COLLAPSE: A representation learning framework for identification and characterization of protein structural sites

Alexander Derry[1,*] & Russ B. Altman[1,2,*]

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA
[2]Departments of Bioengineering, Genetics, and Medicine, Stanford University, Stanford, CA

*Correspondence: aderry@stanford.edu, russ.altman@stanford.edu

Abstract

The identification and characterization of the structural sites which contribute to protein function are crucial for understanding biological mechanisms, evaluating disease risk, and developing targeted therapies. However, the quantity of known protein structures is rapidly outpacing our ability to functionally annotate them. Existing methods for function prediction either do not operate on local sites, suffer from high false positive or false negative rates, or require large site-specific training datasets, necessitating the development of new computational methods for annotating functional sites at scale. We present COLLAPSE (Compressed Latents Learned from Aligned Protein Structural Environments), a framework for learning deep representations of protein sites. COLLAPSE operates directly on the 3D positions of atoms surrounding a site and uses evolutionary relationships between homologous proteins as a self-supervision signal, enabling learned embeddings to implicitly capture structure-function relationships within each site. Our representations generalize across disparate tasks in a transfer learning context, achieving state-of-the-art performance on standardized benchmarks (protein-protein interactions and mutation stability) and on the prediction of functional sites from the PROSITE database. We use COLLAPSE to search for similar sites across large protein datasets and to annotate proteins based on a database of known functional sites. These methods demonstrate that COLLAPSE is computationally efficient, tunable, and interpretable, providing a general-purpose platform for computational protein analysis.

Keywords: Deep learning, structural informatics, functional site annotation, representation learning, protein structure analysis

1. Introduction

The three-dimensional structure of a protein determines its functional characteristics and ability to interact with other molecules, including other proteins, endogenous small molecules, and therapeutic drugs. Biochemical interactions occur at specific regions of the protein known as functional sites. We consider functional sites that range from a few atoms which coordinate an ion or catalyze a reaction to larger regions which binds a cofactor or form a protein-protein interaction surface. The identification of such sites—and accurate modeling of the local structure-function relationship—is critical for determining a protein's biological role, including our understanding of disease pathogenesis and ability to develop targeted therapies or protein engineering technologies. Significant effort has gone into curating databases to catalog these structure-function relationships, [1–3] but this cannot keep up with the rapid increase in proteins in need of annotation. The number of proteins of the Protein Data Bank (PDB) [4] increases each year, and AlphaFold [5] has added high-quality predicted structures for hundreds of thousands more. This explosion of protein structure data necessitates the development of computational methods for identifying, characterizing, and comparing functional sites at proteome scale.

Many widely used methods for protein function identification are based on sequence. Sequence profiles and hidden Markov models built using homologous proteins [6–10] are often used to infer function by membership in a particular family, but these methods do not always identify specific functional residues and can misannotate proteins in mechanistically diverse families. [11] Additionally, structure and function are often conserved even when sequence similarity is very low, resulting in large numbers of false negatives for methods based on sequence alignment. [12,13] Approaches based on identifying conserved sequence motifs within families can help to address these issues. [14,15] However, these methods suffer from similar limitations as sequences diverge, resulting in high false positive and false negative rates, especially when the functional residues are far apart in sequence. [16] More generally, sequence-based methods cannot capture the complex 3D conformations and physicochemical interactions required to accurately define a functional site or inform opportunities to engineer or mutate specific residues.

Recently, methods have applied machine learning to predict function from sequence [17,18] or structure. [19] However, like profile-based methods, these lack the local resolution necessary to identify specific functional sites, and their reliance on non-specific functional labels such as those provided by Gene Ontology terms [20] often limits practical utility. [21] Machine learning approaches that focus on local functional sites are either specific to a particular type of site (e.g. ligand binding [22,23], enzyme active sites [24]) or require building specific models for each functional site of interest, [25,26] which can be computationally expensive and demands sufficient data to train an accurate model.

A major consideration for building generalizable machine learning models for protein sites is the choice of local structure representation. FEATURE, [27] a hand-crafted property-based representation, has shown utility for many functionally-relevant tasks. [25,28,29] However, FEATURE uses heterogeneous features (a mix of counts, binary, and continuous) which are

more difficult to train on and meaningfully compare in high dimensions. Additionally, FEATURE consists of radial features without considering orientation and does not account for interactions between atoms in 3D, leading to loss of information. [26] Deep learning presents an attractive alternative by enabling the extraction of features directly from raw data, [30] but the high complexity of deep learning models means that they require large amounts of labeled data. To address this, a paradigm has emerged in which models are pre-trained on very large unlabeled datasets to extract robust and generalizable features which can then be "transferred" to downstream tasks. [31,32] This approach has been successfully applied to learn representations of small molecules [33,34] and protein sequences, [17,35,36] but there are few examples of representations learned directly from 3D structure. Initial efforts focus on entire proteins rather than sites and operate only at residue-level resolution. [37,38]

We address these issues by developing COLLAPSE (Compressed Latents Learned from Aligned Protein Structural Environments), a framework for functional site characterization, identification, and comparison which (1) focuses on local structural sites, defined as all atoms within a 10 Å radius of a specific residue; (2) captures complex 3D interactions at atom resolution; (3) works with arbitrary sites, regardless of the number of known examples; and (4) enables comparison between sites across proteins. COLLAPSE combines self-supervised methods from computer vision, [39] graph neural networks designed for protein structure, [40,41] and multiple sequence alignments of homologous proteins to learn 512-dimensional protein site embeddings that capture structure-function relationships both within and between proteins.

Self-supervised representation learning refers to the procedure of training a model to extract high-level features from raw data using one or more "pretext tasks" defined using intrinsic characteristics of the input data. The choice of pretext task is critical to the utility of the learned representations. A popular class of methods involves minimizing the distance between the embeddings of two augmented versions of the same data point (for example, cropped and rotated views of the same image), thereby learning a representation that is robust to noise which is independent of the fundamental features of the original data. [39,42,43] Since function is largely conserved within a protein family even as sequences diverge, we draw an analogy between homologous proteins and augmented views of the same image. Specifically, we hypothesized that by pulling together the embeddings of corresponding sites in homologous proteins, we could train the model to learn features which capture the site's structural and functional role. In this scheme, sequence alignments are used to identify correspondences between amino acids, which are then mapped to 3D structures to define the structural site surrounding each residue (Figure 1, Section 2.2).

Pre-trained representations are typically used in one of two settings: (1) transfer learning, which leverages general representations to improve performance on problem-specific supervised tasks where access to labeled data is limited; and (2) extracting insights about the underlying data from the learned embedding space directly (e.g. via visualization or embedding comparisons). [44] In this paper, we illustrate the utility of COLLAPSE protein site in both settings. First, we demonstrate that COLLAPSE generalizes in a transfer learning setting, achieving

competitive or best-in-class results across a range of downstream tasks. Second, we describe two applications that demonstrate the power of our embeddings for protein function analysis without the need to train any downstream models: an iterated search procedure for identifying similar functional sites across large protein databases, and a method for efficiently annotating putative functional sites in an unlabeled protein. All datasets, models, functionality, and source code can be found in our Github repository (https://github.com/awfderry/COLLAPSE).

2. Results

2.1. Intrinsic evaluation of COLLAPSE embeddings

To evaluate the extent to which COLLAPSE embeddings capture relevant structural and functional features, we embedded the environments of all residues in a held-out set consisting of proteins with varying levels of sequence similarity to proteins in the training set. First, we find that the degree of similarity between embeddings of aligned sites is correlated with the level of conservation of that site in the original MSA (Fig. 2a). Even at less than 30% conservation, aligned sites are significantly more similar on average than a randomly sampled background of non-aligned sites ($p < 1 \times 10^{-15}$).

