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, Department of Computational, Quantitative and Synthetic Biology (CQSB, UMR7238), 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

Summary

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 sequences, without exploiting 3D structures. SeaMoon leverages protein Language Model (pLM) embeddings as input to a lightweight convolutional neural network. We assessed SeaMoon against $\sim$ 1 000 collections of experimental conformations exhibiting diverse motions. It predicts at least one ground-truth motion with reasonable accuracy for 40% of the test proteins. SeaMoon captures motions inaccessible to normal mode analysis, an unsupervised physics-based method relying solely on 3D geometry, and generalises to proteins without detectable sequence similarity to the training set. SeaMoon is easily retrainable with novel or updated pLMs.

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[1–3].

Many recent approaches have concentrated on repurposing the protein structure prediction neural network AlphaFold2[4] to generate conformational diversity[5]. Guiding the predictions with state-annotated templates proved successful for modelling the multiple functional states of a couple of protein families[6,7]. In addition, massive sampling strategies have shown promising results for protein complexes[8 9,10] with notable success in the blind CASP15-CAPRI assessment[11]. While they can be deployed seamlessly with parallelized implementations[12], they remain highly resource-intensive.

Other strategies have explored promoting diversity by modulating and disentangling evolutionary signals[13]. The rationale is that amino acid co-variations in evolution reflect 3D structural constraints[14–20]. 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[21–23]. Inputting shallow, masked, corrupted or sub-sampled alignments to AlphaFold2 allowed for modelling distinct conformations for a few protein families[24–27]. Nevertheless, contradictory findings have highlighted difficulties in rationalising the effectiveness of these modifications and interpreting them, particularly for metamorphic proteins[28–30].

More classically, physics-based molecular dynamics (MD) is a method of choice to probe protein conformational landscapes[31]. Nonetheless, the time scales amenable to MD simulations on standard hardware remain much smaller than those spanned by slow molecular processes[32]. This limitation has stimulated the development of hybrid approaches combining MD with machine learning (ML) toward accelerating or enhancing sampling[33]. Deep neural networks can help to identify collective variables from MD simulations as part of importance-sampling strategies[32,34–37]. Or they may directly generate conformations according to a probability distribution learnt from MD trajectories or sets of experimental structures[38–41]. Diffusion-based architectures[38,42,43] and the more general flow-matching framework[44] 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[42].

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[45]. 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[46,47]. The generated conformations are physically plausible, stereochemically realistic, and some of them approximate known biologically relevant intermediate states[46]. 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[48], 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[49]. X-ray crystallography, cryogenic-electron microscopy (cryo-EM), and nuclear magnetic resonance spectroscopy (NMR) have provided invaluable insights into protein diverse conformational states[2,50], but only for a relatively small number of proteins[51]. 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[52–54].

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[50,55–61]. 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[62]. DANCE application to the Protein Data Bank (PDB)[49] revealed that the conformations observed for most protein families lie on a low-dimensional *manifold*. Interpolation trajectories along the manifold can recapitulate known intermediate conformations, supporting its biological significance. Moreover, classical dimensionality reduction techniques can learn this manifold and generate unseen conformations with reasonable accuracy, albeit only in close vicinity of the training set[62].

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[62]. To this end, we devised an objective function invariant to global translations, rotations, and dilatations in 3D space. SeaMoon achieved a predictive performance similar to the normal mode analysis (NMA) when inputting purely sequence-based pLM embeddings[23] 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[63,64] 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.

Results

SeaMoon predicts protein motions from amino acid sequences alone

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 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

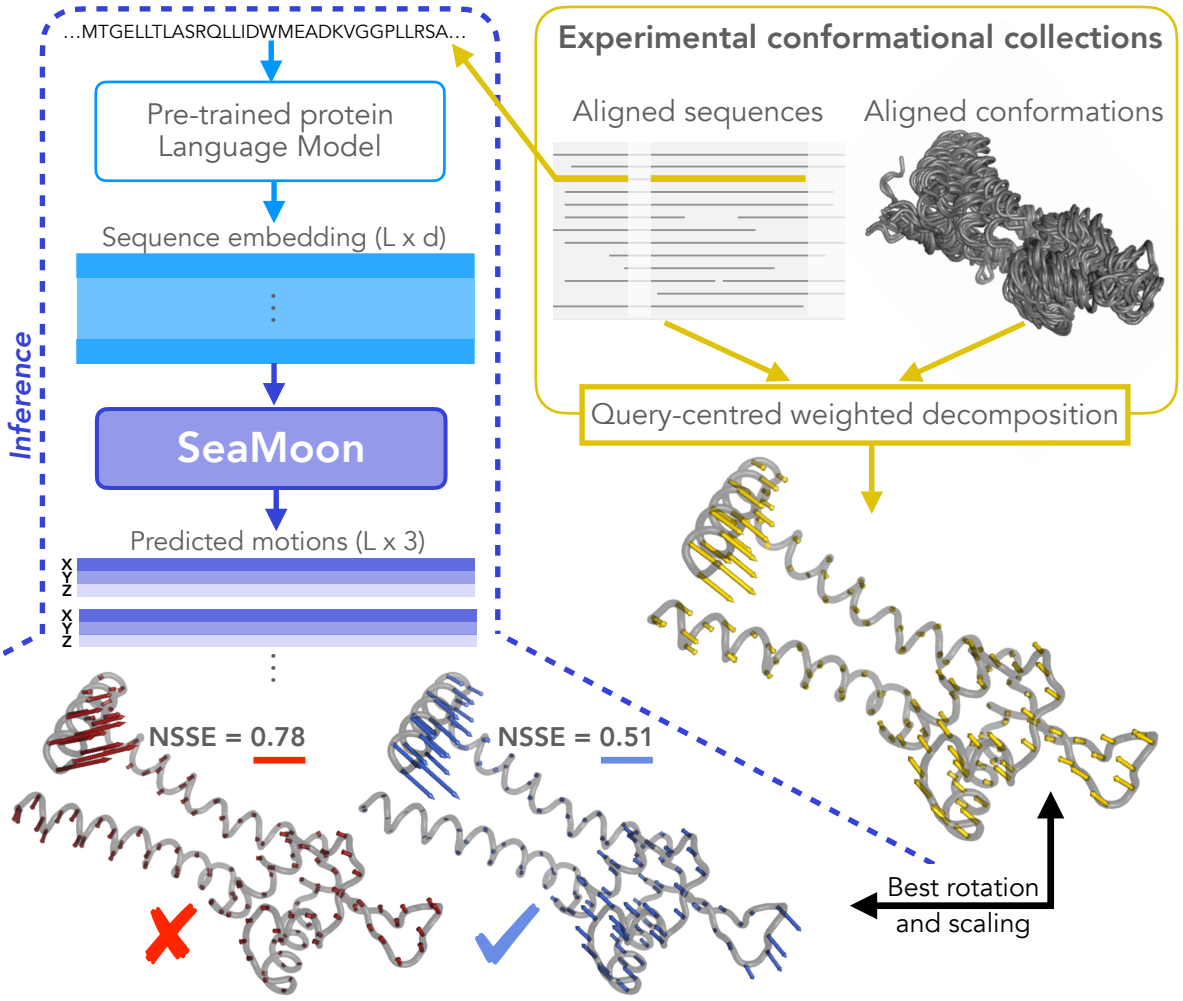

Figure 1: **Outline of SeaMoon's approach.** SeaMoon takes as input a high-dimensional $L \times d$ matrix representation of a protein sequence of length L computed by a pre-trained pLM. It outputs a set of 3D vectors of length L representing linear motions. The training procedure regresses these output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

used the principal components extracted from these collections as ground-truth linear motions to which we compared SeaMoon predicted 3D vectors three per protein by default). 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 a normalised sum-of-squares error (NSSE) between them (see *Methods* for details).

SeaMoon predicted protein motions with similar or better accuracy compared to the purely geometry-based unsupervised NMA when assessed against a test set of 1 121

proteins (**Fig. 2A** and **S1**). SeaMoon performance depends on the pLM used to compute the input embeddings from amino acid sequences. More specifically, the two structure-aware pLMs ESM3[63] and ProstT5[64] yielded a substantial improvement over the purely sequence-based pLM ESM2[23]. Furthermore, ProstT5 outperformed ESM3 (**Fig. 2A**, paired Wilcoxon signed-rank test p-values $< 10^{-6}$), despite having a much smaller number of parameters and embedding dimensions (**Table S1**). ProstT5 is a fine-tuned version of the sequence-only model ProtT5 that translates amino acid sequences into sequences of discrete structural states and reciprocally, while ESM3 is a multi-modal pLM capable of conditioning on and reconstructing several protein sequence and structural properties. Beside the influence of the pLM, we observed a boost in performance by up to 10% upon stimulating the model to learn a one-sequence-to-many-motions mapping (**Fig. 2A**, compare plain and dotted lines). 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[62]. The positive impact of this data augmentation strategy was most visible for the ESM2-based version of SeaMoon (**Fig. 2A**).

Looking at individual predictions, one can appreciate how they degrade as the error (NSSE) increases (**Fig. 3**). A prediction with a NSSE smaller than 0.2 almost perfectly superimposes to the ground-truth motion (**Fig. 3**, 8E7MH). SeaMoon-ProstT5(x5) generated at least one such near-perfect prediction, among the three predicted, for 44 proteins, representing 4% of the test set (**Fig. 2A**). It achieved NSSE smaller than 0.4 for 184 proteins (16%) and smaller than 0.6 for 452 proteins (40%, **Table I**). We consider predictions with NSSE larger than 0.6 as inaccurate as they typically miss or indicate completely wrong directions for a large part of the residues involved in the motion (**Fig. 3**, see 4ZEVB, 6W19p, and 7RTNB). By comparison, the errors computed for random predictions are typically above 0.9 (**Fig. S1B**). Based on a quality NSSE cutoff of 0.6, SeaMoon reached the highest success rate of 40% with ProstT5, followed by ESM3 (39%) and ESM2 (31%), while the NMA success rate is 27% (**Table I**). Average success rates computed over 1 000 bootstrap subsampling simulations, each one containing ∼100 test proteins, deviate by less than 0.5% from the estimates computed over the full test set (**Fig. S2**). The corresponding distributions follow the same trend as the cumulative error curves (**Fig. 2A** and **S1**). In particular, the interquartile range at a NSSE cutoff of 0.6 is 37-44% for SeaMoon-ProstT5(x5), significantly higher than 24-30% for the NMA (**Fig. S2**).

