = \frac{1}{L}\sum_{i=1}^{L} w_i \|\vec{x}_{il}^{\,\text{GT-trans}} - s^{opt}\vec{x}_{ik}\|^2, \tag{6}$$

$$\text{with } \vec{x}_{il}^{\,\text{GT-trans}} = R^{opt}\vec{x}_{il}^{\,\mathrm{GT}} \tag{7}$$

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 3, the scaling coefficient s^{opt} will be equal to c, and thus, the error will be null,

$$SSE_{min} = \frac{1}{L}\sum_{i=1}^{L} w_i \|\vec{x}_{il}^{\,\text{GT-trans}} - c\vec{x}_{ik}\|^2 = \frac{1}{L}\sum_{i=1}^{L} w_i \|c\vec{x}_{ik} - c\vec{x}_{ik}\|^2 = 0. \tag{8}$$

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,

$$SSE_{max} = \frac{1}{L}\sum_{i=1}^{L} w_i \|\vec{x}_{il}^{\,\text{GT-trans}} - \vec{0}\|^2 = \frac{1}{L}\sum_{i=1}^{L} w_i \|\vec{x}_{il}^{\,\text{GT-trans}}\|^2. \tag{9}$$

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,

$$NSSE = \frac{SSE}{SSE_{max}}. \tag{10}$$

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) (Amadei et al., 1999; Leo-Macias et al., 2005; David and Jacobs, 2011). It is computed as an average of the normalised inner products of all the vectors in both subspaces,

$$\text{RMSIP} = \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{i=1}^{L} \frac{(\vec{x}_{ik}^{\text{GT}})^T \vec{x}_{il}^{\text{ortho}}}{\|\vec{x}_{ik}^{\text{GT}}\| \|\vec{x}_{il}^{\text{ortho}}\|} \right), \tag{11}$$

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. S11**). A RMSIP score of 0.70 is considered an excellent correspondence while a score of 0.50 is considered fair (Amadei et al., 1999).

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 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.

Comparison with the normal mode analysis

We compared SeaMoon performance with the physics-based unsupervised normal mode analysis (NMA) (Hayward and Go, 1995). 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 (Hoffmann and Grudinin, 2017) 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 (Laine and Grudinin, 2021). 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.

Motion properties

Contribution

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, $\frac{\lambda_k}{\sum_l \lambda_l}$.

Collectivity

We estimate the collectivity (Brüschweiler, 1995; Tama and Sanejouand, 2001) of the $L \times 3$ predicted or ground-truth motion tensor X_k as,

$$\kappa(X_k) = \frac{1}{L} \exp\left(-\sum_{i=1}^{L} \sum_{j=1}^{3} X_{ijk}^2 \log X_{ijk}^2 \right), \tag{12}$$

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 $1/L$.

Conclusion

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.

Declaration

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 of SeaMoon and SeaMoon predictions are freely available at Zenodo.

Acknowledgments

The Sorbonne Center for Artificial Intelligence (SCAI) provided a salary to VL and computational resources. This work has also been co-funded by the European Union (ERC, PROMISE, 101087830). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

S.G. and E.L. designed research and supervised the project. V.L. designed the model's architecture and carried out its implementation. S.G. and D.T. wrote the proofs and problem formalisation for orienting predictions with respect to a protein conformation

with feedback from E.L.. D.T. implemented the solver. V.L., E.L. and S.G. produced and analysed the results. E.L. wrote the manuscript with input, support and feedback from all authors. All authors edited, read, and approved the final manuscript.

Competing interests

The author(s) declare no competing interests.

References

Lane, T. J. Protein structure prediction has reached the single-structure frontier. *Nature Methods* **2023**, *20*, 170–173.

Miller, M. D.; Phillips, G. N. Moving beyond static snapshots: Protein dynamics and the Protein Data Bank. *Journal of Biological Chemistry* **2021**, *296*.

Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature* **2007**, *450*, 964–972.

Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

Sala, D.; Engelberger, F.; Mchaourab, H.; Meiler, J. Modeling conformational states of proteins with AlphaFold. *Current Opinion in Structural Biology* **2023**, *81*, 102645.

Faezov, B.; Dunbrack Jr, R. L. AlphaFold2 models of the active form of all 437 catalytically-competent typical human kinase domains. *bioRxiv* **2023**, 2023–07.

Heo, L.; Feig, M. Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins: Structure, Function, and Bioinformatics* **2022**, *90*, 1873–1885.

Wallner, B. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **2023**, *39*, btad573.

Wallner, B. Improved multimer prediction using massive sampling with AlphaFold in CASP15. *Proteins: Structure, Function, and Bioinformatics* **2023**, *91*, 1734–1746.