We also confirmed that our embeddings capture local information at a residue-level resolution, meaning that neighboring environments can be effectively distinguished from each other. Indeed, the normalized cosine similarity between residue embeddings decreases between the residues in sequence increases (Fig. 2b). This effect generalizes even to proteins far away from the training set in sequence identity. Finally, among chains with a single fold according to CATH 4.2 [45] ($n = 11{,}270$), the top-level structural class can be distinguished clearly in protein-level embeddings, suggesting that secondary structure is a major feature captured by COLLAPSE (Fig. 2c). Lower levels of the CATH hierarchy also cluster clearly in low-dimensional space (Fig. S1).

2.2. Generalization across ATOM3D benchmarks

To assess the utility of COLLAPSE embeddings in a transfer learning context, we use ATOM3D, a suite of benchmarking tasks and datasets for machine learning in structural biology. [46] We selected two tasks from ATOM3D which focus on protein sites: protein interface prediction (PIP) and mutation stability prediction (MSP). We evaluate performance compared to the ATOM3D reference models and to the task-specific GVP-GNN reported in Jing et al. (2021), [41] which is state-of-the-art for all tasks. Table 1 reports the results both with and without task-specific fine-tuning of the embedding model parameters. Without fine-tuning, COLLAPSE embeddings and a simple classifier achieve results comparable or better than the ATOM3D reference models trained specifically for each task. Fine-tuning improves performance further, achieving state-of-the-art on PIP and comparable performance to the GVP-GNN on MSP.

2.3. Functional site prediction models

COLLAPSE embeddings can also be used to build high-precision functional site prediction models. We train prediction models for 10 functional sites defined by the PROSITE database, [14] which identifies local sites using curated sequence motifs. On sites labeled true positive (TP) by PROSITE, COLLAPSE outperforms the analogous FEATURE models and perform comparably or better than task-specific 3DCNN models trained end-to-end, achieving greater than 86% recall on all sites at a threshold of 99% precision. PROSITE also provides false negatives (FNs; true proteins which are not recognized by the PROSITE pattern) and false positives (FPs; proteins which match the PROSITE pattern but are not members of the functional family). Table 2 summarizes the number of correct predictions at the protein level relative to the total number identified by PROSITE. For all families, COLLAPSE correctly identifies a greater or equal number of FN proteins compared to FEATURE and 3DCNN classifiers. The improvement is notable in some cases, such as a 162.5% increase in proteins detected for IG_MHC, a 37.5% increase for ADH_SHORT, and a 17.6% increase for EF_HAND_1. For four of the seven proteins with FP data, we correctly rule out all FPs. For ADH_SHORT and EF_HAND_1, we perform 9.1% and 4.0% worse relative to 3DCNN, respectively, but this slight increase in FPs is not substantial relative to the improvement in FNs recovered for these families.

2.4. Iterative search for functional sites across protein databases

While COLLAPSE embeddings can be used to train highly accurate models for functional site detection, we can only train such models for those functional sites for which we have sufficient training examples. Another way to understand the possible function of a site is to analyze similar sites retrieved from a structure database. The set of hits retrieved by this search may contain known functional annotations or other information which sheds light on the query site. We use iterative COLLAPSE embedding comparisons to perform such a search across the PDB. We investigate the performance of this method on the PROSITE dataset while varying two parameters: the number of iterations and the p-value cutoff for inclusion at each iteration. The method generally achieves high recall and precision after 2–5 iterations at a p-value cutoff of 5×10^{-3} to 5×10^{-4} (Fig. 4; Fig. S4). Notably, when evaluating on the FN and FP subsets, our search method even outperforms the cross-validated models on some sites (e.g. IG_MHC, Fig. 4a). However, the precision and recall characteristics vary widely across families; in some cases it predicts the same set of proteins as the trained model (e.g. TRYPSIN_HIS; Fig. 4b), while in others it performs worse (e.g. EF_HAND_1; Fig. 4c). Importantly, the method requires no training and is very efficient: runtime per iteration scales linearly with the size of the query set and with database size (Fig. S5).

2.5. Protein structure annotation

Our iterative search method assumes that a site of interest has already been identified. However, when a new protein is discovered and its structure is solved, the locations of functional sites are often unknown. By comparing local environments in the protein's structure to those contained in databases of known functional sites, we can predict which sites are likely to be functional. Figure 5 shows two example annotations using a modified mutual best hit criterion against a reference database consisting of embeddings from PROSITE and the Catalytic

Site Atlas (CSA). First, we show the structure of meizothrombin, a precursor to thrombin and a trypsin-like serine protease with a canonical His-Asp-Ser catalytic triad. Our method correctly identifies all three residues as belonging to the trypsin-like serine protease family in PROSITE (Fig. 5a). Hits against the CSA, which are more specific, also include closely homologous proteins such as C3/C5 convertase. The associated kringle domain is also identified by its characteristic disulfide bond. Second, we show the structure of beta-glucuronidase (Fig. 5b), a validation set protein which has no homologs in the training set. We correctly identify all four catalytic residues defined by the CSA (in yellow), as well as PROSITE signatures corresponding to the glycosyl hydrolases family 2, the family which contains beta-glucuronidase.

3. Discussion

The utility of COLLAPSE embeddings for functional analysis derives from several key features of the training algorithm. First, the use of homology as a source of self-supervision signal allows the model to learn patterns of structural conservation across proteins, imbuing the model with a biological inductive bias towards features that may be important to the protein's function. Such patterns could in theory be learned by a model which sees each protein independently, but it would require much more data and training time to identify subtle signals across disparate proteins. While evolutionary data has proved crucial to the success of sequence-based models, to our knowledge this is the first time sequence alignments have been used to direct the training of a structural model. Second, by focusing on local protein sites, our embeddings are more precise and flexible than models which aim to represent an entire protein. COLLAPSE embeddings can be used for arbitrary tasks on the level of single residues or even individual functional atoms, to detect important regions in proteins, and to identify functional relationships between proteins even if they are divergent in sequence or global fold. Moreover, by aggregating over multiple residues or entire proteins, site-specific embeddings can also be applied to domain-level or full-protein tasks. Finally, by using an atomic graph representation and a GVP-GNN encoder, COLLAPSE captures all inter-atomic interactions (in contrast to methods which operate at a residue level) and produces representations that are fully equivariant to 3D rotation and translation.

As input to machine learning models, COLLAPSE embeddings generalize across tasks that require the model to learn different aspects of the protein structure-function relationship, including identifying protein-protein interactions, predicting stabilizing mutations, and classifying functional sites. On the PROSITE dataset of functional sites, we significantly outperform FEATURE, the closest analog to COLLAPSE as a protein structural site representation. We expect that substituting for COLLAPSE embeddings in other applications addressed by the FEATURE suite of methods [28,29,47] will also lead to improved performance.

Pre-trained COLLAPSE embeddings also perform better than or comparable to end-to-end 3DCNN models despite the use of a much simpler SVM classifier, demonstrating the effectiveness of the transfer learning paradigm. Additionally, we achieve higher sensitivity for detecting PROSITE false negatives than both FEATURE and 3DCNN baselines, regardless of which

baseline performs better on each site. This result suggests that by using our embeddings as input to machine learning tasks, we strike a balance whereby the models are robust to noise through the use of fixed embeddings while still capturing complex physicochemical features through the deep learning–based pre-training process.

