Contrastive learning unites sequence and structure in a global representation of protein space

Guy Yanai[1], Gabriel Axel[2], Liam M. Longo[3,4]* Nir Ben-Tal[2]*, Rachel Kolodny[1]*

[1]Department of Computer Science, University of Haifa, Haifa, Israel

[2]School of Neurobiology, Biochemistry and Biophysics, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel

[3]Earth-Life Science Institute, Institute of Science Tokyo, Tokyo, Japan

[4]Blue Marble Space Institute of Science, Seattle, USA

*To whom correspondence should be addressed. E-mail: llongo@elsi.jp; bental@tauex.tau.ac.il; trachel@cs.haifa.ac.il

Abstract

Amino acid sequence dictates the three-dimensional structure and biological function of proteins. Yet, despite decades of research, our understanding of the interplay between sequence and structure is incomplete. To meet this challenge, we introduce Contrastive Learning Sequence-Structure (CLSS), an AI-based contrastive learning model trained to co-embed sequence and structure information in a self-supervised manner. We trained CLSS on large and diverse sets of protein building blocks called domains. CLSS represents both sequences and structures as vectors in the same high-dimensional space, where distance relates to sequence-structure similarity. Thus, CLSS provides a natural way to represent the protein universe, reflecting evolutionary relationships, as well as structural changes. We find that CLSS refines expert knowledge about the global organization of protein space, and highlights transitional forms that resist hierarchical classification. CLSS reveals linkage between domains of seemingly separate lineages, thereby significantly improving our understanding of evolutionary design.

Introduction

Mapping the protein universe – that is, constructing representations for the relationships between all proteins – is essential to understanding the forces that have shaped protein emergence and evolution over the past 4 billion years [1-3]. To date, the most significant mapping achievements are hierarchical classifications [4-7], where building blocks of protein tertiary structure, called domains, are grouped based on sequence and/or structure similarity. Constructed with expert supervision, these databases are a repository of human insight and intuition. Yet, hierarchical classifications permit only restricted notions of relatedness and similarity, as inter-domain relationships are coarse grained into just a few nested tiers of full domains.

To capture more subtle inter-domain relationships, networks [3,8-10] and Euclidean maps of fixed-size vector representations [11-23] have been employed. Derived from pair-wise comparisons, networks can represent cross-fold similarities, where domains from ostensibly different evolutionary lineages share common sub-domain sized segments [24-28]. Similar sequence segments present in different domains [24], including ones that bridge different folds [25-27], hint at early evolutionary events [27,29]. These shared segments highlight fragment repurposing and recontextualization in the discovery and functional evolution of folds. Euclidian maps, on the other hand, are now generated by Protein Language Models (PLMs) – state-of-the-art computational models that are trained by self-supervision from vast amounts of data. Although both network- and PLM-based representations of protein space can combine sequence and structure modalities in various forms [19-21,23,30-32], they fail to provide an intuitive, joint representation of sequence and structure information.

A meaningful overview of the protein universe must provide a unified account of sequence and structure similarity, while accommodating the many-to-many relationship between them [33-35]. Moreover, by co-embedding multiple modalities, the paired data can guide each other, leading to enhanced performance. Contrastive Language-Image Pre-training (CLIP) [36], for example, co-embeds images and their natural language text descriptions. In CLIP, contrastive loss (CL) guides the image-text co-embedding, enabling state-of-the-art classification of images based on textual descriptions. We reason that the application of CL, and its pairing with protein language models (PLMs) [17,18], presents a unique opportunity to explicitly consider the complex interplay between sequence and structure.

Here, we report *Contrastive Learning Sequence-Structure* (CLSS), a protein language model, that embeds sequence segments using their corresponding structures as a guide, thereby co-embedding protein sequence and structure information. Inspired by CLIP, we used self-supervised contrastive learning on sequence and structure embeddings. Encouragingly, our trained model yields excellent agreement with curated labels, including fine details, for ~30,000 representative protein domains. In cases where seemingly disparate domains have similar embeddings, visual inspection reveals that CLSS embeddings are more reflective of structural similarities than expert-curated hierarchical classification. Our trained model provides meaningful maps of protein sequence and structure, and the relationships between them, offering an unprecedented, holistic view of the protein universe.

Results

The CLSS architecture

We designed and trained CLSS to co-embed protein sequence segments alongside domain structures (**Figure 1;** see **Methods** for more details). Briefly, CLSS is a deep network with two towers [36,37], one for structure and one for sequence. The structure tower is based on the frozen ESM3 structure encoder, and the sequence tower has the same architecture and initial weights as the ESM2 35-million-parameter model. Both encoders are followed by a trained adapter that averages the embeddings, reduces their size, and normalizes the output. Training over 80 epochs, we minimized contrastive loss in batches of 1440 domains culled from the domains predicted by AlphaFold 2 (AF2) [38]. Full domain structures were trained against randomly chosen sub-segments of varying length, from 10 residues to full domain, with the goal of encoding contextual information within domain segments.

CLSS unifies sequence space and structure space

Using the trained model, we calculated sequence and structure embeddings for two sets of representative domains taken from across the protein universe: *Dataset 1* comprises 31,696 ECOD domains taken from the 109 best-characterized protein evolutionary lineages (ECOD X-group), spanning 16 architectures. *Dataset 2* comprises 9,899 CATH domains from Røgen's set of sequence-similar but topologically distinct domain pairs [34]. For each domain, we calculated embeddings for three modalities: structure, full sequence, and a randomly selected sequence segment of length 20-60 residues (95,088 and 29,697 embeddings for Datasets 1 and 2, respectively). We visualized the embeddings as maps: two-dimensional t-SNE projections [39] of CLSS embeddings, where each point represents the embedding of one modality of a single domain.

Maps of Dataset 1 (**Figure 2**) show that, regardless of the data modality, CLSS embeddings agree well with structure labels from ECOD, placing domains from the same architecture, and even the same structure class (*i.e.*, all-α, all-β, α/β, and α+β domains), close to each other. Most importantly, maps from structure and sequence inputs are nearly identical, demonstrating that the model successfully co-embeds these two modalities. The distance distributions between the 32-dimensional CLSS embeddings from different modalities of the same domain in Dataset 1 (**Figure 3A**) show that the embeddings of sequence and structure are close to each other. This is true for all ECOD architectures, as seen in the architecture-dependent cumulative distance distributions between modalities (**Figure S6**). For comparison, embeddings from three other state-of-the-art PLMs – ESM3 (mean of the per-residue embeddings) [19], ProstT5 (mean of per-residue embeddings) [40], and ProTrek (the first embedded vector) [21], were calculated (**Figure 3, Figures S2, S3, S4**) – and found in separate regions of the joint map, rather than intertwined as in CLSS.

CLSS recapitulates expert-curated hierarchical classifications

To assess whether expert-curated hierarchical classifications are reflected in the CLSS embeddings, pairwise distances were calculated for Dataset 1 domains of varying degrees of similarity. Distributions were then estimated using a Gaussian kernel

density function. Both structure (**Figure 4A**) and sequence (**Figure 4C**) modalities yield virtually identical distributions, consistent with their highly correlated embeddings (**Figure 2** and **Figure 3, panel 1A**). Even though ECOD labels were not used during training [41], we find that embedding distance distributions follow the ECOD hierarchy, where domains from successively broader groupings have more and more dissimilar embeddings. Recall that ECOD H-groups, and to some extent X-groups, attempt to capture homologous relationships between domains, while architecture and class (all-α, all-β, α/β, and α+β) are structural classifications only. Nevertheless, both structure and sequence modalities separate the distributions of all levels. In contrast, the embedding distance distributions from other PLMs is significantly less informative of sequence or structure similarity (**Figure S4**).