SeaMoon generalises to unseen proteins across diverse protein families

SeaMoon produced high-quality predictions at different levels of similarity to the training set, which we can interpret as varying difficulty levels (**Table I**, **Fig. 2B**, and **Fig. S3**). It generated acceptable predictions for up to 50% of the easy cases, namely the test proteins sharing over 70% sequence similarity with at least one train protein (**Table I**). The predictions are almost perfect (NSSE < 0.2) for about 20 proteins, representing 10% of this subset (**Fig. 2C**). Most of them are antibodies, a class of proteins well represented in both train and test sets (**Fig. 4A**). Beyond such easy cases, SeaMoon achieved similar success rates on test proteins sharing similar 3D folds with some train proteins (TM-score>0.7), despite highly divergent sequences, below 30% identity (**Table**

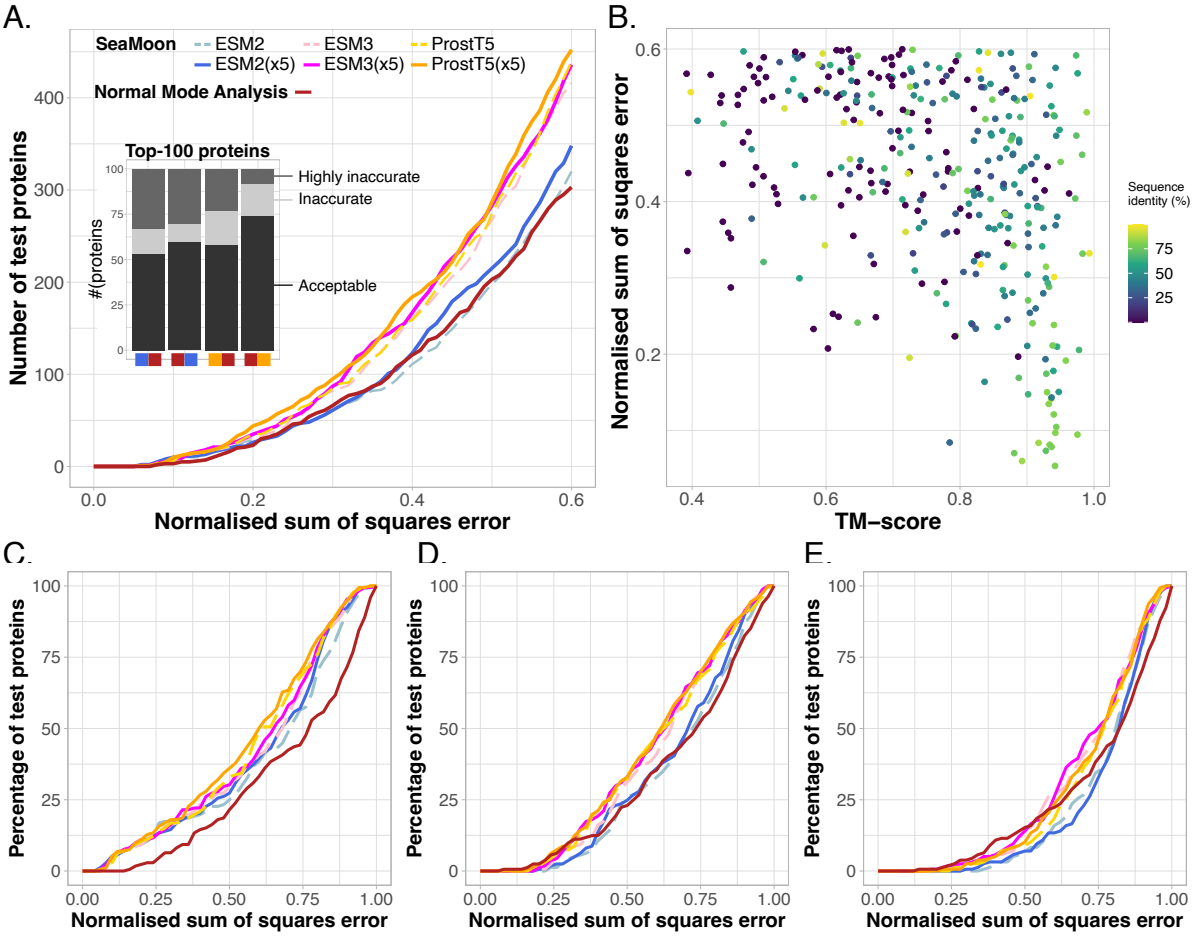

Figure 2: **SeaMoon performance and generalisation capability.** We report the NSSE of the best match between 3 predictions and 3 ground-truth motions for each test protein. **A.** Cumulative NSSE for six different versions of SeaMoon and for the NMA. We tested three pLMs, namely ESM2, ESM3 and ProstT5, 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. S1 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; see Fig. S5 for the other SeaMoon versions and NMA. **C-E.** Cumulative NSSE computed on three subsets of test proteins with increasing difficulty. **A. Easy:** at least 70% sequence identity with at least one train protein. **B. Intermediate:** at most 30% sequence identity with any train set example and a TM-score of at least 0.7 with some train protein. **C. Difficult:** at most 30% sequence identity and 0.5 as TM-score with any train protein. The colour code is the same as in panel A.

I and **Fig. 2D**). These test proteins represent an intermediate level of difficulty for which the ATP-binding cassette (ABC) transporter superfamily provides an illustrative example. In particular, SeaMoon-ESM2(x5) successfully transferred the opening-closing

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

Method	**Protocol**	**Number of proteins with acceptable predictions (NSSE<0.6)**			
		overall	easy	intermediate	difficult
Total		1 121	172	161	158
SeaMoon	ESM2	320 (29%)	66 (38%)	53 (33%)	25 (16%)
	ESM2(x5)	348 (31%)	68 (40%)	58 (36%)	21 (13%)
	ESM3	416 (37%)	70 (41%)	66 (41%)	**43 (27%)**
	ESM3(x5)	436 (39%)	73 (42%)	75 (47%)	41 (26%)
	ProstT5	439 (39%)	84 (49%)	76 (47%)	29 (18%)
	ProstT5(x5)	**452 (40%)**	**86 (50%)**	**77 (48%)**	37 (23%)
NMA		303 (27%)	57 (33%)	57 (35%)	35 (22%)

We report the number of proteins with at least one acceptable prediction over the whole test set (1 121 proteins in total) or over subsets of different levels of difficulty. **A. Easy:** at least 70% sequence identity with at least one train protein. **B. Intermediate:** at most 30% sequence identity with any train set example and a TM-score of at least 0.7 with some train protein. **C. Difficult:** at most 30% sequence identity and 0.5 as TM-score with any train protein. We consider predictions as acceptable if their normalised sum-of-squares error is smaller than 0.6 (see **Fig. 3** for illustrative examples of predictions with different error levels). The highest success rate is highlighted in bold for each category.

motion characteristic of the "Venus Fly-trap" mechanism for transporting sugars[65] from ABC transporters in the training set to the held-out putative ABC transporter from *Campylobacter jejuni* (**Fig. 4B**, 5T1PE, NSSE=0.33). The latter does not have any detectable sequence similarity with any train protein but shares high structural similarity with the sugar ABC transporter from *Thermus thermophilus* (7C68B, TM-score = 0.83). Finally, while SeaMoon predictions tend to display higher errors on the most challenging subset, defined with TM-score below 0.5 and sequence similarity below 30%, its success rate at NSSE=0.6 remains comparable or better than that of the NMA (**Table I** and **Fig. 2E**). Successful cases completely unrelated to the training set include the benzoyl-coenzyme A reductase from *Geobacter metallireducens* (**Fig. 4C**, 4Z3ZF). SeaMoon-ESM2(x5) achieved NSSE=0.37 on this protein, whose structurally closest train protein, the bacterial penicillin-binding protein 1B (7LQ6A, TM-score=0.44), exhibits a different 3D fold and different motions.

The TM-align algorithm[66], which we used to compute TM-scores, compares protein structures by treating them as rigid bodies that cannot deform. This approach might underestimate the similarity between structures that actually represent the same fold in different conformational states. To address this limitation, we re-evaluated the TM-scores between the most structurally similar train-test protein pairs using Kpax flexible fragment-based structural alignment algorithm[67]. Unlike TM-align, Kpax can accommodate conformational changes by aligning protein fragments independently. The cumulative NSSE loss curves obtained on the Kpax-based intermediate and difficult subsets show the same trends as those computed from the TM-align-based subsets (**Fig. S4**). This analysis confirms that SeaMoon is able to recognise previously seen folds and generalise to unseen ones. Furthermore, while the structure-aware pLMs consistently yield the best performance across the different categories, ESM3 becomes more advantageous