One of the most important aspects of COLLAPSE which sets it apart from task-specific machine learning tools is the ability to derive insights from the embedding space itself, without fitting any models. In these cases, the embedding distance provides a functionally relevant distance measure for comparing functional sites. We demonstrate this with our functional site search and annotation methods, both of which rely only on direct comparisons in the embedding space. Both are efficient, generalizable, and offer significance estimates which allow a user to tune the sensitivity and specificity of the results. For example, for discovery applications it may be desirable to optimize for sensitivity at the cost of more false positives, while prioritizing drug targets for experimental validation may require greater specificity.

The choice of background distribution for computing empirical p-values is critical for accurate tuning of the significance threshold. For most general-purpose tasks, the non-redundant subset of the PDB used here is sufficient, but more specific applications may benefit from a different choice to improve statistical power (e.g. a distribution computed only from embeddings of a single residue type). We also note that the range of cosine similarities is relatively small (~0.9–1.0) even for the background distribution, which we attribute to the use of mean pooling over atoms in the encoder's graph aggregation step. A different choice of aggregation function may produce larger dynamic range across the embeddings. However, when normalized the comparisons are robust and locally specific at a resolution of one residue: less than 3% of neighboring environments would be considered significantly similar at $p = 1 \times 10^{-4}$.

The ability of iterative nearest-neighbor searches in the embedding space to identify known sites in PROSITE demonstrates that functional sites cluster meaningfully in the embedding space. The effect of changing input parameters (number of iterations and p-value cutoff) on the sensitivity and specificity of the results varies somewhat across functional families. In some cases (notably IG_MHC), this method achieves better sensitivity for FNs than even machine learning models trained using CV, while in others (EF_HAND_1, PROTEIN_KINASE_TYR) it cannot achieve this without a significant drop in precision. This is likely due to differences in structural conservation between sites, whereby sites which are more structurally heterogeneous are more difficult to fully capture using a query-based approach than a trained model which can learn to recognize diverse structural patterns. However, since training an accurate model requires access to a representative training dataset, which is not always available, we consider our search method to be a powerful complement to training site-specific models for the purpose of functional site identification when labeled data is scarce. We also note that while many methods enable structural search for full proteins [48,49] or binding sites, [29,50,51] ours is the first search tool specifically designed for arbitrary local structural sites.

Functional annotation of novel protein structures is of great value to the structural biology and biochemistry communities, but there are few tools for doing so at the residue level. COLLAPSE

provides a method for residue-level annotation which is efficient and tunable, making it suitable for both screening and discovery purposes. As shown by the examples in Fig. 5, the method identifies known functional annotations while limiting false positives to closely related homologs, even when the input is not related to any protein in the training set (<5% sequence identity for beta-glucuronidase). Importantly, all predictions can be explained and cross-referenced by rich metadata from the reference data sources, enhancing trust and usability. Of the PDBs returned for true positive sites in meizothrombin and beta-glucuronidase, 45.5% (20/44) and 87.5% (14/16), respectively, were not hits in a protein BLAST search with standard parameters, demonstrating the value of local structural comparisons for functional annotation. Additionally, the method is easy to update and extend over time via the addition of new sources of functional data, and reference databases can even be added or removed on a case-by-case basis.

COLLAPSE depends on the availability of solved 3D protein structures in the PDB. This restricts not only the number of homologous proteins that can be compared at each training step, but also the set of protein families which can even be considered—less than one third of alignments in the CDD contained at least two proteins with structures in the PDB. Including structures from AlphaFold Structure Database [52] would dramatically increase the coverage of our training dataset, but the utility of including predicted structures alongside experimentally solved structures in training or evaluation of machine learning models still needs to be evaluated. [53] A preliminary evaluation of our annotation method on the predicted structure for meizothrombin reveals high agreement with the corresponding PDB structure (Fig. S6) despite a root-mean-square deviation of 3.67 Å between the two structures, suggesting that COLLAPSE may already generalize to AlphaFold predictions for some proteins.

In summary, COLLAPSE is a general-purpose protein structure embedding method for functional site analysis. We provide a Python package and command-line tools for generating embeddings for any protein site, conducting functional site searches, and annotating input protein structures. We also provide downloadable databases of embeddings for a non-redundant subset of the PDB and for known functional sites. We anticipate that as more data becomes available, these tools will serve as a catalyst for data-driven biological discovery and become a critical component of the protein research toolkit.

4. Materials and Methods

4.1. Training dataset and data processing

COLLAPSE pre-training relies on a source of high-quality protein families associated with known structures and functions, as well as multiple sequence alignments (MSAs) in order to define site correspondences. We use the NCBI-curated subset of the Conserved Domain Database (CDD), [54,55] which explicitly validates domain boundaries using 3D structural information. We downloaded all curated MSAs from the CDD (n=17,906 as of Sep. 2021) and filtered out those that contained less than two proteins with structures deposited in the PDB. After removing

chains with incomplete data or which could not be processed properly, this resulted in 5,643 alignments for training, corresponding to 16,931 PDB chains (Fig. S2). We then aligned the sequences extracted from the ATOM records in each PDB chain to its MSA, without altering the original alignment, thus establishing the correct mapping from alignment position to PDB residue number. As a held-out set for validation, we select 1,370 families defined by PFAM [6] which do not share a common superfamily cluster (as defined by the CDD) with any training family. We then bin these families based on the average sequence identity to the nearest protein in the training dataset and sample five families from each bin, resulting in 50 validation families with varying levels of similarity to the training data (Table S1).

4.1.1. Definition of sites and environments

In general, we define protein sites relative to the location of the relevant residues. Specifically, we define the environment surrounding a protein site as all atoms within 10 Å radius of the functional center of the central residue. The functional center is defined as the centroid of the functional atoms of the side chain as defined by previous work. [26,27] For residues with two functional centers (Trp and Tyr), during training one is randomly chosen at each iteration, and at inference time the choice depends on the specific application (i.e. if the function being evaluated depends on the aromatic or polar group; see Table S2). If the functional atom is not known (e.g. for annotating unlabeled proteins), we take the average over all heavy side-chain atoms. Protein-level embeddings are computed as the mean over all residue-level embeddings.

4.1.2. Empirical background calculation

To make comparisons more meaningful and to provide a mechanism for calculating statistical significance, we quantile-transform all cosine similarities relative to an empirical cosine similarity distribution. To compute background distributions, we use a high-resolution (<2.0 Å), non-redundant subset of the PDB at 30% sequence similarity provided by the PISCES server [56] (5,833 proteins). We compute the embeddings of 100 sites from each structure, corresponding to five for each amino acid type, sampled with replacement. Exhaustively computing all pairwise similarities is computationally infeasible, so we sample $n = 50{,}000$ pairs of environments and compute the cosine similarity of each. We performed this procedure to generate empirical similarity distributions $(S_1, \ldots, S_n)$ for the entire dataset and for each amino acid individually (Figure S3). Cosine similarities (s) are then quantile-transformed relative to the relevant empirical cumulative distribution function:

$$\mathrm{F(s)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{S_i < s}$$

The p-value for any embedding comparison is then defined as $1 - F(s)$, or the probability that a randomly sampled pair of embeddings is greater than the pair in question. Amino acid–specific empirical backgrounds are used for functional site search and are aggregated into a single combined distribution for annotation. For the functional site–specific background used to filter

hits during annotation, we use an empirical background computed by comparing each functional site embedding to the embeddings of the corresponding amino acid in the 30% non-redundant PDB subset.