CLSS robustly co-embeds the sequence and structure of metamorphic domains

Co-embedding of proteins with similar sequences but dissimilar structures poses a unique challenge to contrastive learning: In the sequence modality, the embeddings should be near each other (due to sequence similarity) but in the structure modality the embeddings should be far apart (due to structural dissimilarity). To understand if and how CLSS resolves these conflicting signals, we mapped the 'metamorphic' CATH domains of Dataset 2. To our surprise, the sequence and structure maps are generally similar to each other (**Figure 5, Figure S10, panel A**) and both modalities successfully group domains from the same architecture and super-architecture (**Figures 5A**, **5B** and **5D**). This result is likely a consequence of training on sequence segments because diverse segments can be extracted from a fold but CLSS still places these segments near the full domain in embedding space, as described below. Although the other PLMs compared here do not co-embed modalities – and thus, one does not expect metamorphic proteins to present a challenge – the resulting groupings are far less pronounced (**Figures S7, S8, S9**; *e.g.*, α-domains). Finally, a joint map of domains from Dataset 1 and Dataset 2 shows that the metamorphic CATH domains fall near their corresponding ECOD X-groups (**Figure S10**; note that t-SNE projections depend on the dataset mapped such that by including the diverse domains of Dataset 1 in the map, some CATH domain clusters contract).

CLSS encodes the context of domain sub-sequences

CLSS training included sequence segments of at least 10 residues. Thus, CLSS should be able to accurately embed domain sub-sequences as a third modality. The maps and distributions in **Figures 1-5** and **S2-S5** include embeddings for the sub-sequence modality. **Figure 2C** shows the map of the sub-sequence CLSS embeddings for Dataset 1, which is broadly similar to the maps obtained using the full sequence (**Figure 2D**) and the full structure (**Figure 2B**). Moreover, the embedding distance distributions from the sub-sequence modality (**Figure 4C**) follow the same trend as the other modalities, with domains from broader groupings resulting in progressively more distant embeddings. Nevertheless, embedding distance distributions for H-group and X-group level similarity are further away from each other than for the other two modalities. The distance distributions between embeddings from different modalities of the same domain in Dataset 1 (**Figure 2, panel 1A** and **Figure S4**) show that sub-sequence and sequence yield more similar embeddings than sub-sequence and structure. Neither ESM3, ProstT5, nor ProTrek yield informative embeddings for the

sub-sequence modality (**Figures 3D** and **S4**); their embedding distance distributions are virtually the same regardless of the level of the shared labels.

There are cases where the embedding of a sub-sequence is far from the embeddings of the full domain from which it came, even for CLSS. Often, these cases are associated with architectures that contain both α and β elements: Primarily mixed α+β and α/β (light green), α/β three-layered-sandwich (dark green), α/β barrel (green), and, to a lesser extent, α+β complex topology (brown) and α+β four layers (light brown; discussed in more detail below). This may reflect that α/β folds have significant fragment similarity between them, such that extracting a short fragment greatly diminishes information about the surrounding domain context. We will explore this.

Examples of CLSS disagreement with hierarchical classifications highlight transitional proteins

Are disagreements with hierarchical classifications due to limitations of CLSS, or do they reveal *bona fide* domain features that are not well-captured within the context of a hierarchical classification? To explore this question, **Figure 6** shows three examples of disagreement between ECOD and CLSS.

The first example highlights the challenges facing sequence-based HMMs when used for classification. We have previously observed sequence and structure connections between the β-trefoils (X-group 6) and β-propellers (X-group 5) [42,43]. Consistent with these results, the CLSS embeddings of these two β-folds form adjacent but distinct clusters. The domain e6tjeA1 annotated by the automatic ECOD pipeline as a β-trefoil, clusters with the β-propellers in the CLSS embedding space across all three modalities (**Figure 6A**). Inspection of the structure reveals that 6tje is a Rosetta-designed 9-bladed β-propeller called Cake5 [44], consistent with the CLSS embedding. The classification of e6tjeA1 by the automatic ECOD pipeline as a β-trefoil suggests that these folds are indeed evolutionarily related and provides an instructive example: Briefly, Cake5 must dimerize to form a complete β-propeller and the ECOD pipeline divided one chain of Cake5 into two separate domains: The first “β-trefoil domain” (e6tjeA1) comprises 3 β-propeller blades with a total of 12 β-strands, the same number as in a canonical β-trefoil. The second “domain” (e6tjeA2) encompasses just 7 β-strands and is correctly classified by ECOD as a β-propeller, despite being a smaller segment. As Cake 5 is a designed sequence (which can differ significantly from known protein families of the same fold), a repetitive sequence (where the number of repeats can vary between related domains, complicating the identification of domain boundaries), and a homo-oligomer (where the full domain is split across two chains), it is an exceptionally difficult protein to classify automatically.

A second example of disagreement is the localization of α+β four-layer sandwiches, within the greater α/β cluster on the CLSS map (**Figure 6B**; marked in ochre). Here, too, inspection affirms the CLSS clustering by revealing significant α/β-character. Domain e2nxfA1 (X-group 246), for example, has two juxtaposed β-α-β-α-β motifs that form the core of the fold. However, unlike a Rossmann domain – in which the two β-α-β-α-β motifs form a single extended β-sheet with a “doubly wound” 3-2-1-4-5-6 strand order – the β-α-β-α-β motifs of this domain pack against each other, face-to-face, to form the four-layer α-β-β-α sandwich structure [27]. The localization of the α+β four-layer sandwiches within the larger embedding space occupied by the Rossmannoid folds suggests the evolutionary potential for β-α-β-α-β motifs to interconvert between the two possible packing arrangements.

A final example of disagreement between CLSS embeddings and the ECOD hierarchy relates to the classification of single-stranded right-handed β-helices (X-group 207), hereafter referred to as β-helices. Although β-helices are classified as β proteins by ECOD, β-helix domains belonging to H-group 207.1 are located near α/β proteins on the CLSS map (green colored embeddings), whereas β-helix domains belonging to H-group 207.2 are located near other β-proteins, as would be expected (**Figure 6C**). Inspection of the structures once again affirms the CLSS embeddings: 207.1 β-helices do, in fact, have significant α/β-character. The ECOD hierarchical classifications scheme permits only a single architecture classification per grouping, and the β-helix highlights how this limitation can be misleading for groups with diverse structures. If H-groups 207.1 and 207.2 are indeed related, as ECOD suggests, the β-helices represent a remarkable case of protein structure exploration, bridging the space of β-folds with that of α/β-folds.

In summary, disagreement between hierarchical classification schemes, represented here by the ECOD database, and CLSS seem to reflect the weaknesses of hierarchical classification. Moreover, these examples suggest that over long timescales, protein space that is more continuous than hierarchal in nature.