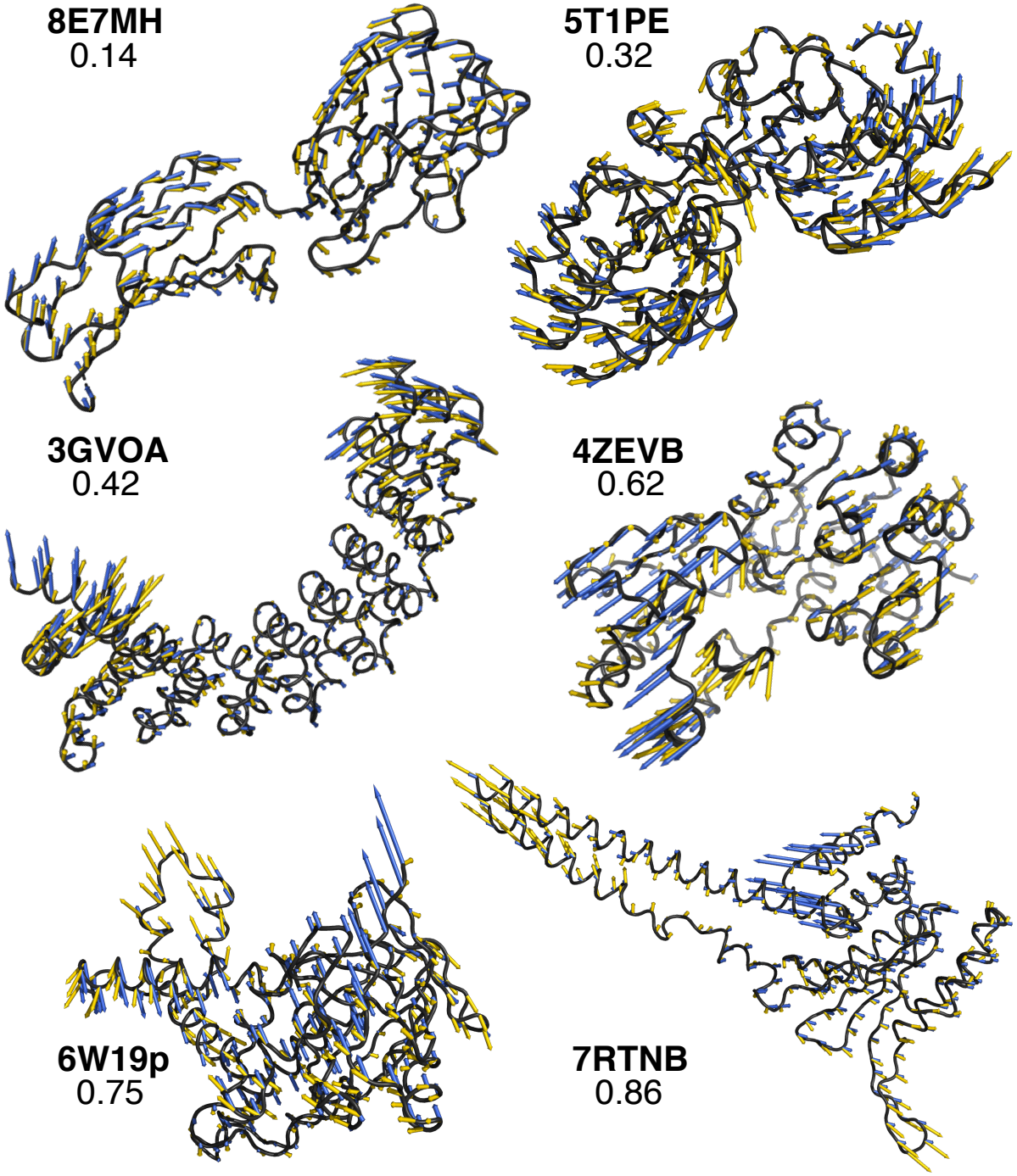

Figure 3: **Examples of predictions.** They allow for a visual assessment of how well the predicted vectors (in blue) approximate the ground-truth motions (in yellow) at different levels of NSSE (indicated on each panel). For each example, the query conformation is shown in black cartoons and labelled with its PDB chain identifier (in bold). We obtained the predicted vectors with SeaMoon-ESM2(x5).

compared to ProstT5 as the similarity to the training set decreases (**Table I**, **Fig. 2**, and **Fig. S4**). This observation suggests that ESM3 larger parameter count and access to functional properties during pre-training may provide it with better generalisation capabilities.

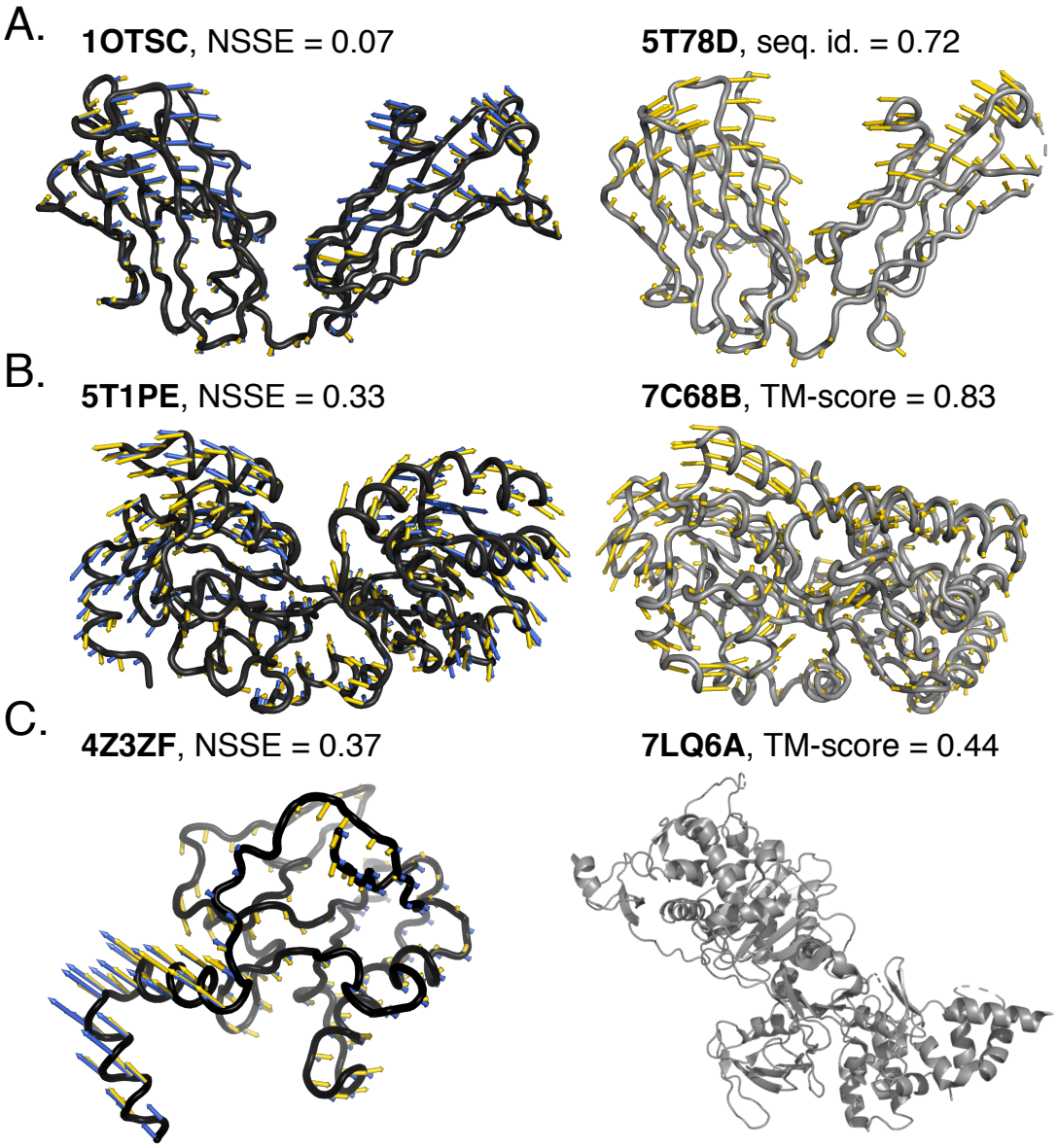

Figure 4: **Examples of predictions for test proteins with decreasing similarity to the training set.** The conformations are shown in cartoons and labelled with their PDB chain identifier. The ground-truth and SeaMoon-ESM2(x5) predicted motions are depicted with yellow and blue arrows, respectively. 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.

SeaMoon is complementary to the normal mode analysis

We observed a substantial overlap between the sets of successful predictions generated by SeaMoon and NMA, as well as between their respective NSSE distributions (**Fig. 2A**, inset, and **Fig. S5**). Specifically, SeaMoon base version with ESM2 embeddings generated acceptable predictions for 60% of the top-100 test proteins best-predicted by the NMA, and this proportion reaches 75% with ProstT5 embeddings (**Fig. 2A**, inset). Using implicit structural knowledge allowed recovering elastic motions, such as that exhibited by the mammalian plexin A4 ectodomain (**Fig. S6**, 5L5LB, NSSE=0.28). Reciprocally, about half of the top-100 proteins best-predicted by SeaMoon exhibit motions accessible to the NMA (**Fig. 2A**, inset).

Most of the motions well captured by the two approaches involve a large portion of the protein and correspond to large conformational changes. They include functional opening-closing motions of virulence factors, thermophilic proteins, metalloenzymes, periplasmic binding proteins, dehydrogenases, glutamate receptors, and antibodies (**Fig. S7**). To further validate SeaMoon's ability to predict a wide range of opening-closing motions, we evaluated its performance on the iMod benchmark[68], which was specifically designed to assess coarse-grained elastic network models (**Fig. S8**). SeaMoon-ProstT5(x5) matched the success rate of the NMA on this dataset, capturing the motions of 12 out of 21 proteins with high accuracy (NSSE<0.4), including adenylate kinase hinge motion (1LAFE) and the chaperone GroEL complex motion (1AONA). In addition, SeaMoon-ProstT5(x5) outperformed the NMA on the two shear motions in the dataset, exhibited by the *E. coli* DNA Polymerase III (1MMIA) and the adenosyl-cobinamide kinase (1C9KA), and on the allosteric motion of Aspartate carbamoyltransferase (1EKXA).