4.2. COLLAPSE training algorithm

Each iteration of the COLLAPSE pre-training algorithm consists of the following steps, as shown in Fig. 1. We trained our final model using the Adam optimizer [57] with a learning rate of 1e-4 and a batch size of 48 pairs for 1,200 epochs on a single TESLA V100 GPU. Model selection and hyperparameter tuning (e.g. environment radius, edge distance cutoff, learning rate schedule, pooling strategy, inclusion/exclusion of atoms) was evaluated using intrinsic embedding characteristics (see Section 2.1) and ATOM3D validation set performance (Section 4.3).

Step 1. Randomly sample one pair of proteins from the MSA and one aligned position from each protein (i.e. there is not a gap in either protein). Map MSA column position to PDB residue number using the pre-computed alignment described in Section 2.2. Note that this step ensures that each epoch, a different pair of residues is sampled from each CDD family, effectively increasing the size of the training dataset by many orders of magnitude relative to a strategy which trains on individual proteins or MSAs.

Step 2. Extract 3D environment around each selected residue (Section 4.1.1). Only atoms from the same chain are considered. Waters and hydrogens are excluded but ligands, metal ions, and cofactors are included.

Step 3. Convert each environment into a spatial graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node in the graph represents an atom and is featurized by a one-hot encoding of the atom type $\mathcal{V} \in$ { carbon (C), nitrogen (N), oxygen (O), fluorine (F), sulfur (S), chlorine (Cl), phosphorus (P), selenium (Se), iron (Fe), zinc (Zn), calcium (Ca), magnesium (Mg), and "other" }, representing the most common elements found in the PDB. Edges in the graph are defined between any pair of atoms separated by than 4.5 Å. Following Jing et al. (2021), [41] edges between atoms (i, j) with coordinates (x_i, x_j) are featurized using (1) a 16-dimensional Gaussian radial basis function encoding of distance $r(||x_j - x_i||)$ and (2) a unit vector $< x_j - x_i >$ encoding orientation.

Step 4. Compute embeddings of each site. We embed each pair of structural graphs $(\mathcal{G}_1, \mathcal{G}_2)$ using a pair of graph neural networks, each composed of three layers of Geometric Vector Perceptrons (GVPs), [40,41] which learn rotationally-equivariant representations of each atom and have proved to be state-of-the-art in a variety of tasks involving protein structure. [41,58] We adopt all network hyperparameters (e.g. number of hidden dimensions) from Jing et al. (2021). [41] Formally, each GVP learns a transformation of the input graph into 512-dimensional embeddings of each node:

$$f(\mathcal{G}_1;\ \theta) \ \rightarrow\ z_\theta \in \mathbb{R}^{|\mathcal{V}_1| \times 512}$$
$$f(\mathcal{G}_2;\ \phi) \ \rightarrow\ z_\phi \in \mathbb{R}^{|\mathcal{V}_2| \times 512}$$

The final embedding of the entire graph is then computed by global mean pooling over the embeddings of each atom. While in principle, the two networks could be direct copies of each other (i.e. have tied parameters $\theta = \phi$), we adopt the approach proposed by Grill et al (2020) [39] which refers to the two networks as the *online encoder* and the *target encoder*, respectively. Only the online network parameters θ are updated by gradient descent, while the target network parameters ϕ are updated as an exponential moving average of θ:

$$\phi \leftarrow \mu\phi + (1-\mu)\phi ,$$

where μ is a momentum parameter which we set equal to 0.99. No gradients are propagated back through the target network. Intuitively, the target network produces a regression target based on a "decayed" representation, while the online network is trained to continually improve this representation over the course of training. The online network is used to generate embeddings for all downstream applications.

Step 5. Compute loss and update parameters. The loss function is defined directly in the embedding space using the cosine similarity between the target network embedding $z_\phi \in \mathbb{R}^{512}$ and the online network embedding $z_\theta \in \mathbb{R}^{512}$ projected through a simple linear predictor network $pred(z_\theta) \in \mathbb{R}^{512}$. To increase the signal-to-noise ratio and encourage the model to learn functionally relevant information, we weight the loss at each iteration by the sequence conservation w_{cons} of that column in the original MSA (defined by the inverse of the Shannon's entropy of amino acids at that position, ignoring gaps). To reduce bias in computing the conservation, we include all proteins in the alignment curated by CDD, even those without corresponding structures. As a result of this, the loss function is expressed as:

$$\mathcal{L} = w_{cons} \cdot [2 - 2 \cdot \frac{\langle pred(z_\theta), z_\phi \rangle}{||pred(z_\theta)||_2 \cdot ||z_\phi||_2}] \text{ , where}$$

$$w_{cons} = \frac{1}{-\sum_{i \in AA} p_i log(p_i)}$$

Finally, we symmetrize the loss by passing each site in the input pair through both online and target networks and summing the loss from each. This symmetrized loss is then used to optimize the parameters of the online network using gradient descent.

4.3. Benchmarking on ATOM3D tasks

We evaluate COLLAPSE on two ATOM3D tasks concerned with local sites in one or more protein structures: protein interface prediction (PIP), and mutation stability prediction (MSP). We do not evaluate on residue identity (RES), which concerns predicting the identity of a masked central amino acid because the central amino acid of the environment is a key component of the COLLAPSE training procedure, resulting in almost perfect performance. See Townshend et al. [46] for details on dataset construction and reference architectures. Below we briefly describe each task and our fine-tuning procedure.

4.3.1 Protein Interface Prediction (PIP)

The PIP dataset contains protein-protein interactions mined from the PDB and split by 30% sequence identity. The task is set up as a binary classification of whether or not a pair of residues, one from each interacting chain, are in contact in the bound interface. For each pair, we embed the environments around each residue separately and concatenate the embeddings. We then train a feed-forward neural network on the combined embeddings to predict whether the residues are in contact. We use one hidden layer with dimension 2048, followed by ReLU activation and dropout with 50% probability.

4.3.2 Mutation Stability Prediction (MSP)

The MSP dataset consists of pairs of wild-type and mutant protein complexes, split by 30% sequence identity. The task is set up as a binary classification of whether or not the introduction of the mutation increases or decreases the stability of the complex. Like PIP, we embed the environments around each residue in the pair, concatenate, and train a feed-forward network to predict the binary outcome.

4.4. Training site-specific models on PROSITE data

We choose 10 sites presented in Torng et al. (2019), [26] selected because they are the most challenging to predict using FEATURE-based approaches. [25,26] For each functional site, we train a binary classifier on fixed COLLAPSE embeddings in five-fold nested cross-validation (CV). The classifiers are support vector machines (SVMs) with radial basis function kernels and weighted by class frequency. Within each training fold, the inner CV is used to select the regularization hyperparameter $C \in \{0.1, 1, 10, 100, 1000, 5000\}$ and the outer CV is used for model evaluation. To enable more accurate comparisons, we use the same dataset, evaluation procedures as Torng et al. (2019). [26] We benchmark against reported results for SVMs trained on FEATURE vectors (a direct comparison to our procedure) and 3D convolutional neural networks (3DCNNs) trained end-to-end on the functional site structures (the current state of the art for this task). We use PROSITE FN/FP sites as an independent validation of our trained models, using an ensemble of the models trained on each CV fold and the classification threshold determined above. A site is considered positive if the probability estimate from any of the five fold models is greater than the threshold. Some proteins contain more than one site; in these cases, the protein is considered to be positive if any sites are predicted to be positive.