Johansson-Åkhe, I.; Wallner, B. Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Frontiers in Bioinformatics* **2022**, *2*, 85.

others,, et al. Impact of AlphaFold on structure prediction of protein complexes: The CASP15-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics* **2023**, *91*, 1658–1683.

Brysbaert, G.; Raouraoua, N.; Mirabello, C.; Véry, T.; Blanchet, C.; Wallner, B.; Lensink, M. MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. **2024**,

Sfriso, P.; Duran-Frigola, M.; Mosca, R.; Emperador, A.; Aloy, P.; Orozco, M. Residues coevolution guides the systematic identification of alternative functional conformations in proteins. *Structure* **2016**, *24*, 116–126.

573 Benner, S. A.; Gerloff, D. Patterns of divergence in homologous proteins as indicators of
574 secondary and tertiary structure: a prediction of the structure of the catalytic domain
575 of protein kinases. *Advances in Enzyme Regulation* **1991**, *31*, 121–181.

576 Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue
577 contacts in proteins. *Proteins: Structure, Function, and Bioinformatics* **1994**, *18*, 309–
578 317.

579 Ortiz, A. R.; Kolinski, A.; Rotkiewicz, P.; Ilkowski, B.; Skolnick, J. Ab initio folding of
580 proteins using restraints derived from evolutionary information. *Proteins: Structure,*
581 *Function, and Bioinformatics* **1999**, *37*, 177–185.

582 Lapedes, A. S.; Giraud, B. G.; Liu, L.; Stormo, G. D. Correlated mutations in models of
583 protein sequences: phylogenetic and structural effects. *Lecture Notes-Monograph Series*
584 **1999**, 236–256.

585 Giraud, B.; Heumann, J. M.; Lapedes, A. S. Superadditive correlation. *Physical Review*
586 *E* **1999**, *59*, 4983.

587 Thomas, J.; Ramakrishnan, N.; Bailey-Kellogg, C. Graphical models of residue coupling
588 in protein families. Proceedings of the 5th international workshop on Bioinformatics.
589 2005; pp 12–20.

590 Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct
591 residue contacts in protein–protein interaction by message passing. *Proceedings of the*
592 *National Academy of Sciences* **2009**, *106*, 67–72.

593 Bepler, T.; Berger, B. Learning the protein language: Evolution, structure, and function.
594 *Cell systems* **2021**, *12*, 654–669.

595 Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.;
596 Shmueli, Y.; Dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A.
597 Evolutionary-scale prediction of atomic-level protein structure with a language model.
598 *Science* **2023**, *379*, 1123–1130.

599 Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.;
600 Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Toward
601 Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans-*
602 *actions on Pattern Analysis and Machine Intelligence* **2022**, *44*, 7112–7127.

603 Kalakoti, Y.; Wallner, B. AFsample2: Predicting multiple conformations and ensembles
604 with AlphaFold2. *bioRxiv* **2024**, 2024–05.

605 Wayment-Steele, H. K.; Ojoawo, A.; Otten, R.; Apitz, J. M.; Pitsawong, W.;
606 Hömberger, M.; Ovchinnikov, S.; Colwell, L.; Kern, D. Predicting multiple confor-
607 mations via sequence clustering and AlphaFold2. *Nature* **2023**, 1–3.

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

610 Stein, R. A.; Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and con-
611 formational heterogeneity with Alphafold2. *PLOS Computational Biology* **2022**, *18*,
612 e1010483.

Porter, L. L.; Artsimovitch, I.; Ramírez-Sarmiento, C. A. Metamorphic proteins and how to find them. *Current opinion in structural biology* **2024**, *86*, 102807.

Chakravarty, D.; Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Science* **2022**, *31*, e4353.

Chakravarty, D.; Schafer, J. W.; Chen, E. A.; Thole, J.; Porter, L. AlphaFold2 has more to learn about protein energy landscapes. *bioRxiv* **2023**, 2023–12.

Hollingsworth, S. A.; Dror, R. O. Molecular dynamics simulation for all. *Neuron* **2018**, *99*, 1129–1143.

Chen, H.; Roux, B.; Chipot, C. Discovering reaction pathways, slow variables, and committor probabilities with machine learning. *Journal of Chemical Theory and Computation* **2023**, *19*, 4414–4426.

Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annual review of physical chemistry* **2020**, *71*, 361–390.

Belkacemi, Z.; Gkeka, P.; Lelièvre, T.; Stoltz, G. Chasing collective variables using autoencoders and biased trajectories. *Journal of chemical theory and computation* **2021**, *18*, 59–78.

Bonati, L.; Piccini, G.; Parrinello, M. Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2113533118.

Wang, Y.; Ribeiro, J. M. L.; Tiwary, P. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current opinion in structural biology* **2020**, *61*, 139–145.

Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *The Journal of chemical physics* **2018**, *149*.

others,, et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence* **2024**, 1–10.

Lu, J.; Zhong, B.; Tang, J. Score-based enhanced sampling for protein molecular dynamics. ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling. 2023.

Ramaswamy, V. K.; Musson, S. C.; Willcocks, C. G.; Degiacomi, M. T. Deep learning protein conformational space with convolutions and latent interpolations. *Physical Review X* **2021**, *11*, 011052.

Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*, eaaw1147.

others,, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, 1–3.

Jing, B.; Erives, E.; Pao-Huang, P.; Corso, G.; Berger, B.; Jaakkola, T. EigenFold: Generative Protein Structure Prediction with Diffusion Models. *arXiv preprint arXiv:2304.02198* **2023**,

651 Jing, B.; Berger, B.; Jaakkola, T. AlphaFold meets flow matching for generating protein
652 ensembles. *arXiv preprint arXiv:2402.04845* **2024**,

653 Hayward, S.; Go, N. Collective variable description of native protein dynamics. *Annual*
654 *review of physical chemistry* **1995**, *46*, 223–250.

655 Grudinin, S.; Laine, E.; Hoffmann, A. Predicting protein functional motions: an old recipe
656 with a new twist. *Biophysical journal* **2020**, *118*, 2513–2525.

657 Hoffmann, A.; Grudinin, S. NOLB: Nonlinear rigid block normal-mode analysis method.
658 *Journal of chemical theory and computation* **2017**, *13*, 2123–2134.

659 Laine, E.; Grudinin, S. HOPMA: Boosting protein functional dynamics with colored
660 contact maps. *The Journal of Physical Chemistry B* **2021**, *125*, 2577–2588.

661 Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.;
662 Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic acids research* **2000**,
663 *28*, 235–242.

664 Ramelot, T. A.; Tejero, R.; Montelione, G. T. Representing structures of the multiple
665 conformational states of proteins. *Current Opinion in Structural Biology* **2023**, *83*,
666 102703.

667 Bryant, P. Structure prediction of alternative protein conformations. *bioRxiv* **2023**, 2023–
668 09.

669 Trewhella, J. Recent advances in small-angle scattering and its expanding impact in
670 structural biology. *Structure* **2022**, *30*, 15–23.

671 Martel, A.; Gabel, F. Time-resolved small-angle neutron scattering (TR-SANS) for struc-
672 tural biology of dynamic systems: Principles, recent developments, and practical guide-
673 lines. *Methods in enzymology* **2022**, *677*, 263–290.

674 Flechsig, H.; Ando, T. Protein dynamics by the combination of high-speed AFM and
675 computational modeling. *Current Opinion in Structural Biology* **2023**, *80*, 102591.

676 Wankowicz, S.; Fraser, J. Comprehensive Encoding of Conformational and Compositional
677 Protein Structural Ensembles through mmCIF Data Structure. *ChemRxiv* **2023**,

678 Ellaway, J. I.; Anyango, S.; Nair, S.; Zaki, H. A.; Nadzirin, N.; Powell, H. R.; Gut-
679 manas, A.; Varadi, M.; Velankar, S. Identifying Protein Conformational States in the
680 PDB and Comparison to AlphaFold2 Predictions. *bioRxiv* **2023**, 2023–07.

681 Varadi, M. et al. PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated
682 resources of macromolecular structures to support basic and applied research and ed-
683 ucation. *Protein Science* **2022**, *31*.

684 Modi, V.; Dunbrack Jr, R. L. Kincore: a web resource for structural classification of
685 protein kinases and their inhibitors. *Nucleic Acids Research* **2022**, *50*, D654–D664.

686 Parker, M. I.; Meyer, J. E.; Golemis, E. A.; Dunbrack Jr, R. L. Delineating the RAS
687 conformational landscape. *Cancer research* **2022**, *82*, 2485–2498.

Tordai, H.; Suhajda, E.; Sillitoe, I.; Nair, S.; Varadi, M.; Hegedus, T. Comprehensive collection and prediction of ABC transmembrane protein structures in the AI era of structural biology. *International Journal of Molecular Sciences* **2022**, *23*, 8877.

Pándy-Szekeres, G.; Caroli, J.; Mamyrbekov, A.; Kermani, A. A.; Keserű, G. M.; Kooistra, A. J.; Gloriam, D. E. GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Research* **2023**, *51*, D395–D402.

Lombard, V.; Grudinin, S.; Laine, E. Explaining Conformational Diversity in Protein Families through Molecular Motions. *Scientific Data* **2024**, *11*, 752.

others,, et al. Simulating 500 million years of evolution with a language model. *bioRxiv* **2024**, 2024–07.