Despite the overall agreement between the two approaches, the NMA performed extremely poorly for a quarter to a third of SeaMoon top-100 test proteins (**Fig. 2A**, inset). The associated motions tend to be localised, with a median value of collectivity $\kappa = 0.20$. More broadly, we confirmed this trend by statistically assessing whether the set of motions well captured by the NMA were enriched or depleted in motions of different sizes (**Fig. 5**). The set of acceptable NMA predictions contains about twice more global motions ($\kappa > 0.60$) than the full test set and four times fewer localised motions ($\kappa < 0.30$, **Fig. 5A**). By contrast, the sets of acceptable SeaMoon predictions do contain roughly the same proportions of localised motions as in the full test set. These results indicate that SeaMoon is more capable of handling localised motions than the NMA. The bacterial toxin PemK provides an illustrative example of this capability (**Fig. 6A**). SeaMoon-ESM2(x5) captured the PemK's loop L12 motion with high precision (**Fig. 6A**, NSSE=0.24) whereas the NMA failed to delineate the mobile region in the protein and to infer its direction of movement (**Fig. 6A**, 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[69].

In addition, SeaMoon allows to go beyond the elastic approximation of the NMA and hence, may be a better option for modeling deformations that imply large secondary structure rearrangements, such as fold-switches. This advantage is revealed by the protective antigen (PA) from anthrax (**Fig. 6B**). 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. 6B**). By contrast, the NMA predicted a breathing motion poorly approximating the ground-truth one (**Fig. 6B**). PA's ∼30Å-large conformational transition is essential for the translocation of the bacterium's edema and lethal factors to the host cell[70]. Neither PemK nor PA have any

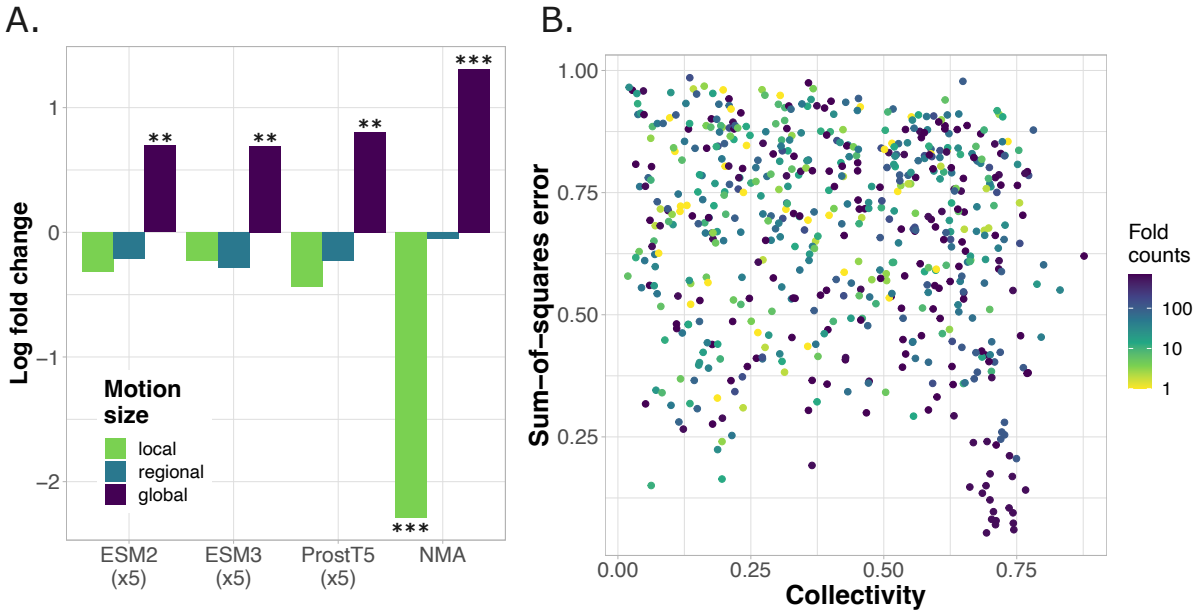

Figure 5: **SeaMoon performance depending on the type of motion and fold representativity. A.** Log_2 fold change in the proportion of local, regional and global motions between the set of acceptable predictions (NSSE<0.6), for each method, and the full test set. The categories of motions are defined based on collectivity computed on the ground-truth motions. Local: $\kappa \leq 0.30$. Regional: $0.30 < \kappa \leq 0.60$. Global: $\kappa > 0.60$. The stars indicate the significance of the enrichment or depletion, as estimated by a hypergeometric test, namely ** for a p-value below 10^{-8} and *** below 10^{-40} (see also **Table S3**). **B.** NSSE computed for SeaMoon-ESM2(x5) in function of motion collectivity and fold (CATH topology) counts in the training set. Each dot correspond to a protein and we report collectivity for the ground-truth motion best approximated by the predictions. Only test proteins with at least one annotated CATH domain are shown. In case of multiple domains, we consider the maximum fold count value.

detectable sequence similarity to the training set. SeaMoon likely leveraged information coming from training proteins with similar folds and functions from other bacteria[71,72].

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)[73–75] 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. 7** 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. 7A**). The associated NSSE were all smaller than 0.5 (**Fig. 7B**). By inspecting the training set, we could identify several major capsid proteins from other bacteriophages sharing the same HK97-like fold

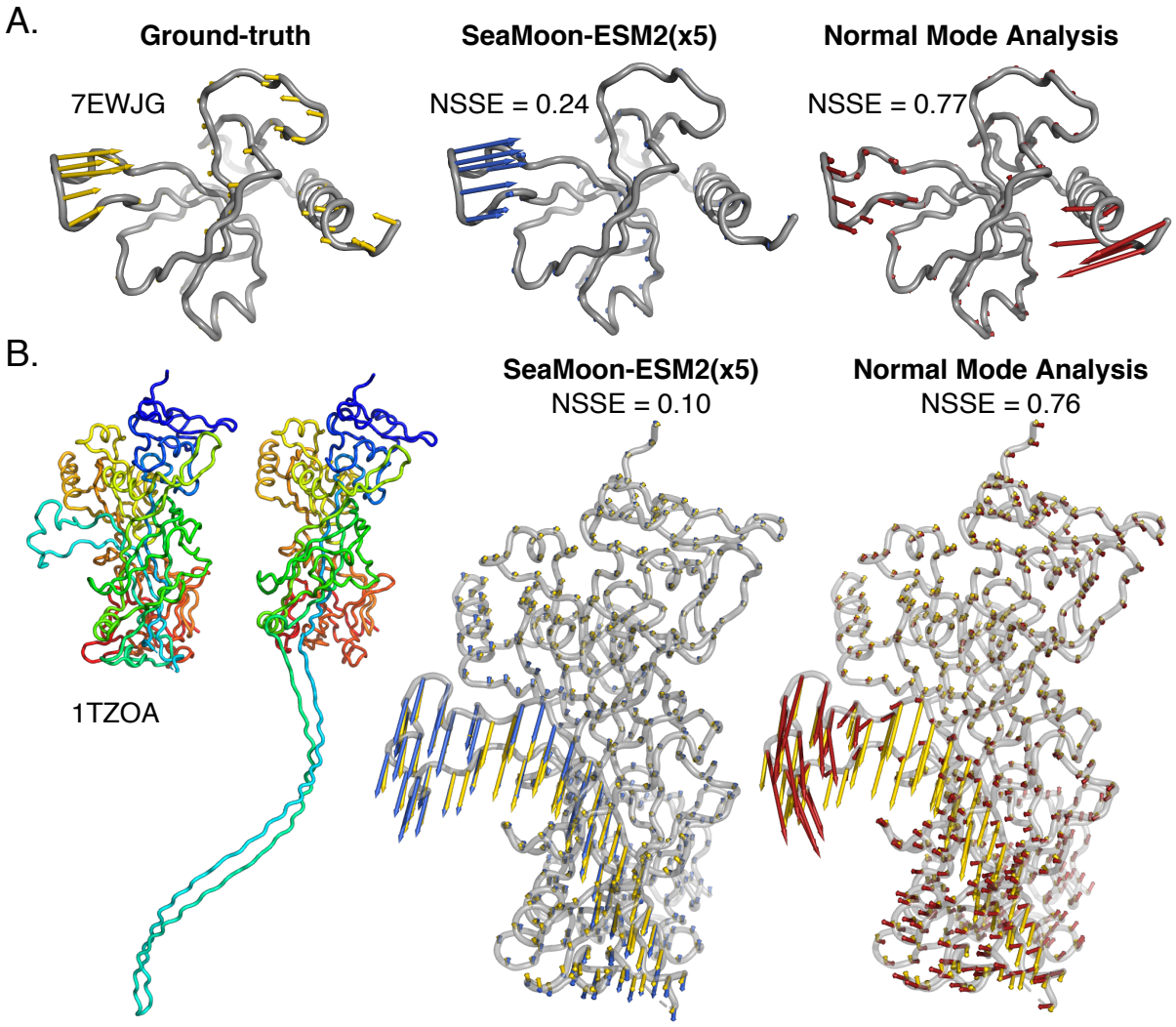

Figure 6: **Examples of motions well predicted by SeaMoon and not by the NMA.** The arrows depicted in yellow, blue and red on the grey 3D structures represent the ground-truth motions and the best-matching predictions from SeaMoon-ESM2(x5) and the NMA, respectively. **A.** Bacterial toxin PemK (PDB 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.

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[76].