4.5. Iterated functional site search

First, we embed the database to be searched against using the pre-trained COLLAPSE model. For the results presented in Section 2.4, we use the same PROSITE dataset used to train our cross-validated models to enable accurate comparisons. However, we also provide an embedding dataset for the entire PDB and scripts for generating databases for any set of protein structures. Then, we index the embedding database using FAISS, [59] which enables efficient similarity searches for high-dimensional data. For each site, we then perform the

following procedure five times with different random seeds in order to assess the variability of results under different query sites. The input parameters are the number of iterations n_{iter} and the p-value cutoff for selecting sites at each iteration p_{cutoff}.

1. Sample a single site from the PROSITE TP dataset (to simulate querying a known functional site), generate COLLAPSE embedding, and add to query set.
2. Compute effective cosine similarity cutoff s_{cutoff} using the $(1 - p_{cutoff})$ quantile of the empirical background for the functional amino acid of the query site (e.g. cysteine for an EGF_1 site).
3. Compare embedding(s) of query to database and retrieve all neighbors within s_{cutoff} of the query.
4. Add all neighbors to query set and repeat Step 3 n_{iter} times. Note that when there is more than one query point, neighbors to *any* point in the query are returned.
5. Compute precision and recall of final query set, using PROSITE data as ground truth.

4.6. Protein site annotation

Instead of a database of all protein sites, the annotation method requires a database of known functional sites. We use all true positive sites defined in PROSITE. For each pattern, we identify all matching PDBs using the ScanProsite tool [14] and extract the residues corresponding to all fully conserved positions in the pattern (i.e. where only one residue is allowed). The environment around each residue is embedded using COLLAPSE. We also embed all residues in the Catalytic Site Atlas (CSA), a curated dataset of catalytic residues responsible for an enzyme's function. All data processing matches the pre-training procedure. The final dataset consists of 25,407 embeddings representing 1,870 unique functional sites.

The annotation method operates in a similar fashion to the search method, where each residue in the input protein is embedded and compared to the functional site database. Any residue that has a hit with a p-value below the pre-specified cutoff is returned as a potential functional site. To filter out false positives due to common or non-specific features (e.g. small polar residues in alpha-helices), we also remove hits which are not significant against the empirical distribution specific to that functional site (Section 4.1.2). This results in a modified mutual best hit criterion with two user-specified parameters: the residue-level and site-level significance thresholds. Along with each hit is the metadata associated with the corresponding database entry (PDB ID, functional site description, etc.) so each result can be examined in more detail. For the examples presented we remove all ligand atoms from the input structure to reduce the influence of non-protein atoms on the embeddings.

5. Acknowledgments

We thank Kristy Carpenter, Delaney Smith, Adam Lavertu, and Wen Torng for useful discussions. Computing for this project was performed on the Sherlock cluster; we would like to thank Stanford University and the Stanford Research Computing Center for providing

computational resources and support. A.D. is supported by LM012409 and R.B.A. is supported by NIH GM102365 and Chan Zuckerberg Biohub.

Figures

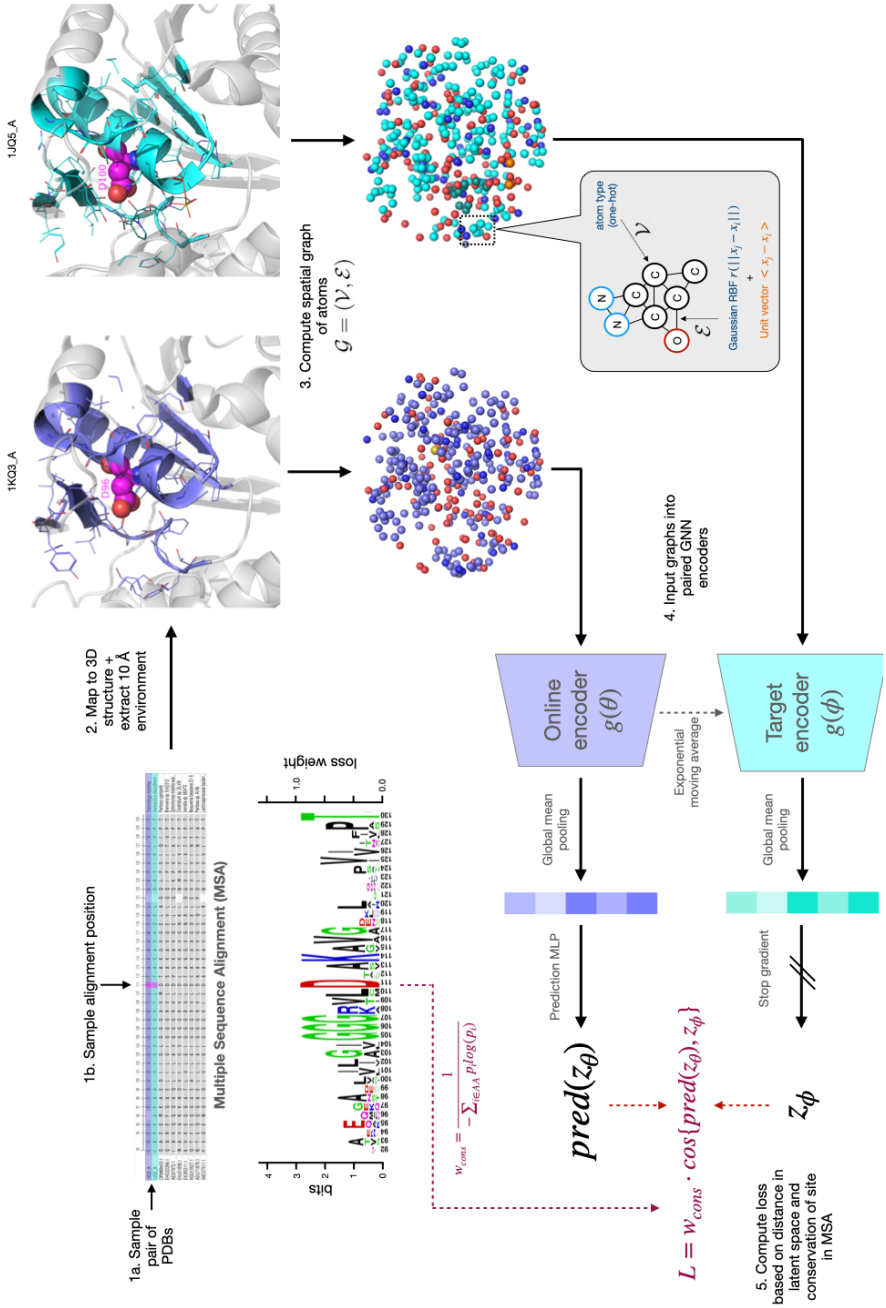

Figure 1. Schematic of a single iteration of COLLAPSE algorithm. Clockwise from the top left, we show **(1a,b)** the process of sampling a pair of sites from the MSA, **(2)** extracting the corresponding structural environments, and **(3)** converting into a spatial graph. The inset shows the node and edge featurization scheme. Finally, we show **(4)** a schematic of the network architecture, consisting of paired graph neural networks followed by mean pooling over all nodes to produce site embeddings. **(5)** These embeddings are then compared using a loss function weighted by the conservation of the position in the MSA, as shown by the sequence logo in center left.