The CLSS viewer

We provide a web interface for exploring a t-SNE projection of the CLSS embeddings of almost all ECOD domains from the ECOD-AF2 F100 dataset (https://gabiaxel.github.io/clss-viewer/). Individual domains or groups of domains can be searched for by their domain identifier (e.g., “e6tjeA1”), evolutionary grouping (e.g., “207”, “207.1”), and partial name (e.g., “lectin”). When an area of interest on the plot is selected, the list of associated domains, along with their ECOD classifications and (x, y) coordinates, are displayed. To facilitate data sharing, the URL is updated to stably reflect the current selection state. In addition, selected domain data can be downloaded as a CSV file for offline analysis.

Discussion

We introduce CLSS, a self-supervised sequence-structure co-embedding model, and demonstrate its use in mapping the protein universe. Given a query protein sequence, or structure, or even only a sequence segment, CLSS can embed it next to similar or related domains. The CLSS viewer provides an easy-to-use, interactive interface for exploring CLSS embeddings.

CLSS differs from established PLMs in several important ways: First, CLSS was trained on sub-domain sequence segments and the structure of the full domain within which the segments are situated; CLSS is not a masked protein language model. Consequently, CLSS embeddings contain meaningful domain, and even sub-domain, representations. Second, the contrastive learning procedure successfully co-embeds sequence and structure information. Thus, CLSS embeddings capture the interplay between these two modalities. Because of these two novel qualities, CLSS-derived representations of the protein universe are superior to those of ESM3, ProstT5, and ProTrek, as judged by its ability to reproduce known evolutionary relationships. CLSS also includes several quality-of-life improvements: With only 36M trainable parameters, it is considerably smaller than other state-of-the-art models. Thus, training purpose-built models with CLSS-like architectures is possible with only modest computational resources. Finally, CLSS uses a compact embedding space of only 32-

dimensions [45], facilitating data storage and searching [46]. These design choices provide an alternative view on the future of PLMs, with a better bias-variance tradeoff that creates more efficient and compact models [47].

CLSS contrastive loss aims to co-embed sequence segments, including ones far shorter than the full domain, near the structures that host them. This property can be useful when embedding metamorphic proteins because their sequence segments must be compatible with different structures. Indeed, even though co-embedding should be challenging for metamorphic domains, CLSS successfully maps CATH metamorphic domains (Dataset 2). By characterizing the sequence segments, or motifs, that can fit within the context of a specific structure, CLSS may aid in high-throughput library-based design approaches, and thus may facilitate protein design [48-51]. The motivation for sub-segment training is similar to that of protein threading [52] and early protein structure prediction approaches that quantified the compatibility between a sequence and a (potentially much larger) structure [53-55].

We argue that hierarchical domain classifications, like ECOD, are relevant and challenging benchmarks for PLMs because they curate expert knowledge about protein evolutionary interrelationships, which is informative for functional characterization [7,56,57]. Our view is a departure from current practice, which centers more on the success of a supervised downstream task, such as the prediction of Enzyme Commission (EC) numbers, gene ontology (GO) terms [20,22,23,30-32,41], or the identity of masked residues [19]. Assessment of features that are largely defined at the protein family level provide little information about the global organization of the embedding space. Instead, by painting PLM-based maps with expert knowledge of the protein universe at various degrees of abstraction [17,19,20,22,23,31,32,58], we can assess the large-scale and fine-scale structure of the embedding space with an unprecedented degree of explainability.

The agreement of the CLSS maps with hierarchical labels is striking, especially in light of its self-supervised training, which did not have access to this information [41]. CLSS maps from the sequence, structure, and sub-sequence modalities capture documented similarities of sequence and structure. Cases where hierarchical classifications and CLSS appear to disagree seem to highlight potential events of structural exploration: connections between different folds and possible examples of protein universe-spanning structure divergence. Although these types of relationships are not necessarily needed for protein structure prediction *per se* [53-55], they provide clues into the nature of protein sequence-structure exploration. In short, CLSS offers a new, AI-based, perspective on the continuous, malleable nature of protein sequences and structures, a long-sought scientific challenge [38]. When paired with studies of cross-fold fragment sharing and the observation of metamorphic protein folds, our data support the view that the protein universe, despite its staggering diversity, is readily traversable over long timescales.

Looking forward, we anticipate that CLSS, or models like it, will be a valuable resource for molecular biologists. CLSS maps can provide new contextualizing information – a view of protein similarity that encompasses aspects of sequence and structure. Ultimately, a unified sequence-structure embedding space will create new possibilities for database searching, sequence alignment, protein engineering, and evolutionary trajectory analysis.

Methods

CLSS Architecture

Figure S1 illustrates the architecture of CLSS. Our contrastive model has a two-tower architecture: The sequence tower is a trainable ESM2-like model with 35M parameters [26]; the structure tower is a frozen ESM3 [2] model with 1.4B parameters. Given a protein (or domain) with N residues, the sequence and structure towers infer a matrix of size $N \times K$, where K is a model-dependent constant (480 for ESM2 and 1536 for ESM3). Taking the mean over the matrix columns for each input results in vectors of length 480 and 1536 for ESM2 and ESM3, respectively. Finally, an adapter (a single linear layer) of size $K \times 32$ is appended at the end of each model, followed by L2 normalization. For every sequence-structure pair, CLSS outputs two vectors of length 32.

Datasets

Our training/validation dataset comprises 1 million randomly sampled domains with replacement from the publicly accessible ECOD domain classification of AlphaFold 2 predicted structures [38]. CLSS uses ESM3's `ProteinChain` API and ignores domains that have PDB files the API failed to open. Following Machine Learning best practices, we randomly split the dataset into a training set with 950,000 domains, and a validation set with the remaining 50,000 domains. Dataset 1 comprises 31,696 ECOD domains taken from the 109 most-characterized protein evolutionary lineages (ECOD X-group), spanning 16 architectures. Dataset 2 comprises 9,899 CATH domains from Røgen's set of sequence-similar but topologically distinct domain pairs [34]. The CLSS viewer shows all domains in the ECOD-AF2 F100 dataset with a valid PDB file (approximately 1.81 million) except for 19 domains with missing/corrupt PDB files our download pipeline failed to fix. In cases where the provided PDB files were invalid, we applied the PDBFixer python library. In cases where the provided PDB files were empty, we used the PDB file from the ECOD-PDB dataset instead, and if it was empty as well, we used the relevant PDB file from RCSB and cut it manually according to the domain's start and end residues as listed on ECOD.

Training Process

In each training step, 1440 domains (illustrated with a batch of size 4) were randomly selected from the full training dataset. The sampled batch was then split into 8 minibatches of 180 domains. Each minibatch was assigned to a different GPU. From the sequence of each domain, we sampled a random subsequence (highlighted in **Figure 1** in light blue) and calculated its ESM2-like sequence-segment embedding and projection to a normalized vector of size 32 (annotated in **Figure 1** as $Se_1, \dots, Se_4$). We then calculated the ESM3 structure embedding for the full domain and its projection to a normalized vectors of size 32 (annotated as $St_1, \dots, St_4$). Our contrastive loss function maximized the dot product between the structure and sequence segment embeddings from the same domain (marked in green), while minimizing it between different domains (marked in gray). Finally, we applied the `SoftMax` function to each row of the matrix, and computed the cross-entropy loss on the resulting distribution of each row, while labeling values on the diagonal as the "correct" class. Overall, we trained the

network for a fixed number of 80 epochs on 8 NVIDIA A100 (40GB) GPUs using the PyTorch Lightning framework and `DDPstrategy` for approximately 4.5 days.