Influence of fold representativity

Protein folds are not evenly represented in the PDB, with some folds being significantly more abundant than others. Yet, we chose not to correct for this bias in our PDB-wide training set of conformational collections, because the same fold may exhibit different

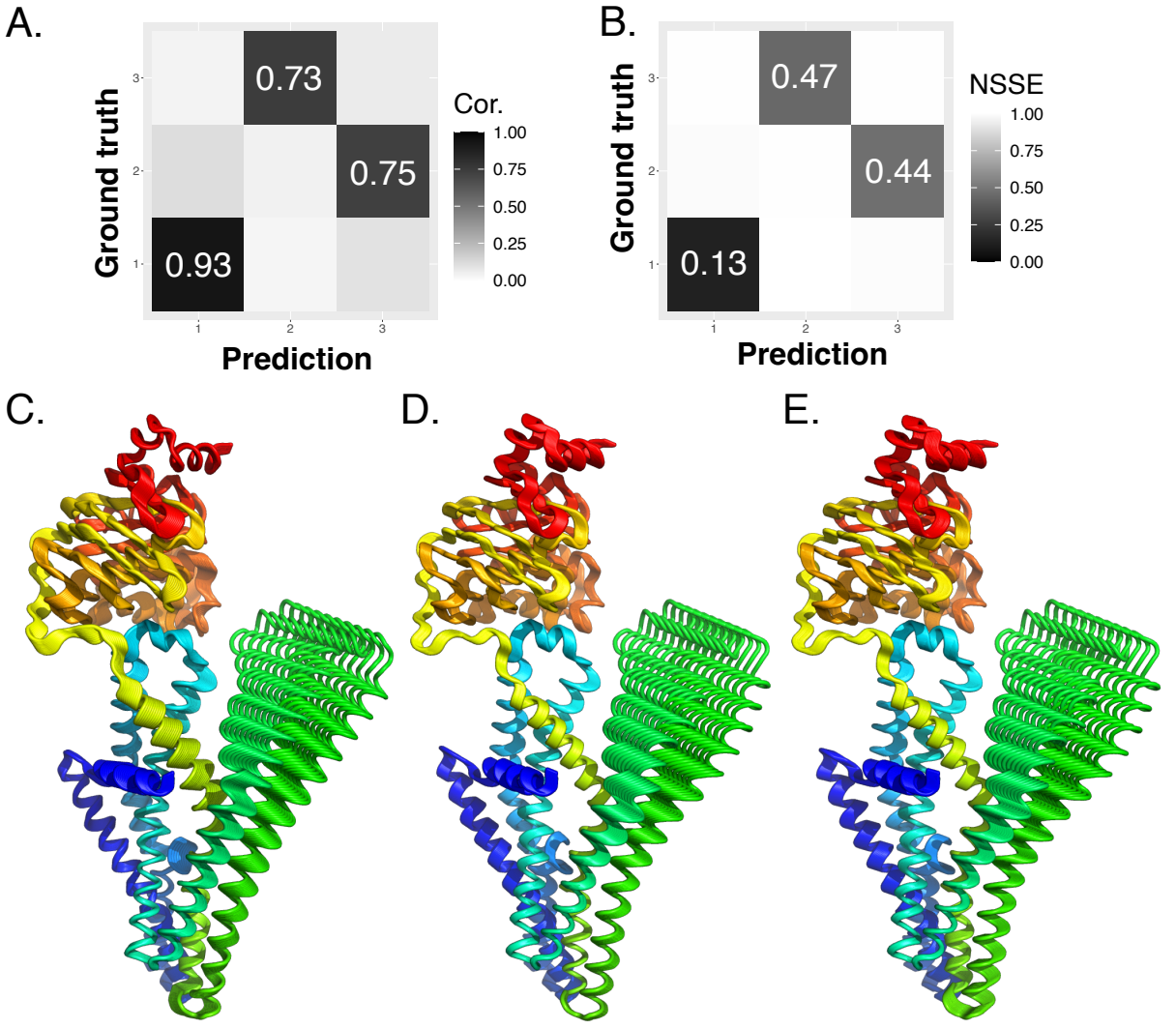

Figure 7: **Motion subspace comparison and deformation trajectories. A-B.** Ogopogo major capsid protein motion subspace. PDB code: 8ECN, chain B. **A.** Pairwise similarities measured as Pearson correlations between the ground-truth motions and SeaMoon-ProstT5(x5) predictions. **B.** Pairwise discrepancies 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.

motions in different collections. This design choice raises two important questions. First, does SeaMoon performance remain consistent on a redundancy-reduced version of the test set? To address this, we computed average best-matching motion pair errors for each of the 252 folds represented in the test set, using CATH topologies as fold definitions[77]. This analysis showed that the per-fold performance estimates align well with those computed at the protein level (**Fig. 8**, compare panels A and B). The success rate computed for SeaMoon-ProstT5(x5) is 37% (at NSSE<0.6) compared to 40% without redundancy reduction. Moreover, SeaMoon performance remains similar to or better than the NMA

(**Fig. 8A-B**). The second question concerns whether the motions of the folds that are over-represented in the training set are consistently better captured than those of under-represented folds (**Fig. 5B**). We tested this hypothesis by fitting a linear regression between the per-fold NSSEs and the fold frequency counts in our training set. We found no significant association between fold representativity and the per-fold average NSSEs, and only a weak association with the per-fold minimum NSSEs (**Fig. 8C-D**, adjusted R^2 of 0.20 for SeaMoon-ProstT5(x5)). These results demonstrate that SeaMoon predictions are not significantly biased by the unbalanced representation of protein folds in the training set. Furthermore, SeaMoon-ProstT5(x5) success rate at NSSE<0.6 computed on the test proteins that do not share any fold with the training set is 43% (12 out of 28 proteins), similar to the success rate computed over the full test set (40%).

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 S4** and **Fig. 9**). 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. 9**, 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 S4** and **Fig. 9**). Moreover, a 7-layer transformer architecture (see *Methods*) underperformed SeaMoon's convolutional neural network, despite having roughly the same number of free parameters (**Fig. 9**, 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. 9**, 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

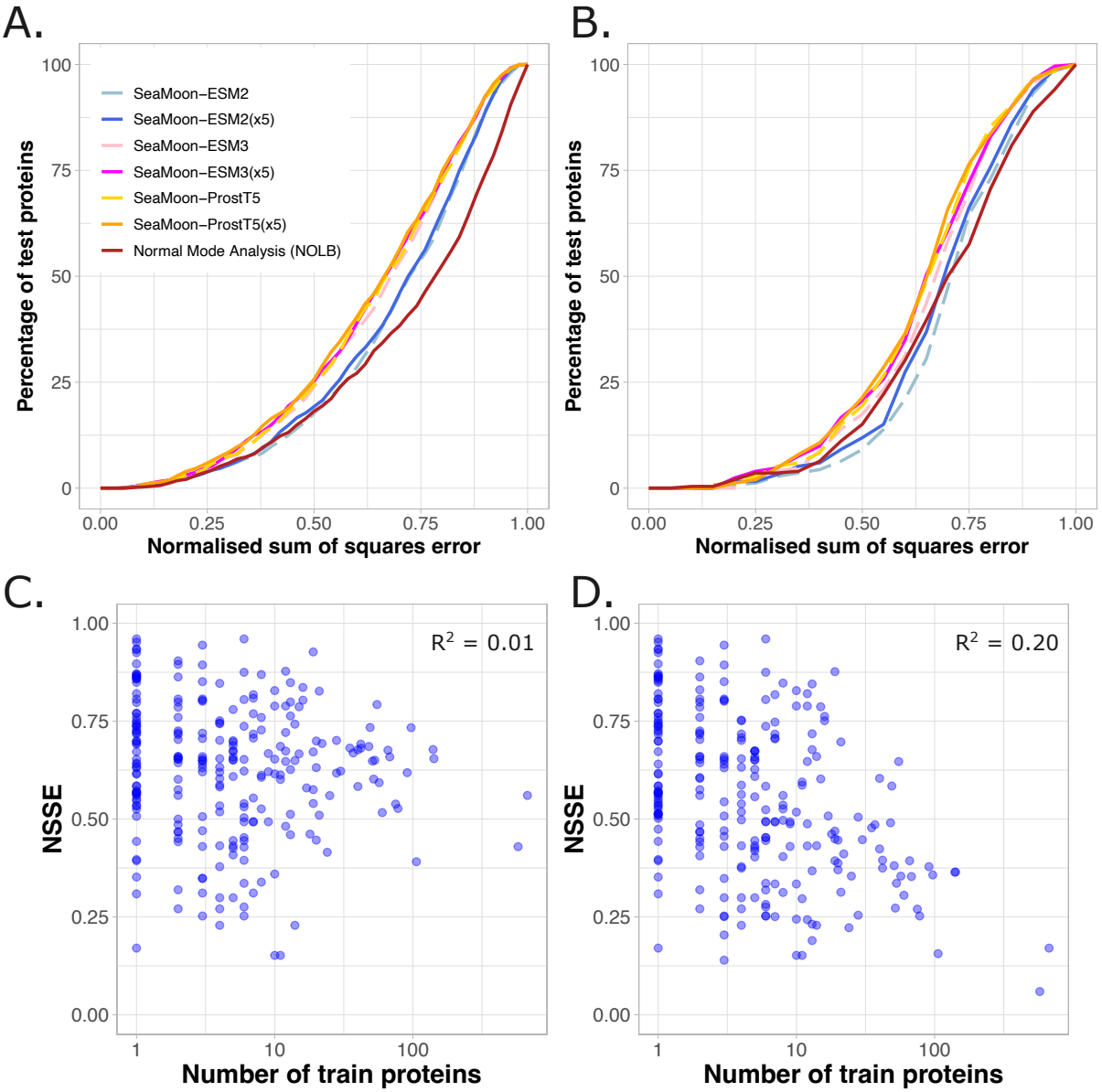

Figure 8: **Influence of fold representativity. A-B.** Cumulative NSSE computed per protein (A) and per fold (B), defined as topology in the CATH classification. The per-fold NSSEs are computed as the average NSSEs over the proteins associated to each fold. **C-D.** Per-fold NSSE in function of the number of train proteins containing at least one domain from this fold. The per-fold NSSEs are computed as the average NSSEs (C) of the minimum NSSEs (D) over the proteins associated to each fold. We report the adjusted R-squared values estimated by fitting a linear regression between the per-fold NSSEs and the log of the per-fold train protein counts.

velocity minimisation (**Fig. 7C-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.