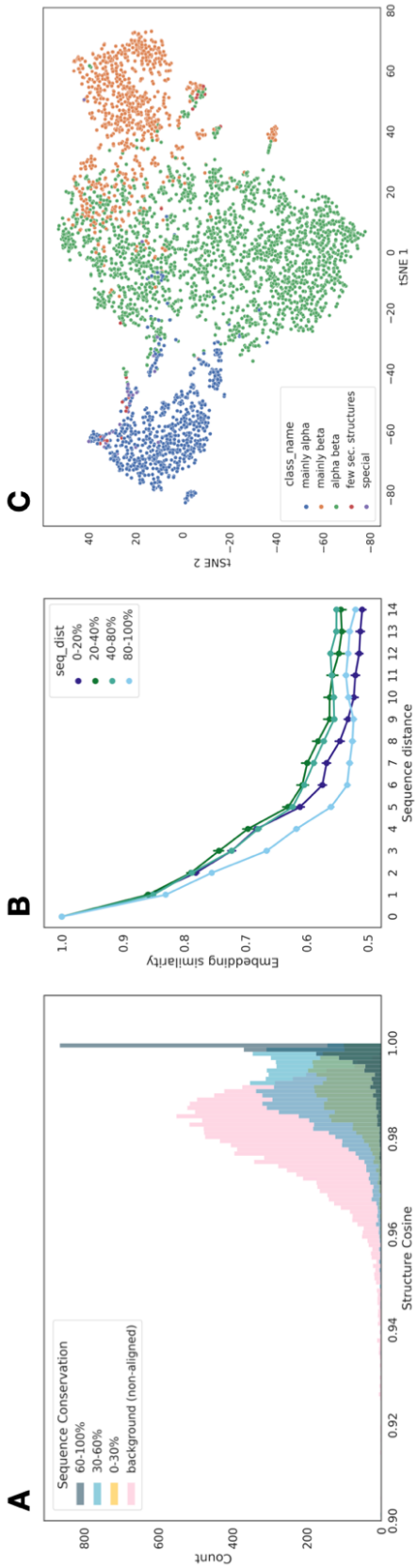

Figure 2. Analysis of learned embeddings. **(a)** Raw cosine similarity distributions (i.e. before quantile transformation) of aligned sites, binned by the sequence conservation of the corresponding column of the MSA. Highly conserved positions also have highly similar embeddings, but even less conserved positions have more similar embeddings than randomly sampled non-aligned sites (in pink). **(b)** Spatial sensitivity of embedding similarity, as measured by the sequence distance between two sites. Results are stratified by the average distance to the closest training protein, demonstrating that neighboring embeddings can be readily distinguished even for proteins with very low similarity to the training set. **(d)** tSNE projection of average protein-level embeddings for single-domain chains, colored by the highest-level CATH class, showing that embeddings effectively capture secondary structure patterns.

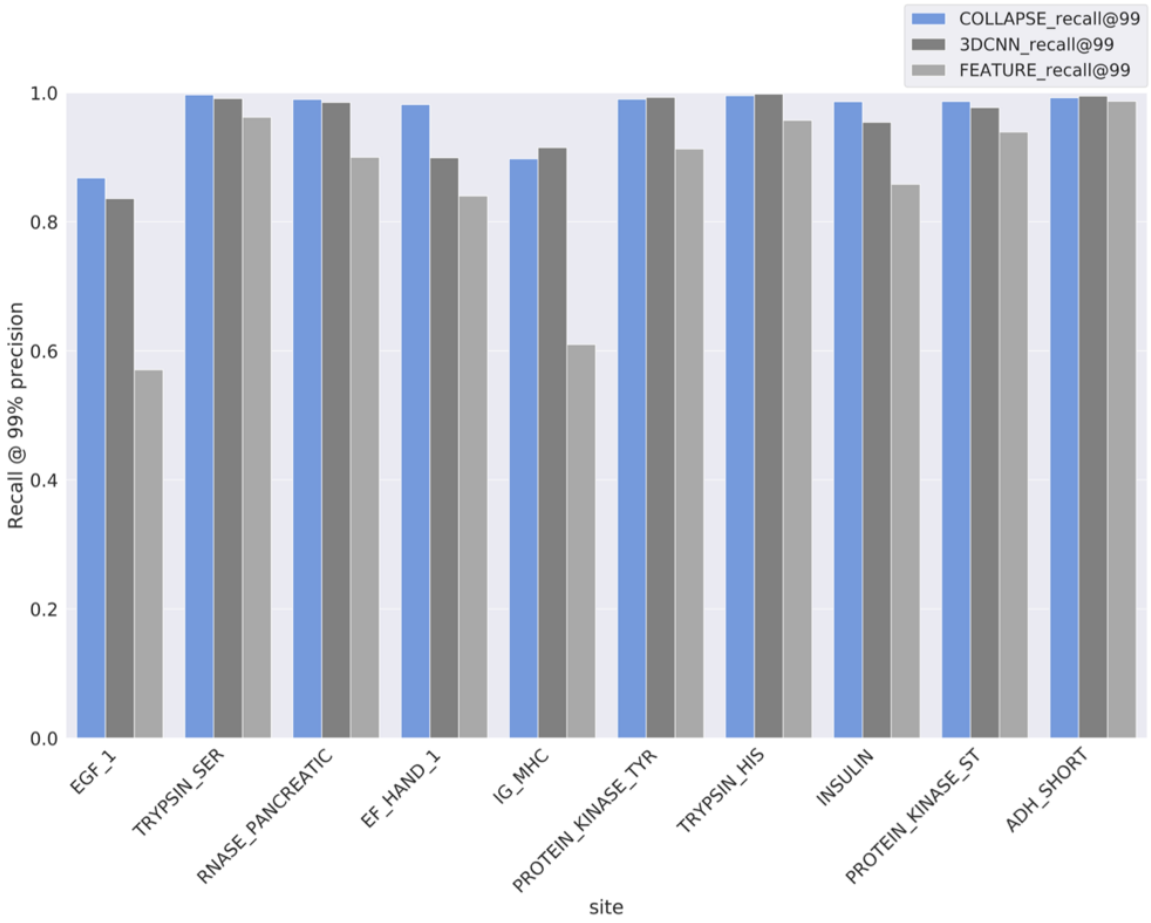

Figure 3. Performance of models trained on true positives from 10 PROSITE functional sites in 5-fold cross-validation: COLLAPSE embeddings + SVM (blue), 3DCNN trained end-to-end (dark gray), and FEATURE vectors + SVM (light gray). Metric is the recall for all TP annotations at a threshold which produces 99% precision. COLLAPSE achieves better recall than FEATURE and better or comparable recall to the 3DCNN.