Heinzinger, M.; Weissenow, K.; Sanchez, J. G.; Henkel, A.; Steinegger, M.; Rost, B. ProstT5: Bilingual language model for protein sequence and structure. *bioRxiv* **2023**, 2023–07.

Kim, D.-H.; Kang, S.-M.; Baek, S.-M.; Yoon, H.-J.; Jang, D. M.; Kim, H. S.; Lee, S. J.; Lee, B.-J. Role of PemI in the Staphylococcus aureus PemIK toxin–antitoxin complex: PemI controls PemK by acting as a PemK loop mimic. *Nucleic Acids Research* **2022**, *50*, 2319–2333.

Machen, A. J.; Fisher, M. T.; Freudenthal, B. D. Anthrax toxin translocation complex reveals insight into the lethal factor unfolding and refolding mechanism. *Scientific Reports* **2021**, *11*, 13038.

Anderson, D. M.; Sheedlo, M. J.; Jensen, J. L.; Lacy, D. B. Structural insights into the transition of Clostridioides difficile binary toxin from prepore to pore. *Nature microbiology* **2020**, *5*, 102–107.

Dhanasingh, I.; Choi, E.; Lee, J.; Lee, S. H.; Hwang, J. Functional and structural characterization of Deinococcus radiodurans R1 MazEF toxin-antitoxin system, Dr0416-Dr0417. *Journal of Microbiology* **2021**, *59*, 186–201.

Chandravanshi, M.; Samanta, R.; Kanaujia, S. P. Conformational trapping of a β-glucosides-binding protein unveils the selective two-step ligand-binding mechanism of ABC importers. *Journal of molecular biology* **2020**, *432*, 5711–5734.

Hou, X.; Burstein, S. R.; Long, S. B. Structures reveal opening of the store-operated calcium channel Orai. *Elife* **2018**, *7*, e36758.

Amadei, A.; Ceruso, M. A.; Di Nola, A. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics* **1999**, *36*, 419–424.

Leo-Macias, A.; Lopez-Romero, P.; Lupyan, D.; Zerbino, D.; Ortiz, A. R. An analysis of core deformations in protein superfamilies. *Biophysical journal* **2005**, *88*, 1291–1299.

David, C. C.; Jacobs, D. J. Characterizing protein motions from structure. *Journal of Molecular Graphics and Modelling* **2011**, *31*, 41–56.

others,, et al. A structural dendrogram of the actinobacteriophage major capsid proteins provides important structural insights into the evolution of capsid stability. *Structure* **2023**, *31*, 282–294.

Joosten, R. P.; Long, F.; Murshudov, G. N.; Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **2014**, *1*, 213–220.

Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**,

Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; Consortium, U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015**, *31*, 926–932.

Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **2007**, *23*, 1282–1288.

Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **2022**, *51*, D523–D531.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.

others,, et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research* **2023**, *51*, D753–D759.

Olsen, T. H.; Boyles, F.; Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science* **2022**, *31*, 141–146.

Barrio-Hernandez, I.; Yeo, J.; Jänes, J.; Mirdita, M.; Gilchrist, C. L.; Wein, T.; Varadi, M.; Velankar, S.; Beltrao, P.; Steinegger, M. Clustering predicted structures at the scale of the known protein universe. *Nature* **2023**, *622*, 637–645.

Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **2021**, *50*, D439–D444.

Van Kempen, M.; Kim, S. S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C. L.; Söding, J.; Steinegger, M. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **2024**, *42*, 243–246.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **2020**, *21*, 1–67.

Steinegger, M.; Mirdita, M.; Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods* **2019**, *16*, 603–606.

LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.

others,, et al. Rectifier nonlinearities improve neural network acoustic models. Proc. icml. 2013; p 3.

766 Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a
767 simple way to prevent neural networks from overfitting. *The journal of machine learning*
768 *research* **2014**, *15*, 1929–1958.

769 others„ et al. Pytorch: An imperative style, high-performance deep learning library. *Ad-*
770 *vances in neural information processing systems* **2019**, *32*.

771 Gower, J. C.; Dijksterhuis, G. B. *Procrustes problems*; OUP Oxford, 2004; Vol. 30.

772 Schönemann, P. H. A generalized solution of the orthogonal procrustes problem. *Psy-*
773 *chometrika* **1966**, *31*, 1–10.

774 Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint*
775 *arXiv:1412.6980* **2014**,

776 Brüschweiler, R. Collective protein dynamics and nuclear spin relaxation. *The Journal of*
777 *Chemical Physics* **1995**, *102*, 3396–3403.

778 Tama, F.; Sanejouand, Y. H. Conformational change of proteins arising from normal
779 mode calculations. *Protein Engineering* **2001**, *14*, 1–6.