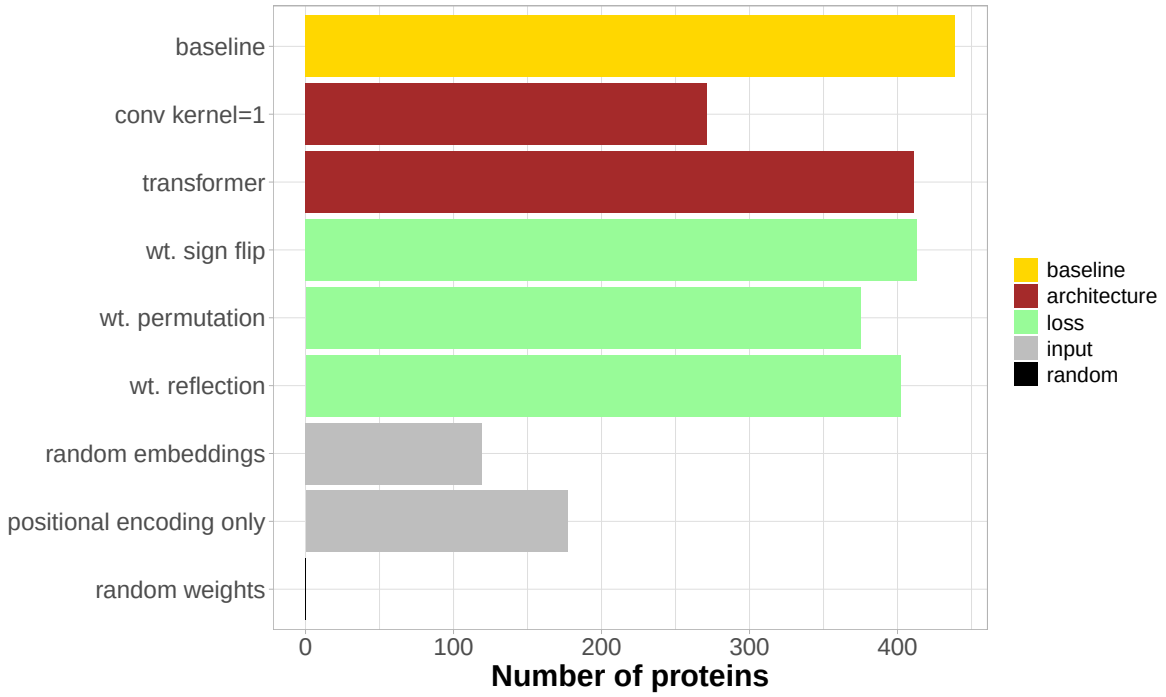

Figure 9: **Ablation study** We report the number of test proteins with at least one acceptable prediction. The baseline model is SeaMoon-ProstT5.

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 pLM embeddings generated from amino acid sequences alone. Our results demonstrated two key findings. First, pLMs that had been exposed to structural information during their pre-training outperformed those trained exclusively for sequence reconstruction. Second, they highlighted SeaMoon's ability to transfer knowledge about motions across distant homologs, leveraging the universal representation space of pLMs.

SeaMoon's transfer learning approach makes it suitable for systematically assessing the evolutionary conservation of protein motions. Moreover, it complements unsupervised methods that rely entirely on the 3D geometry of protein structures, such as Normal Mode Analysis (NMA). Additionally, SeaMoon is highly computationally efficient, and thus applicable on a large scale. 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.

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. They may result from experimental artifacts, most often of crystallographic origin. Nevertheless, establishing objective criteria to reliably distinguish artifactual conformations from biologically relevant ones remains infeasible. Our working assumption is that a subset of the conformational manifold in a protein collection always represents functional motions. To address this challenge, we designed our training loss function specifically to evaluate submanifolds by calculating the minimum error between each reference motion and the set of predicted motions, allowing the model to capture conformational diversity while mitigating the impact of potential artefacts.

Future work will explore incorporating explicit structural information. This should resolve the ambiguities in orienting predicted motions at inference time and could unlock SeaMoon's potential for sampling or deforming conformations. Furthermore, we will test the potential of augmenting both training and evaluation sets with *in silico* generated data. This could include motions derived from MD and NMA simulations, as well as protein conformations predicted by AlphaFold. In addition, future improvements will include more explicit descriptions of the protein environment. For instance, we plan to condition predictions on the presence of other molecules. This is particularly important because only 15% of all 750K protein chains available in the PDB are monomeric[62]. Consequently, motions induced by binding partners or ligands likely constitute a significant portion of our training dataset.

Despite these limitations and avenues for improvement, the current findings offer valuable insights for integrative structural biology. SeaMoon provides compact representations of continuous structural heterogeneity in proteins. Such representations have high interpretability and explanatory power for most of the cases. The estimated motion subspaces can readily be used to compute protein conformational entropy. Lastly, the SeaMoon 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.

Acknowledgments

The Sorbonne Center for Artificial Intelligence (SCAI) provided a salary to VL and computational resources. The authors thank Institut de Biologie Paris-Seine (IBPS) at Sorbonne Université for funding via a Collaborative Grant (Action Incitative) to DT. This work was 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. For the purpose of Open Access, a CC-BY public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission.

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.

Declaration of Interests

The authors declare no competing interests.

Methods

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[78].

Methods details

Datasets

To generate training data, we constructed a non-redundant set of conformational collections representing the whole PDB (as of June 2023) using DANCE[62]. To ensure high quality of the data, we replaced the raw PDB coordinates with their updated and optimised versions from PDB-REDO whenever possible[79]. 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 5 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[62]. 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 7 335 collections, randomly split between train (70%), validation (15%) and test (15%) sets.

In addition, we manually selected the conformational collections corresponding to the proteins from the iMod benchmark[68] and put them in our test set. The iMod benchmark is derived from the molecular motions database MolMovDB[80] and was previously used to assess coarse-grained elastic network model-based flexible fitting methods[81]. It comprises pairs of open-closed conformations for a couple of tens of proteins that represent a wide variety of motions, predominantly hinge motions but also shear and other complex motions. We could identify 21 conformational collections containing both open and closed conformations and complying with all our criteria, except for the percentage of variance explained by the first principal component (which can go down to 60% on this benchmark set).

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[62]. 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[23] and ESM3-small (1.4B)[63], as well as Protein sequence-structure T5[64]. ESM2-T33-650M-UR50 is a BERT[82] style 650-million-parameter encoder-only transformer architecture trained on all clusters from Uniref50[83,84], a version of UniProt[85] 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[86] 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[87], OAS[88] and the PDB[49], augmented with synthetic sequences generated by an inverse folding model[63]. Protein sequence-structure T5 is a bilingual pLM trained on a high-quality clustered version of the AlphaFold Protein Structure Database[89,90] to translate 1D sequences of amino acids into 1D sequences of 3Di tokens representing 3D structural states[91] and vice versa. The 3Di alphabet, introduced by the 3D-alignment method Foldseek[91], 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[22], an encoder-decoder transformer architecture[92] trained on reconstructing corrupted amino acids from the Big Fantastic Database[93] 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[94] 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[95], as well as 80% dropout[96]. 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 S4** and **Fig. 9** for more details. We implemented the models in PyTorch[97] 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,

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

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[62]. 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[98,99]. 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[100] 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 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.

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,

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

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)[73–75]. It is computed as an average of the normalised inner products of all the vectors in both subspaces,

$$\text{RMSIP} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{L}\left(\frac{(\vec{x}_{ik}^{\text{GT}})^T \vec{x}_{il}^{\text{ortho}}}{\|\vec{x}_{ik}^{\text{GT}}\|\|\vec{x}_{il}^{\text{ortho}}\|}\right)^2}, \tag{11}$$

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. S10**). A RMSIP score of 0.70 is considered an excellent correspondence while a score of 0.50 is considered fair[73].

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)[45]. 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[47] 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[48]. 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.

Protein fold-centered analysis

We defined protein folds as the topologies from the CATH classification[77]. We estimated the NSSE of each fold as the average NSSE computed over all test proteins containing the corresponding topology. At the protein level, we consider the NSSE of the best matching pair of predicted and ground-truth motions.

Protein properties

Sequence and structure similarity. We estimated sequence similarity between train and test proteins using MMseqs2[101] 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[66]. TM-score measures the topological similarity of protein structures. It ranges between 0 and 1, and a score higher than 0.5 assumes roughly the same fold. Furthermore, to account for protein flexibility in our estimation of structural similarity, we used the flexible structural alignment functionality from Kpax (version 5.1.1)[67,102]. Kpax implements a tiled dynamic programming algorithm that optimally partitions the input structures into rigid segments and aligns them. This strategy makes the TM-score estimation robust to changes in the relative positions and orientations of the identified segments, and thus to conformational changes. For each test protein, we computed the flexible Kpax TM-score with its best training hit identified by TM-align.

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[103,104] 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 $1/L$.

Visualisation

Three-dimensional protein structures were visualised and rendered using PyMOL[105]. All plots and statistical visualisations were generated using the ggplot2 R package version

4.2[106]. Motion directions were represented as 3D arrows overlaid on the structures using a custom Python script developed in-house that utilises PyMOL's CGO module and NumPy[107] for vector calculations.

Quantification and statistical analysis

We assessed the statistical significance of performance differences between ESM3 and ProstT5 with a paired Wilcoxon signed-rank test[108,109] performed using the *wilcox.test* function implemented in the R Stats Package version 4.2.2[110]. We assessed the statistical significance of the enrichments and depletions in motion types reported in Figure 3A with hypergeometric tests using the *phyper* function from the same package. Furthermore, we quantified the significance and statistical reliability of SeaMoon performance by comparing it with a random baseline and by performing bootstrap resampling, as described below.

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.

Success rate estimation. We estimated the success rate of a method as the percentage of test proteins for which it approximated at least one ground-truth motion with normalised sum-of-squares error NSSE<NSSE_{cut}, where NSSE_{cut} is set to 0.6 by default. To assess the statistical reliability of our success rate estimates, we used bootstrap resampling. Specifically, we generated 1 000 bootstrap random samples of $N^{2/3}$ test proteins, where N =1 121 is the cardinality of the full test set, allowing for replacement. For each method and each NSSE cutoff considered, we computed the average success rate, as well as the interquartile range and 95% central range over the 1 000 bootstrap samples.