The training process updated all parameters of ESM2-like model and the adapter networks of both models, while keeping the parameters of the ESM3 network frozen. This design implemented our goal of injecting structure information into the sequence model. Overall, the number of trained parameters in our network was only ~36M, slightly more than the small ESM2 (35M) model, and dramatically less than ProstT5 (3B), ESM3 (1.4B), and ProTrek (930M).

Ablation Studies and Hyperparameter Search

Several ablation studies and hyperparameter searches were performed. We trained the same model architecture with adapters of varying sizes: 16, 32, and 64. We also varied the minimum length of the random sub-sequences sampled from each full domain sequence (considering 1, 10, 15, 20, 25, and 30 residues), as well as training using the full-length sequences only, without any sub-sequence sampling (note that this model was only trained for 42 epochs). In addition to testing multiple architectures, we also ablated our loss function: optimizing the temperature parameter during training (similar to CLIP), versus using a fixed temperature with multiple values (0.25, 0.5, 0.75, 1, and 1.25); calculating the loss on both the rows and columns of our similarity matrix, or just on the rows; and computing the loss globally, on entire batches, compared to locally on each GPU.

To evaluate the trained models, we used the distributions of distances between each domain sequence, structure, and sub-sequence embeddings for avalidation set of 50,000 ECOD domains. We find that increasing the size of the adapter networks yields higher distances between the different modalities (**Figure S12A**), diminishing performance, while decreasing it results in a slight improvement over our chosen architecture (**Figure S12B**). **Figure S13** shows that, generally, training on shorter sequence segments results in slightly higher distances between the sequence and structure modalities, but significantly lower distances between the sequence segment modality to both the sequence and structure modalities. This holds true for all tested models, except for the model trained on segments of any length (minimum length of 1 residue) (**Figure S13 E**), which performs worse than our chosen model, which was trained on segments with 10 or more residues. This exception suggests that sequences with less than 10 residues do not contain enough information to associate them with their parent domain structures. Conversely, training on full sequences (**Figure S13F**) yields vastly lower distances between the sequence and structure modalities than our chosen architecture, but has the opposite effect with respect to the sub-sequence modality, for which distance distributions to both the sequence and structure modalities are shifted significantly towards greater values. This demonstrates the significance of training on sequence segments to the success of CLSS.

Interestingly, optimizing the temperature during training yields worse results than our chosen architecture, which uses a fixed temperature of 0.5 (**Figure S14A**). Increasing the temperature to 1 generally results in better performance, while further increasing it does not have a significant effect (**Figure S14**, panels B-E). The models trained with a fixed temperature of 1 and 1.25 showed the best results, and are slightly better than our chosen model. Lastly, training with global and local loss yielded similar results (**Figure S15A**), and so too did computing the loss on both the rows and columns compared to only considering the rows (**Figure S15 B**).

Embedding Visualizations

To create the t-SNE plots in **Figures 2**, **3**, **5**, **S2-S4**, and **S7-S10** and in the CLSS viewer, we used 1000 iterations and a perplexity value of 30. We visualized the probability density function of the distances between embeddings using the Gaussian kernel density estimator (gaussian_kde routines) from the scipy.stats package.

Code Availability

Our codebase (including training code) and the model weights are available on our GitHub (https://github.com/guyyanai/CLSS).

Acknowledgements

This work was supported by Israel Science Foundation (ISF) grant 1764/21, and by the Minerva Foundation. N.B.T. is supported by the Abraham E. Kazan Chair in Structural Biology, Tel Aviv University. G.A. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. L.M.L acknowledges support from Human Frontiers Science Program (HFSP) grant RGEC29/2025.

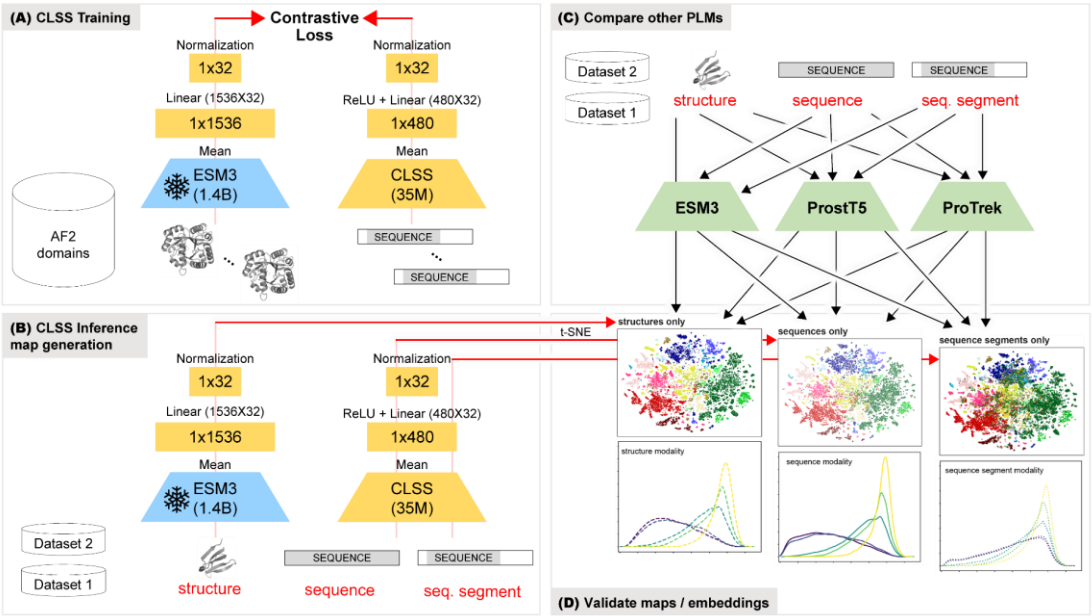

Figure 1: Overview of training, validating, and testing of CLSS to create unified maps of protein sequence and structure space. (A) Overview of the two-tower CLSS architecture. On the left is a structure tower based on the frozen, pre-trained ESM3 model (light blue) followed by a trained (yellow) adapter that averages, reduces dimension, and normalizes the embedding. On the right is the trained CLSS sequence tower, build upon a pre-trained ESM2 model, and its adapter network (yellow). The networks are trained using contrastive loss on batches of randomly chosen structures and sequence segments from the ECOD-AF2 domain database. Once trained, we calculate the embeddings of the structures, sequences, and sequence segments from Datasets 1 and 2 using CLSS (B) and other PLMs (C). (D) Dimensionality reduction by t-SNE was used to create visual maps of protein space (upper images). Pairwise distance distributions (lower images) were calculated from the embeddings directly, rather than the t-SNE reduced space.