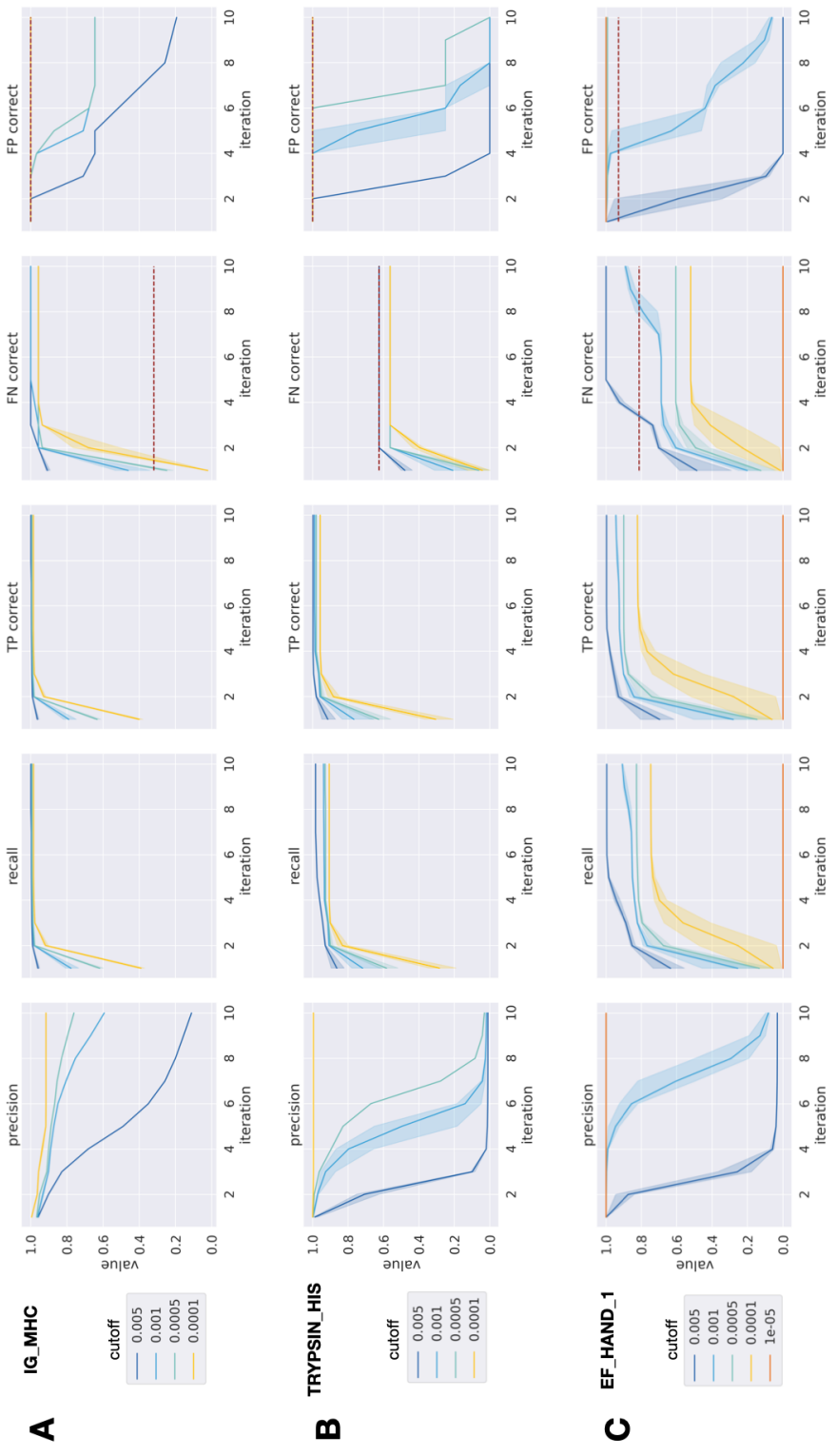

Figure 4. Iterated functional site search performance per iteration for three PROSITE families. Colors denote different user-specified empirical p-value cutoffs and error bars represent variance over three randomly sampled queries. From left to right, metrics shown are: precision across all results (including TP, FP, and FN), recall across all results, proportion of TP sites predicted correctly, proportion of FN sites predicted correctly, and proportion of FP sites predicted correctly. For FN and FP, the performance of our CV-trained models is shown as a red dashed line. Sample error is shown for three random starting queries. The three families shown are **(a)** IG_MHC, **(b)** TRYPSIN_HIS, and **(c)** EF_HAND_1, in order of relative performance compared to CV-trained models. While the performance characteristics vary across sites, the number of iterations and p-value cutoff can be tuned to achieve good performance.

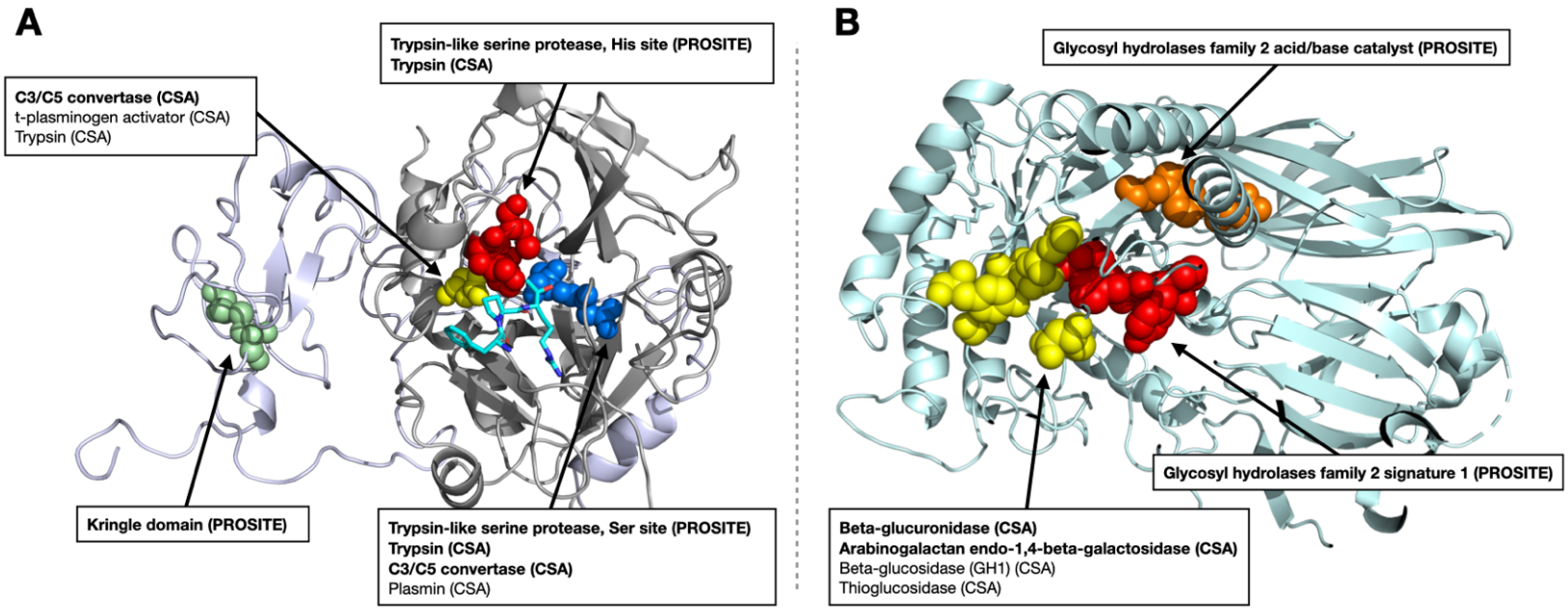

Figure 5. Results of functional annotation tool applied to **(a)** meizothrombin (PDB ID 1A0H) and **(b)** beta-glucuronidase (PDB ID 3HN3), both at $p < 1 \times 10^{-4}$. No member of the beta-glucuronidase family is in the training set (maximum sequence identity 2.8%). Functional residues identified by our method are shown as spheres, with colors corresponding to the functional site. Hits labeled in bold are also significant at a more stringent cutoff ($p < 5 \times 10^{-5}$). All hits represent either the correct function or those of very closely related proteins, showing that COLLAPSE is effective for annotation of proteins whether similar proteins are present in the training set.

Tables

Table 1. Performance of models trained on ATOM3D benchmark tasks. Comparisons are made with ATOM3D reference architectures (3DCNN, GNN, and ENN) as well as the GVP-GNN results reported in Jing et al. (2021) [41], which is state-of-the-art for these datasets. We report mean and standard deviation across three training runs. Numbers in bold indicate best performance on each task (within one standard deviation).