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER	ADDITIONAL INFORMATION
Deposited data			
Protein motion dataset	This paper	10.5281/zenodo.13833309	https://zenodo.org/records/13833309
SeaMoon trained models	This paper	10.5281/zenodo.15616637	https://github.com/PhyloSofS-Team/seamoon
Software and algorithms			
Python	Python Software Foundation	RRID:SCR_008394	Version 3.8+
PyTorch	[97]	RRID:SCR_018536	Version 2.1.0
PyMOL	[105]	RRID:SCR_000305	Version 2.5
R	R Core Team	RRID:SCR_001905	Version 4.2.2
ggplot2	[106]	RRID:SCR_014601	Version 4.2
TM-align	[66]	N/A	Version 20220412
Kpax	[67,102]	N/A	Version 5.1.1
ESM-2	[23]	N/A	https://huggingface.co/facebook/esm2_t33_650M_UR50D
ESM-3	[63]	N/A	https://huggingface.co/EvolutionaryScale/esm3-sm-open-v1
ProstT5	[64]	N/A	https://huggingface.co/Rostlab/ProstT5
DANCE	[62]	10.5281/zenodo.15617758	https://github.com/PhyloSofS-Team/DANCE
HOPMA	[48]	10.5281/zenodo.15616796	https://github.com/elolaine/HOPMA
NOLB	[47]	N/A	https://team.inria.fr/nano-d/software/nolb-normal-modes/
SeaMoon code	This paper	10.5281/zenodo.15616637	https://github.com/PhyloSofS-Team/seamoon
Other			
Protein Data Bank	wwPDB	RRID:SCR_012820	https://www.rcsb.org/
PDB-redo	[79]	N/A	https://pdb-redo.eu

References

[1] Thomas J Lane. Protein structure prediction has reached the single-structure frontier. *Nature Methods*, 20(2):170–173, 2023.

[2] Mitchell D Miller and George N Phillips. Moving beyond static snapshots: Protein dynamics and the protein data bank. *Journal of Biological Chemistry*, 296, 2021.

[3] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[5] D Sala, F Engelberger, HS Mchaourab, and J Meiler. Modeling conformational states of proteins with alphafold. *Current Opinion in Structural Biology*, 81:102645, 2023.

[6] Bulat Faezov and Roland L Dunbrack Jr. Alphafold2 models of the active form of all 437 catalytically-competent typical human kinase domains. *bioRxiv, doi: 10.1101/2023.07.21.550125*, 2023.

[7] Lim Heo and Michael Feig. Multi-state modeling of g-protein coupled receptors at experimental accuracy. *Proteins: Structure, Function, and Bioinformatics*, 90(11):1873–1885, 2022.

[8] Björn Wallner. Afsample: improving multimer prediction with alphafold using massive sampling. *Bioinformatics*, 39(9):btad573, 2023.

[9] Björn Wallner. Improved multimer prediction using massive sampling with alphafold in casp15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1734–1746, 2023.

[10] Isak Johansson-Åkhe and Björn Wallner. Improving peptide-protein docking with alphafold-multimer using forced sampling. *Frontiers in Bioinformatics*, 2:85, 2022.

[11] Marc F Lensink, Guillaume Brysbaert, Nessim Raouraoua, Paul A Bates, Marco Giulini, Rodrigo V Honorato, Charlotte van Noort, Joao M C Teixeira, Alexandre M J J Bonvin, Ren Kong, Hang Shi, Xufeng Lu, Shan Chang, Jian Liu, Zhiye Guo, Xiao Chen, Alex Morehead, Raj S Roy, Tianqi Wu, Nabin Giri, Farhan Quadir, Chen Chen, Jianlin Cheng, Carlos A Del Carpio, Eichiro Ichiishi, Luis A Rodriguez-Lumbreras, Juan Fernandez-Recio, Ameya Harmalkar, Lee-Shin Chu, Sam Canner, Rituparna Smanta, Jeffrey J Gray, Hao Li, Peicong Lin, Jiahua He, Huanyu Tao, Sheng-You Huang, Jorge Roel-Touris, Brian Jimenez-Garcia,

Charles W Christoffer, Anika J Jain, Yuki Kagaya, Harini Kannan, Tsukasa Nakamura, Genki Terashi, Jacob C Verburgt, Yuanyuan Zhang, Zicong Zhang, Hayato Fujuta, Masakazu Sekijima, Daisuke Kihara, Omeir Khan, Sergei Kotelnikov, Usman Ghani, Dzmitry Padhorny, Dmitri Beglov, Sandor Vajda, Dima Kozakov, Surendra S Negi, Tiziana Ricciardelli, Didier Barradas-Bautista, Zhen Cao, Mohit Chawla, Luigi Cavallo, Romina Oliva, Rui Yin, Melyssa Cheung, Johnathan D Guest, Jessica Lee, Brian G Pierce, Ben Shor, Tomer Cohen, Matan Halfon, Dina Schneidman-Duhovny, Shaowen Zhu, Rujie Yin, Yuanfei Sun, Yang Shen, Martyna Maszota-Zieleniak, Krzysztof K Bojarski, Emilia A Lubecka, Mateusz Marcisz, Annemarie Danielsson, Lukasz Dziadek, Margrethe Gaardlos, Artur Gieldon, Adam Liwo, Sergey A Samsonov, Rafal Slusarz, Karolina Zieba, Adam K Sieradzan, Cezary Czaplewski, Shinpei Kobayashi, Yuta Miyakawa, Yasuomi Kiyota, Mayuko Takeda-Shitaka, Kliment Olechnovic, Lukas Valancauskas, Justas Dapkunas, Ceslovas Venclovas, Bjorn Wallner, Lin Yang, Chengyu Hou, Xiaodong He, Shuai Guo, Shenda Jiang, Xiaoliang Ma, Rui Duan, Liming Qui, Xianjin Xu, Xiaoqin Zou, Sameer Velankar, and Shoshana J Wodak. Impact of alphafold on structure prediction of protein complexes: The casp15-capri experiment. *Proteins*, 91(12):1658–1683, Dec 2023.

[12] Nessim Raouraoua, Claudio Mirabello, Thibaut Véry, Christophe Blanchet, Björn Wallner, Marc F Lensink, and Guillaume Brysbaert. Massivefold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. *Nature Computational Science*, pages 1–5, 2024.

[13] Pedro Sfriso, Miquel Duran-Frigola, Roberto Mosca, Agustí Emperador, Patrick Aloy, and Modesto Orozco. Residues coevolution guides the systematic identification of alternative functional conformations in proteins. *Structure*, 24(1):116–126, 2016.

[14] 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. *Advances in Enzyme Regulation*, 31:121–181, 1991.

[15] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317, 1994.

[16] Angel R Ortiz, Andrzej Kolinski, Piotr Rotkiewicz, Bartosz Ilkowski, and Jeffrey Skolnick. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins: Structure, Function, and Bioinformatics*, 37(S3):177–185, 1999.

[17] Alan S Lapedes, Bertrand G Giraud, LonChang Liu, and Gary D Stormo. Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Lecture Notes-Monograph Series*, pages 236–256, 1999.

[18] BG Giraud, John M Heumann, and Alan S Lapedes. Superadditive correlation. *Physical Review E*, 59(5):4983, 1999.

[19] 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*, pages 12–20, 2005.

[20] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[21] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669, 2021.

[22] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022.

[23] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[24] Yogesh Kalakoti and Björn Wallner. AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Communications Biology*, 8(1):373, 2025.

[25] Hannah K Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, pages 1–3, 2023.

[26] Diego Del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with alphafold2. *Elife*, 11:e75751, 2022.

[27] Richard A Stein and Hassane S Mchaourab. Speach_af: Sampling protein ensembles and conformational heterogeneity with alphafold2. *PLOS Computational Biology*, 18(8):e1010483, 2022.

[28] Lauren L Porter, Irina Artsimovitch, and César A Ramírez-Sarmiento. Metamorphic proteins and how to find them. *Current Opinion in Structural Biology*, 86:102807, 2024.

[29] Devlina Chakravarty and Lauren L Porter. Alphafold2 fails to predict protein fold switching. *Protein Science*, 31(6):e4353, 2022.

[30] Devlina Chakravarty, Joseph W Schafer, Ethan A Chen, Joseph Thole, and Lauren Porter. Alphafold2 has more to learn about protein energy landscapes. *bioRxiv*, pages 2023–12, 2023.

[31] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

[32] Haochuan Chen, Benoît Roux, and Christophe Chipot. Discovering reaction pathways, slow variables, and committor probabilities with machine learning. *Journal of Chemical Theory and Computation*, 19(14):4414–4426, 2023.

[33] Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, 71(1):361–390, 2020.

[34] Zineb Belkacemi, Paraskevi Gkeka, Tony Lelièvre, and Gabriel Stoltz. Chasing collective variables using autoencoders and biased trajectories. *Journal of Chemical Theory and Computation*, 18(1):59–78, 2021.

[35] Luigi Bonati, GiovanniMaria Piccini, and Michele Parrinello. Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences*, 118(44):e2113533118, 2021.

[36] Yihang Wang, Joao Marcelo Lamim Ribeiro, and Pratyush Tiwary. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current Opinion in Structural Biology*, 61:139–145, 2020.

[37] João Marcelo Lamim Ribeiro, Pablo Bravo, Yihang Wang, and Pratyush Tiwary. Reweighted autoencoded variational bayes for enhanced sampling (rave). *The Journal of Chemical Physics*, 149(7), 2018.

[38] Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, He Zhang, Shidi Tang, Hongxia Hao, Peiran Jin, Chi Chen, Frank Noé, Haiguang Liu, and Tie-Yan Liu. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, 6(5):558–567, 2024.

[39] Jiarui Lu, Bozitao Zhong, and Jian Tang. Score-based enhanced sampling for protein molecular dynamics. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

[40] Venkata K Ramaswamy, Samuel C Musson, Chris G Willcocks, and Matteo T Degiacomi. Deep learning protein conformational space with convolutions and latent interpolations. *Physical Review X*, 11(1):011052, 2021.

[41] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[42] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis

Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[43] Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, pages published at the MLDD workshop, ICLR 2023, 2023.

[44] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, pages published at the GenBio workshop, NeurIPS 2023, 2024.