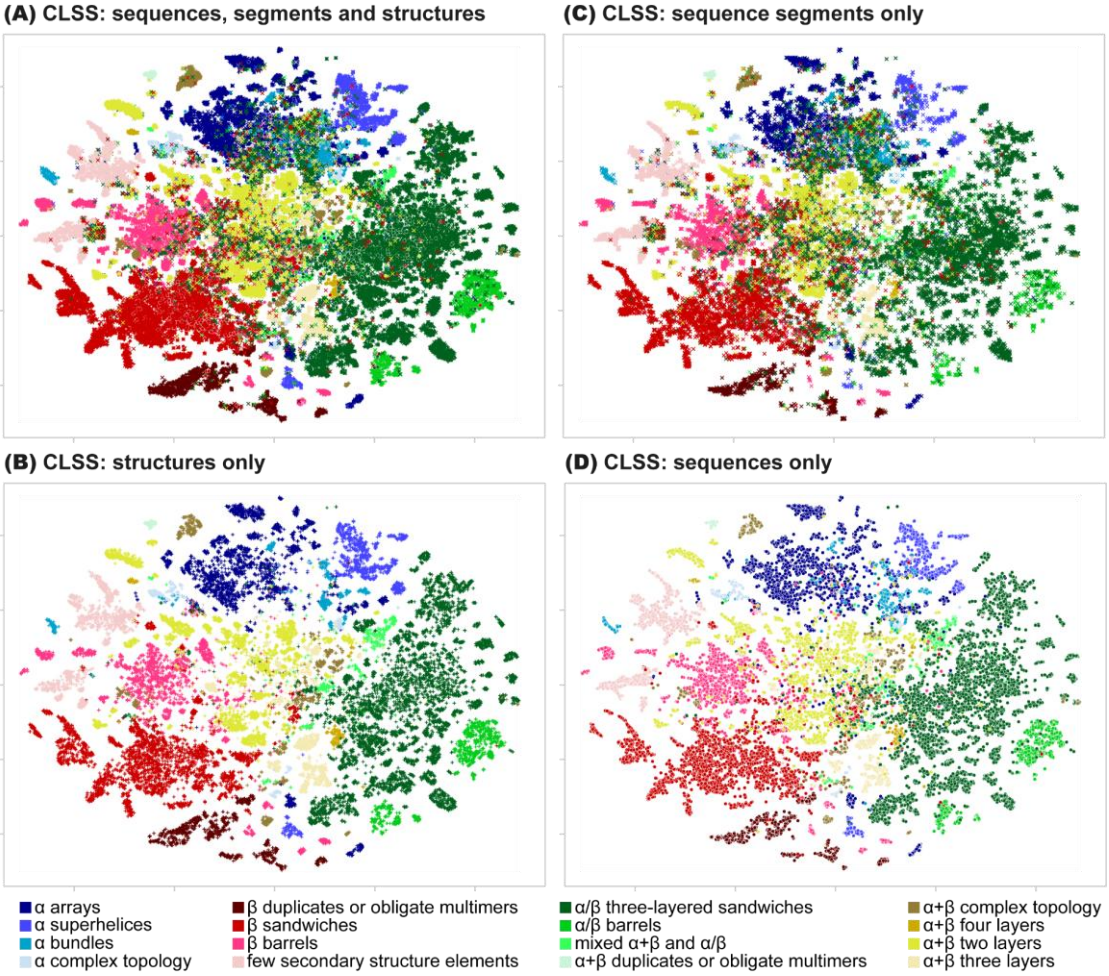

Figure 2: CLSS embedding maps of ECOD domains (Dataset 1). For each domain, we calculate the embeddings by three modalities – structure, sequence, and a random sequence segment – and then compute a t-SNE projection of the embeddings. Each point represents one of the modalities of a domain colored according to the label of its ECOD architecture. (A) An overlay of all three modalities. Sequences are marked by circles, structures by '+', and random sequence segments by 'x'. (B) Structure embeddings. (C) sequence segment embeddings. (D) Sequence embeddings. We find that the maps of all three modalities are very similar to each other, with the sequence (D) and structure (B) embeddings being the closest. This shows that CLSS successfully injected structure information into the sequence modality. The global organization of the CLSS embedding space positions domains with the same ECOD architecture, and even the same structure class, near each other.

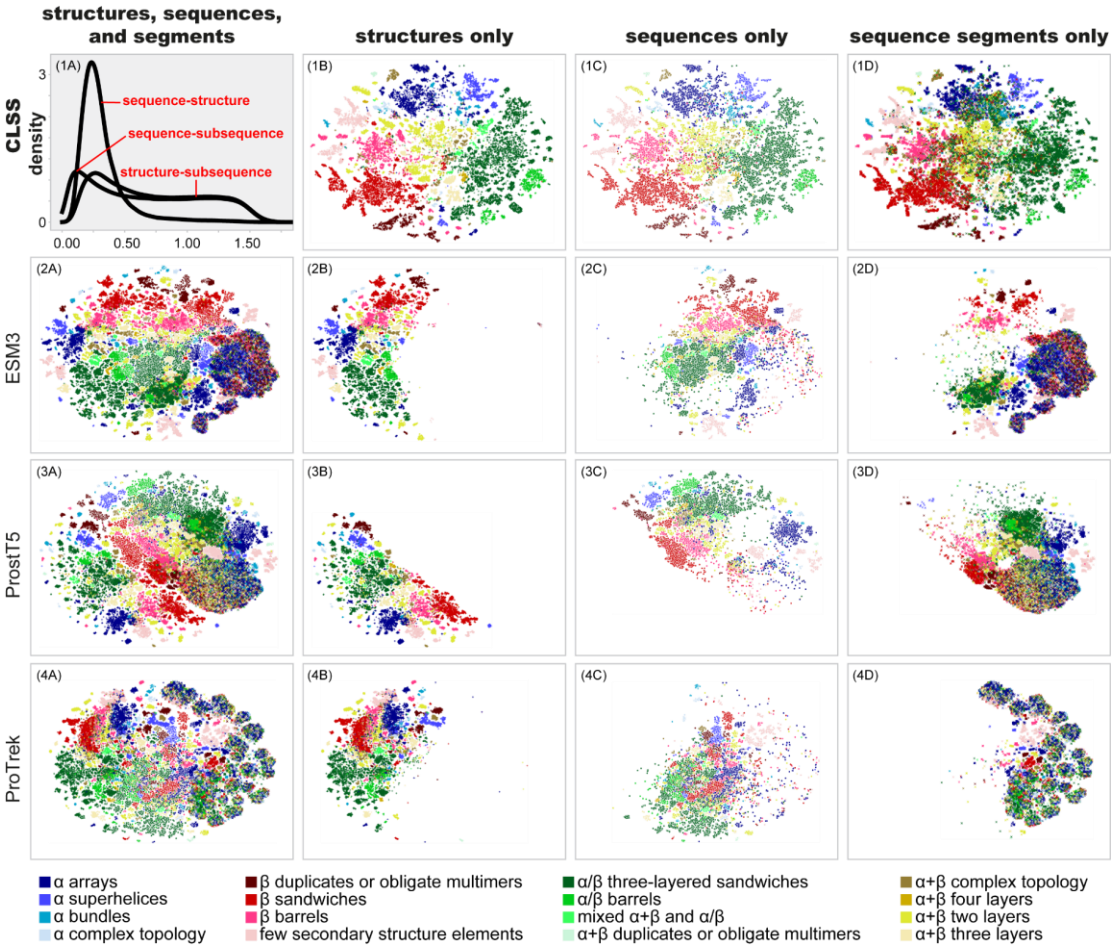

Figure 3: Comparison of maps generated for embeddings of ECOD domains (Dataset 1) with the PLMs (from top to bottom): CLSS, ESM3, ProstT5, and ProTrek. In each row, the maps (left to right) are an overlay of all three modalities, followed by the single modalities of structure, sequence, and sequence segments. The top-left panel shows the distribution of distances between CLSS embeddings of the same domain, in different modalities, demonstrating that the sequence and structure embeddings are indeed near each other; ***Figure S6*** *shows a breakdown of this data by ECOD architecture. We find that in ESM3, ProstT5, and ProTrek, the embeddings of different modalities occupy distinct regions of the map. The structure modality embeddings are closer to the sequence modality embeddings, which in turn are closer to the embeddings of the sub-sequence modality. Embeddings from the structure modality generally place domains of the same architecture, and even structure class close to each other; the sequence modality performs less well (see supplementary* ***Figures S2-S4*** *for results for each PLM).*

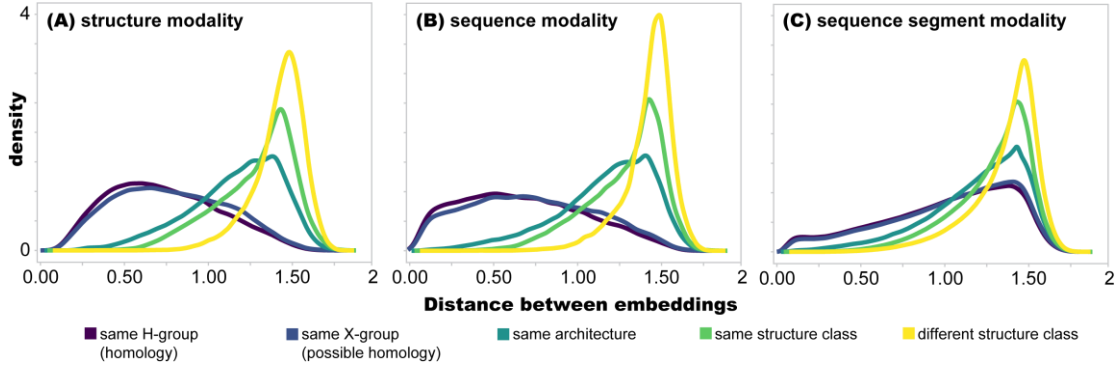

Figure 4: Estimated distributions of distances between pairs of embeddings for domains in Dataset 1. (A-C) The distributions of distances of domain pairs with the same ECOD label at different levels of the hierarchy. The color palette indicates the shared level in the ECOD hierarchy, from pairs with the same H-group label in purple through pairs with different structure class in yellow. In all three modalities, the distances follow the ECOD hierarchy, with the domain pairs from the same H-group being closest. Thus, CLSS embeddings for all three modalities agree with the human-curated labels of ECOD. Notice that the distance distributions of the H- and X-group pairs for the sequence modality (panel B) are slightly shifted to the left compared to the structure modality (panel A), implying that there are pairs within these groups whose sequence embeddings are closer to each other than their structure embeddings.

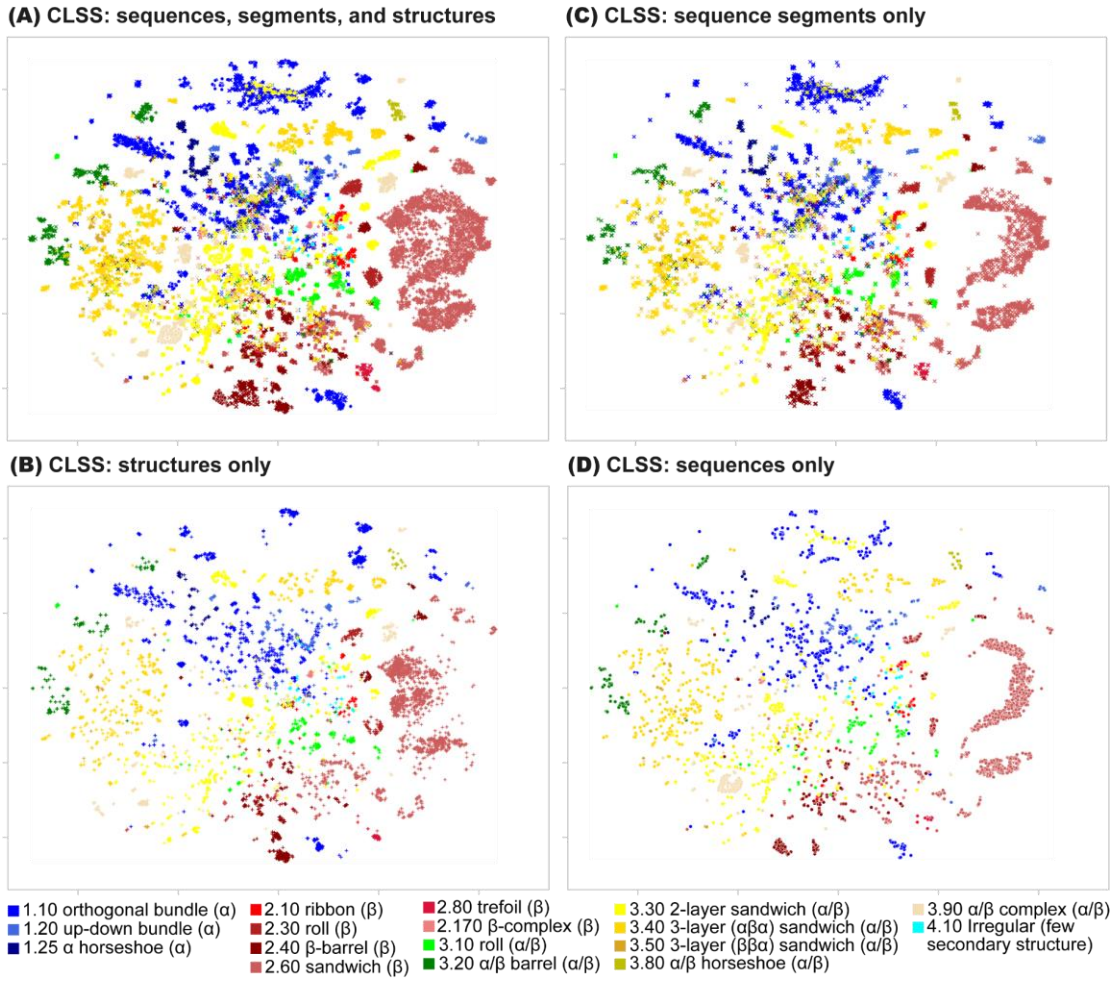

Figure 5: Maps based on CLSS embeddings of metamorphic CATH domains, sharing high sequence similarity but topologically different structures (Dataset 2). For each domain, we consider three modalities: sequence, structure, and a random sequence segment and compute a t-SNE projection of the embeddings. Each point represents one of the modalities of one domain colored according to the label of its CATH architecture. (A) All three modalities. Sequences are marked by circles, structures by '+', and random sequence segments by 'x'. (B) Structure embeddings. (C) sequence segment embeddings. (D) Sequence embeddings.