[45] Steven Hayward and Nobuhiro Go. Collective variable description of native protein dynamics. *Annual Review of Physical Chemistry*, 46(1):223–250, 1995.

[46] Sergei Grudinin, Elodie Laine, and Alexandre Hoffmann. Predicting protein functional motions: an old recipe with a new twist. *Biophysical Journal*, 118(10):2513–2525, 2020.

[47] Alexandre Hoffmann and Sergei Grudinin. Nolb: Nonlinear rigid block normal-mode analysis method. *Journal of Chemical Theory and Computation*, 13(5):2123–2134, 2017.

[48] Elodie Laine and Sergei Grudinin. Hopma: Boosting protein functional dynamics with colored contact maps. *The Journal of Physical Chemistry B*, 125(10):2577–2588, 2021.

[49] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[50] Theresa A Ramelot, Roberto Tejero, and Gaetano T Montelione. Representing structures of the multiple conformational states of proteins. *Current Opinion in Structural Biology*, 83:102703, 2023.

[51] Patrick Bryant and Frank Noé. Structure prediction of alternative protein conformations. *Nature Communications*, 15(1):7328, 2024.

[52] Jill Trewhella. Recent advances in small-angle scattering and its expanding impact in structural biology. *Structure*, 30(1):15–23, 2022.

[53] 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. *Methods in Enzymology*, 677:263–290, 2022.

[54] Holger Flechsig and Toshio Ando. Protein dynamics by the combination of high-speed afm and computational modeling. *Current Opinion in Structural Biology*, 80:102591, 2023.

[55] Stephanie Wankowicz and James Fraser. Comprehensive encoding of conformational and compositional protein structural ensembles through mmcif data structure. *ChemRxiv*, 2023.

[56] Joseph IJ Ellaway, Stephen Anyango, Sreenath Nair, Hossam A Zaki, Nurul Nadzirin, Harold R Powell, Aleksandras Gutmanas, Mihaly Varadi, and Sameer Velankar. Identifying protein conformational states in the protein data bank: toward unlocking the potential of integrative dynamics studies. *Structural Dynamics*, 11(3), 2024.

[57] Mihaly Varadi, Stephen Anyango, Sri Devan Appasamy, David Armstrong, Marcus Bage, John Berrisford, Preeti Choudhary, Damian Bertoni, Mandar Deshpande, Grisell Diaz Leines, Joseph Ellaway, Genevieve Evans, Romana Gaborova, Deepti Gupta, Aleksandras Gutmanas, Deborah Harrus, Gerard J. Kleywegt, Weslley Morellato Bueno, Nurul Nadzirin, Sreenath Nair, Lukas Pravda, Marcelo Querino Lima Afonso, David Sehnal, Ahsan Tanweer, James Tolchard, Charlotte Abrams, Roisin Dunlop, and Sameer Velankar. PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education. *Protein Science*, 31(10), September 2022.

[58] Vivek Modi and Roland L Dunbrack Jr. Kincore: a web resource for structural classification of protein kinases and their inhibitors. *Nucleic Acids Research*, 50(D1):D654–D664, 2022.

[59] Mitchell I Parker, Joshua E Meyer, Erica A Golemis, and Roland L Dunbrack Jr. Delineating the RAS conformational landscape. *Cancer Research*, 82(13):2485–2498, 2022.

[60] Hedvig Tordai, Erzsebet Suhajda, Ian Sillitoe, Sreenath Nair, Mihaly Varadi, and Tamas Hegedus. Comprehensive collection and prediction of abc transmembrane protein structures in the ai era of structural biology. *International Journal of Molecular Sciences*, 23(16):8877, 2022.

[61] Gáspár Pándy-Szekeres, Jimmy Caroli, Alibek Mamyrbekov, Ali A Kermani, György M Keserű, Albert J Kooistra, and David E Gloriam. Gpcrdb in 2023: state-specific structure models using alphafold2 and new ligand resources. *Nucleic Acids Research*, 51(D1):D395–D402, 2023.

[62] Valentin Lombard, Sergei Grudinin, and Elodie Laine. Explaining conformational diversity in protein families through molecular motions. *Scientific Data*, 11(1):752, 2024.

[63] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, page eads0018, 2025.

[64] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024.

[65] Monika Chandravanshi, Reshama Samanta, and Shankar Prasad Kanaujia. Conformational trapping of a β-glucosides-binding protein unveils the selective two-step ligand-binding mechanism of abc importers. *Journal of Molecular Biology*, 432(20):5711–5734, 2020.

[66] Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005.

[67] David W Ritchie. Calculating and scoring high quality multiple flexible protein structure alignments. *Bioinformatics*, 32(17):2650–2658, 2016.

[68] José Ramón Lopéz-Blanco, José Ignacio Garzón, and Pablo Chacón. iMod: Multipurpose normal mode analysis in internal coordinates. *Bioinformatics*, 27(20):2843–2850, 2011.

[69] Do-Hee Kim, Sung-Min Kang, Sung-Min Baek, Hye-Jin Yoon, Dong Man Jang, Hyoun Sook Kim, Sang Jae Lee, and Bong-Jin Lee. Role of pemi in the staphylococcus aureus pemik toxin–antitoxin complex: Pemi controls pemk by acting as a pemk loop mimic. *Nucleic Acids Research*, 50(4):2319–2333, 2022.

[70] Alexandra J Machen, Mark T Fisher, and Bret D Freudenthal. Anthrax toxin translocation complex reveals insight into the lethal factor unfolding and refolding mechanism. *Scientific Reports*, 11(1):13038, 2021.

[71] Immanuel Dhanasingh, Eunsil Choi, Jeongeun Lee, Sung Haeng Lee, and Jihwan Hwang. Functional and structural characterization of deinococcus radiodurans r1 mazef toxin-antitoxin system, dr0416-dr0417. *Journal of Microbiology*, 59:186–201, 2021.

[72] David M Anderson, Michael J Sheedlo, Jaime L Jensen, and D Borden Lacy. Structural insights into the transition of clostridioides difficile binary toxin from prepore to pore. *Nature Microbiology*, 5(1):102–107, 2020.

[73] 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. *Proteins: Structure, Function, and Bioinformatics*, 36(4):419–424, 1999.

[74] Alejandra Leo-Macias, Pedro Lopez-Romero, Dmitry Lupyan, Daniel Zerbino, and Angel R Ortiz. An analysis of core deformations in protein superfamilies. *Biophysical Journal*, 88(2):1291–1299, 2005.

[75] Charles C David and Donald J Jacobs. Characterizing protein motions from structure. *Journal of Molecular Graphics and Modelling*, 31:41–56, 2011.

[76] Jennifer M Podgorski, Krista Freeman, Sophia Gosselin, Alexis Huet, James F Conway, Mary Bird, John Grecco, Shreya Patel, Deborah Jacobs-Sera, Graham Hatfull, Johann Peter Gogarten, Janne Ravantti, and Simon J White. A structural dendrogram of the actinobacteriophage major capsid proteins provides important structural insights into the evolution of capsid stability. *Structure*, 31(3):282–294, Mar 2023.

[77] Vaishali P Waman, Nicola Bordin, Andy Lau, Shaun Kandathil, Jude Wells, David Miller, Sameer Velankar, David T Jones, Ian Sillitoe, and Christine Orengo. Cath v4. 4: major expansion of cath by experimental and predicted structural data. *Nucleic Acids Research*, 53(D1):D348–D355, 2025.

[78] Valentin Lombard, Sergei Grudinin, and Elodie Laine. Data for "SeaMoon: from protein language models to continuous structural heterogeneity". `https://doi.org/10.5281/zenodo.13833309`, 2024.

[79] Robbie P Joosten, Fei Long, Garib N Murshudov, and Anastassis Perrakis. The pdb_redo server for macromolecular structure model optimization. *IUCrJ*, 1(4):213–220, 2014.

[80] Nathaniel Echols, Duncan Milburn, and Mark Gerstein. MolMovDB: Analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Research*, 31(1):478–482, jan 2003.

[81] Mustafa Tekpinar. Flexible fitting to cryo-electron microscopy maps with coarse-grained elastic network models. *Molecular Simulation*, pages 1–9, 2018.

[82] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[83] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[84] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

[85] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022.

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[87] Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, Tom Curtis, Alejandra Escobar-Zepeda, Tatiana A Gurbich, Varsha Kale, Anton Korobeynikov, Shriya Raj, Alexander B Rogers, Ekaterina Sakharova, Santiago Sanchez, Darren J Wilkinson, and Robert D Finn. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51(D1):D753–D759, Jan 2023.

[88] 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. *Protein Science*, 31(1):141–146, 2022.

[89] Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron LM Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, and Martin Steinegger. Clustering predicted structures at the scale of the known protein universe. *Nature*, 622(7983):637–645, 2023.

[90] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021.

[91] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, 2024.

[92] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[93] Martin Steinegger, Milot Mirdita, and Johannes Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*, 16(7):603–606, 2019.

[94] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[95] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *30th International Conference on Machine Learning*, volume 30, page 3. Atlanta, GA, 2013.

[96] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[97] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library.* Curran Associates Inc., Red Hook, NY, USA, 2019.

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

[99] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[100] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[101] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, Nov 2017.

[102] David W Ritchie, Anisah W Ghoorah, Lazaros Mavridis, and Vishwesh Venkatraman. Fast protein structure alignment using gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, 28(24):3274–3281, 2012.

[103] Rafael Brüschweiler. Collective protein dynamics and nuclear spin relaxation. *The Journal of Chemical Physics*, 102(8):3396–3403, 1995.

[104] F. Tama and Y. H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Engineering*, 14(1):1–6, Jan 2001.

[105] Warren L DeLano. The pymol molecular graphics system, 2002.

[106] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[107] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[108] David F Bauer. Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, 67(339):687–690, 1972.

[109] Hollander Myles, Douglas A Wolfe, and Eric Chicken. Nonparametric statistical methods. *Nonparametric Statistical Methods*, pages 27–33, 1973.

[110] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.