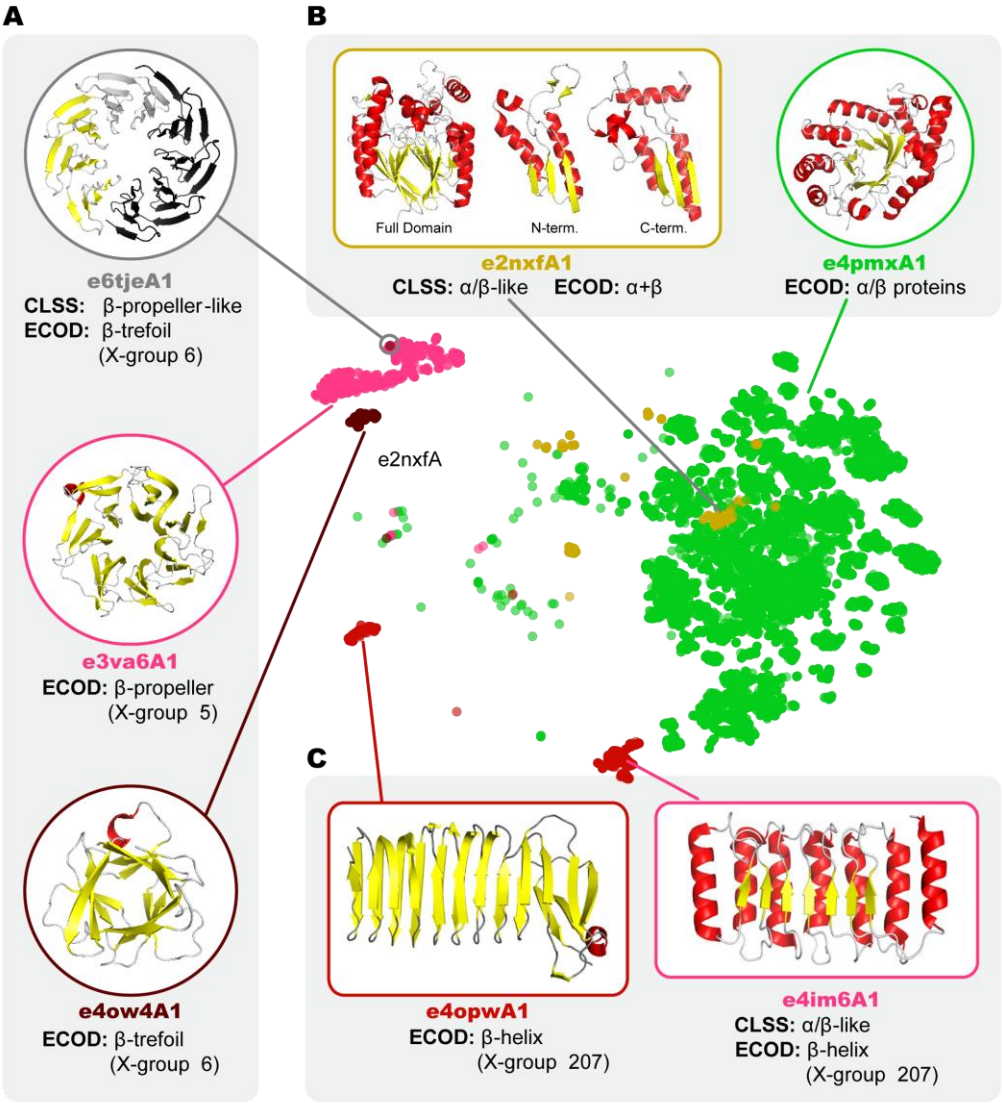

Figure 6 Example comparisons between the ECOD hierarchy and CLSS maps. (A) The challenge of automatic classification of a designed protein with sequence based HMMs. e6tjeA1 is part of a designed β-propeller with β-trefoil-like sequence characteristics (B) The localization of some α+β 4-layer sandwiches (ochre) with α/β-folds (green). The CLSS classification reflects the β-α-β-α-β elements that are contained within their structures. (C) The splitting of X-group 207, which is classified by ECOD as having a β-architecture. Domains positioned near the α/β region (green) have clear signatures of α/β-structure. These examples highlight the limitations of hierarchical classifications and present three candidate events of structural evolution within and between protein X-groups.

References

1. Holm L, Sander C: **Mapping the protein universe**. *Science* 1996, **273**:595-603.
2. Hou J, Sims GE, Zhang C, Kim SH: **A global representation of the protein fold space**. *Proc Natl Acad Sci U S A* 2003, **100**:2386-2390.
3. Nepomnyachiy S, Ben-Tal N, Kolodny R: **Global view of the protein universe**. *Proceedings of the National Academy of Sciences* 2014, 10.1073/pnas.1403395111.
4. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ: **Pfam: The protein families database in 2021**. *Nucleic acids research* 2021, **49**:D412-D419.
5. Lo Conte L, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C: **SCOP: a Structural Classification of Proteins database**. *Nucl. Acids Res.* 2000, **28**:257-259.
6. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH--a hierarchic classification of protein domain structures**. *Structure* 1997, **5**:1093-1108.
7. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim B-H, Grishin NV: **ECOD: An Evolutionary Classification of Protein Domains**. *PLoS Comput Biol* 2014, **10**:e1003926.
8. Edwards H, Deane CM: **Structural bridges through fold space**. *PLoS computational biology* 2015, **11**:e1004466.
9. Alva V, Remmert M, Biegert A, Lupas AN, Söding J: **A galaxy of folds**. *Protein Science* 2010, **19**:124-130.
10. Ben-Tal N, Kolodny R: **Representation of the Protein Universe using Classifications, Maps, and Networks**. *Israel Journal of Chemistry* 2014.
11. Osadchy M, Kolodny R: **Maps of protein structure space reveal a fundamental relationship between protein structure and function**. *Proceedings of the National Academy of Sciences* 2011, **108**:12301-12306.
12. Rogen P, Fain B: **Automatic classification of protein structure by using Gauss integrals**. *Proc Natl Acad Sci U S A* 2003, **100**:119-124.
13. Han X, Sit A, Christoffer C, Chen S, Kihara D: **A global map of the protein shape universe**. *PLOS Computational Biology* 2019, **15**:e1006969.
14. Budowski-Tal I, Nov Y, Kolodny R: **FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately**. *Proc Natl Acad Sci U S A* 2010, **107**:3481-3486.
15. Liu Y, Ye Q, Wang L, Peng J: **Learning structural motif representations for efficient protein structure search**. *Bioinformatics* 2018, **34**:i773-i780.
16. Hie BL, Yang KK, Kim PS: **Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins**. *Cell Syst* 2022, **13**:274-285.e276.
17. Heinzinger M, Rost B: **Teaching AI to speak protein**. *Current Opinion in Structural Biology* 2025, **91**:102986.
18. Bjerregaard A, Groth PM, Hauberg S, Krogh A, Boomsma W: **Foundation models of protein sequences: A brief overview**. *Current Opinion in Structural Biology* 2025, **91**:103004.
19. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M: **Simulating 500 million years of evolution with a language model**. *Science* 2025:eads0018.
20. Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F: **Saprot: Protein language modeling with structure-aware vocabulary. bioRxiv**. Edited by: Cold Spring Harbor Laboratory; 2023.

21. Su J, Zhou X, Zhang X, Yuan F: **ProTrek: Illuminating the Protein Universe with Tri-Modal**. *structure* **1**:f1.
22. Flöge K, Udayakumar S, Sommer J, Piraud M, Kesselheim S, Fortuin V, Günneman S, van der Weg KJ, Gohlke H, Bazarova A: **OneProt: Towards Multi-Modal Protein Foundation Models**. *arXiv preprint arXiv:2411.04863* 2024.
23. Wang D, Pourmirzaei M, Abbas UL, Zeng S, Manshour N, Esmaili F, Poudel B, Jiang Y, Shao Q, Chen J: **S-PLM: Structure-Aware Protein Language Model via Contrastive Learning Between Sequence and Structure**. *Advanced Science* 2025, **12**:2404212.
24. Nepomnyachiy S, Ben-Tal N, Kolodny R: **Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths**. *Proceedings of the National Academy of Sciences* 2017, **114**:11703-11708.
25. Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N: **Bridging Themes: Short Protein Segments Found in Different Architectures**. *Molecular Biology and Evolution* 2021, **38**:2191-2208.
26. Qiu K, Ben-Tal N, Kolodny R: **Similar protein segments shared between domains of different evolutionary lineages**. 2022, **31**:e4407.
27. Longo LM, Jabłońska J, Vyas P, Kanade M, Kolodny R, Ben-Tal N, Tawfik DS: **On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment**. *Elife* 2020, **9**:e64415.
28. Kessel A, Ben-Tal N: *Introduction to proteins: structure, function, and motion* edn 2: CRC Press; 2018.
29. Alva V, Söding J, Lupas AN: **A vocabulary of ancient peptides at the origin of folded proteins**. *eLife* 2015, **4**:e09410.
30. Tan Y, Li M, Zhou B, Zhong B, Zheng L, Tan P, Zhou Z, Yu H, Fan G, Hong L: **Simple, efficient, and scalable structure-aware adapter boosts protein language models**. *Journal of Chemical Information and Modeling* 2024, **64**:6338-6349.
31. Zhang Z, Wang C, Xu M, Chenthamarakshan V, Lozano A, Das P, Tang J: **A systematic study of joint representation learning on protein sequences and structures**. *arXiv preprint arXiv:2303.06275* 2023.
32. Chen D, Hartout P, Pellizzoni P, Oliver C, Borgwardt K: **Endowing protein language models with structural knowledge**. *arXiv preprint arXiv:2401.14819* 2024.
33. Kosloff M, Kolodny R: **Sequence-similar, structure-dissimilar protein pairs in the PDB**. *Proteins* 2008, **71**:891-902.
34. Røgen P: **Sequence-Similar Protein Domain Pairs With Structural or Topological Dissimilarity**. *Proteins: Structure, Function, and Bioinformatics* 2024.
35. Narunsky A, Nepomnyachiy S, Ashkenazy H, Kolodny R, Ben-Tal N: **ConTemplate Suggests Possible Alternative Conformations for a Query Protein of Known Structure**. *Structure* 2015, **23**:2162-2170.
36. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J: **Learning transferable visual models from natural language supervision**. In *International conference on machine learning*: PMLR: 2021:8748-8763.
37. Zhai X, Wang X, Mustafa B, Steiner A, Keysers D, Kolesnikov A, Beyer L: **Lit: Zero-shot transfer with locked-image text tuning**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*: 2022:18123-18133.
38. Schaeffer RD, Zhang J, Medvedev KE, Kinch LN, Cong Q, Grishin NV: **ECOD domain classification of 48 whole proteomes from AlphaFold Structure Database using DPAM2**. *PLoS Computational Biology* 2024, **20**:e1011586.

39. Van der Maaten L, Hinton G: **Visualizing data using t-SNE**. *Journal of machine learning research* 2008, **9**.
40. Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Mirdita M, Steinegger M, Rost B: **Bilingual language model for protein sequence and structure**. *NAR Genomics and Bioinformatics* 2024, **6**:lqae150.
41. Zhang Z, Lu J, Chenthamarakshan V, Lozano A, Das P, Tang J: **Structure-informed protein language model**. *arXiv preprint arXiv:2402.05856* 2024.
42. Tenorio CA, Longo LM, Parker JB, Lee J, Blaber M: **Ab initio folding of a trefoil-fold motif reveals structural similarity with a β-propeller blade motif**. *Protein Science* 2020, **29**:1172-1185.
43. Longo LM, Kolodny R, McGlynn SE: **Evidence for the emergence of β-trefoils by 'Peptide Budding'from an IgG-like β-sandwich**. *PLoS computational biology* 2022, **18**:e1009833.
44. Mylemans B, Laier I, Kamata K, Akashi S, Noguchi H, Tame JRH, Voet ARD: **Structural plasticity of a designer protein sheds light on β-propeller protein evolution**. *The FEBS Journal* 2021, **288**:530-545.
45. Lu AX, Yan W, Yang KK, Gligorijevic V, Cho K, Abbeel P, Bonneau R, Frey N: **Tokenized and Continuous Embedding Compressions of Protein Sequence and Structure**. *bioRxiv* 2024, 10.1101/2024.08.06.606920:2024.2008.2006.606920.
46. **Orthogonal Range Searching**. In *Computational Geometry: Algorithms and Applications*. Edited by de Berg M, Cheong O, van Kreveld M, Overmars M: Springer Berlin Heidelberg; 2008:95-120. 10.1007/978-3-540-77974-2_5
47. Geman S, Bienenstock E, Doursat R: **Neural networks and the bias/variance dilemma**. *Neural computation* 1992, **4**:1-58.
48. Mackenzie CO, Grigoryan G: **Protein structural motifs in prediction and design**. *Current opinion in structural biology* 2017, **44**:161-167.
49. Zhou J, Panaitiu AE, Grigoryan G: **A general-purpose protein design framework based on mining sequence–structure relationships in known protein structures**. *Proceedings of the National Academy of Sciences* 2020, **117**:1059-1068.
50. Khersonsky O, Fleishman SJ: **Why reinvent the wheel? Building new proteins based on ready-made parts**. *Protein science : a publication of the Protein Society* 2016, **25**:1179-1187.
51. Lipsh-Sokolik R, Khersonsky O, Schröder SP, de Boer C, Hoch S-Y, Davies GJ, Overkleeft HS, Fleishman SJ: **Combinatorial assembly and design of enzymes**. *Science* 2023, **379**:195-201.
52. Thiele R, Zimmer R, Lengauer T: **Protein threading by recursive dynamic programming**. *Journal of Molecular Biology* 1999, **290**:757-779.
53. Bowie J, Luthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure**. *Science* 1991, **253**:164-170.
54. Jones DT, Taylort WR, Thornton JM: **A new approach to protein fold recognition**. *Nature* 1992, **358**:86-89.
55. Zhang Y: **Progress and challenges in protein structure prediction**. *Current Opinion in Structural Biology* 2008, **18**:342-348.
56. Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J: **CATH-a hierarchic classification of protein domain structures**. *Structure* 1997, **5**:1093 - 1108.
57. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536-540.
58. Zhang Z, Xu M, Jamasb A, Chenthamarakshan V, Lozano A, Das P, Tang J: **Protein representation learning by geometric structure pretraining**. *arXiv preprint arXiv:2203.06125* 2022.