Task (metric)	*COLLAPSE (fixed)*	*COLLAPSE (fine-tuned)*	*ATOM3D 3DCNN*	*ATOM3D GNN*	*ATOM3D ENN*	*GVP-GNN*
PIP (AUROC)	0.848 ± 0.018	**0.881 ± 0.004**	0.844 ± 0.002	0.669 ± 0.001	N/A	0.866 ± 0.004
MSP (AUROC)	0.616 ± 0.006	**0.668 ± 0.018**	0.574 ± 0.005	0.621 ± 0.009	0.574 ± 0.040	**0.680 ± 0.015**

Table 2. Performance of models trained on PROSITE TP/TN on held-out PROSITE FP/FN annotations. Comparisons are made with FEATURE and 3DCNN numbers as reported in Torng et al. (2019) [26]. Numbers in bold indicate best performance on each site.

Site	*PROSITE label*	*COLLAPSE*	*FEATURE*	*3DCNN*	*PROSITE total*
ADH_SHORT	**FN**	**11**	8	7	14
	FP	30	**33**	**33**	33
EF_HAND_1	**FN**	**40**	28	34	48
	FP	120	106	**125**	128
EGF_1	**FN**	**60**	34	58	90
	FP	**19**	**19**	**19**	19
IG_MHC	**FN**	**21**	8	8	47
	FP	**31**	**31**	**31**	31
PROTEIN_KINASE_ST	**FN**	**269**	264	268	271
*PROTEIN_*	**FN**	**3**	**3**	**3**	3
KINASE_TYR	**FP**	14	**20**	**20**	20
TRYPSIN_HIS	**FN**	**10**	3	**10**	16
	FP	**4**	**4**	**4**	4
TRYPSIN_SER	**FN**	**9**	**9**	**9**	12
	FP	**1**	**1**	**1**	1

References

1. Akiva, E. *et al.* The Structure-Function Linkage Database. *Nucleic Acids Res.* **42**, D521-30 (2014).
2. Furnham, N. *et al.* The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **42**, D485-9 (2014).
3. Ribeiro, A. J. M. *et al.* Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623 (2018).
4. Berman, H. M. *et al.* The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 899–907 (2002).
5. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
6. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
7. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
8. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33**, D284-8 (2005).
9. Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* **41**, D387-95 (2013).
10. Bernhofer, M. *et al.* PredictProtein - Predicting Protein Structure and Function for 29 Years. *Nucleic Acids Res.* **49**, W535–W540 (2021).
11. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).
12. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
13. Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333**, 863–882 (2003).
14. Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344-7 (2013).
15. Attwood, T. K. The PRINTS database: a resource for identification of protein families. *Brief. Bioinform.* **3**, 252–263 (2002).
16. Fetrow, J. S. & Skolnick, J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281**, 949–968 (1998).
17. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer: deep networks for protein functional inference. *bioRxiv* 2021.09.20.461077 (2021) doi:10.1101/2021.09.20.461077.
18. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).
19. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 1–14 (2021).

20. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
21. Ramola, R., Friedberg, I. & Radivojac, P. The field of protein function prediction as viewed by different domain scientists. *bioRxiv* 2022.04.18.488641 (2022) doi:10.1101/2022.04.18.488641.
22. Zhao, J., Cao, Y. & Zhang, L. Exploring the computational methods for protein-ligand binding site prediction. *Comput. Struct. Biotechnol. J.* **18**, 417–426 (2020).
23. Tubiana, J., Schneidman-Duhovny, D. & Wolfson, H. J. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods* (2022) doi:10.1038/s41592-022-01490-7.
24. Moraes, J. P. A., Pappa, G. L., Pires, D. E. V. & Izidoro, S. C. GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms. *Nucleic Acids Res.* **45**, W315–W319 (2017).
25. Buturovic, L., Wong, M., Tang, G. W., Altman, R. B. & Petkovic, D. High precision prediction of functional sites in protein structures. *PLoS One* **9**, e91240 (2014).
26. Torng, W. & Altman, R. B. High precision protein functional site detection using 3D convolutional neural networks. *Bioinformatics* **35**, 1503–1512 (2019).
27. Bagley, S. C. & Altman, R. B. Characterizing the microenvironment surrounding protein sites. *Protein Sci.* **4**, 622–635 (2008).
28. Tang, G. W. & Altman, R. B. Knowledge-based fragment binding prediction. *PLoS Comput. Biol.* **10**, e1003589 (2014).
29. Liu, T. & Altman, R. B. Using multiple microenvironments to find similar ligand-binding sites: application to kinase inhibitor binding. *PLoS Comput. Biol.* **7**, e1002326 (2011).
30. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
31. Oquab, M., Bottou, L., Laptev, I. & Sivic, J. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. in *2014 IEEE Conference on Computer Vision and Pattern Recognition* 1717–1724 (2014).
32. Hu, W. *et al.* Strategies for Pre-training Graph Neural Networks. *arXiv [cs.LG]* (2019).
33. Duvenaud, D. *et al.* Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv [cs.LG]* (2015).
34. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv [cs.LG]* (2017).
35. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv* 622803 (2019) doi:10.1101/622803.
36. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
37. Zhang, Z. *et al.* Protein Representation Learning by Geometric Structure Pretraining. *arXiv [cs.LG]* (2022).
38. Hermosilla, P. & Ropinski, T. Contrastive Representation Learning for 3D Protein Structures. (2021).
39. Grill, J.-B. *et al.* Bootstrap your own latent - A new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **33**, 21271–21284 (2020).

40. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. & Dror, R. Learning from Protein Structure with Geometric Vector Perceptrons. *arXiv [q-bio.BM]* (2020).
41. Jing, B., Eismann, S., Soni, P. N. & Dror, R. O. Equivariant Graph Neural Networks for 3D Macromolecular Structure. *arXiv [cs.LG]* (2021).
42. Chen, X. & He, K. Exploring Simple Siamese Representation Learning. *arXiv [cs.CV]* (2020).
43. Che, F. *et al.* Self-supervised Graph Representation Learning via Bootstrapping. *arXiv [cs.LG]* (2020).
44. Detlefsen, N. S., Hauberg, S. & Boomsma, W. Learning meaningful representations of protein sequences. *Nat. Commun.* **13**, 1914 (2022).
45. Orengo, C. A. *et al.* CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
46. Townshend, R. J. L. *et al.* ATOM3D: Tasks on Molecules in Three Dimensions. in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)* (2021).
47. Torng, W. & Altman, R. B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *Journal of Chemical Information and Modeling* 473074 (2019) doi:10.1021/acs.jcim.9b00628.
48. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545-9 (2010).
49. van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. *bioRxiv* 2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.
50. Zemla, A., Allen, J. E., Kirshner, D. & Lightstone, F. C. PDBspheres - a method for finding 3D similarities in local regions in proteins. *bioRxiv* 2022.01.04.474934 (2022) doi:10.1101/2022.01.04.474934.
51. Valasatava, Y., Rosato, A., Cavallaro, G. & Andreini, C. MetalS3, a database-mining tool for the identification of structurally similar metal sites. *J. Biol. Inorg. Chem.* **19**, 937–945 (2014).
52. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
53. Derry, A., Carpenter, K. A. & Altman, R. B. Training data composition affects performance of protein structure analysis algorithms. *Pac. Symp. Biocomput.* **27**, 10–21 (2022).
54. Marchler-Bauer, A. *et al.* CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31**, 383–387 (2003).
55. Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).
56. Wang, G. & Dunbrack, R. L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).
57. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).
58. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. *bioRxiv* 2022.04.10.487779 (2022) doi:10.1101/2022.04.10.487779.
59. Johnson, J., Douze, M. & Jegou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **7**, 535–547 (2021).

Supplementary Materials

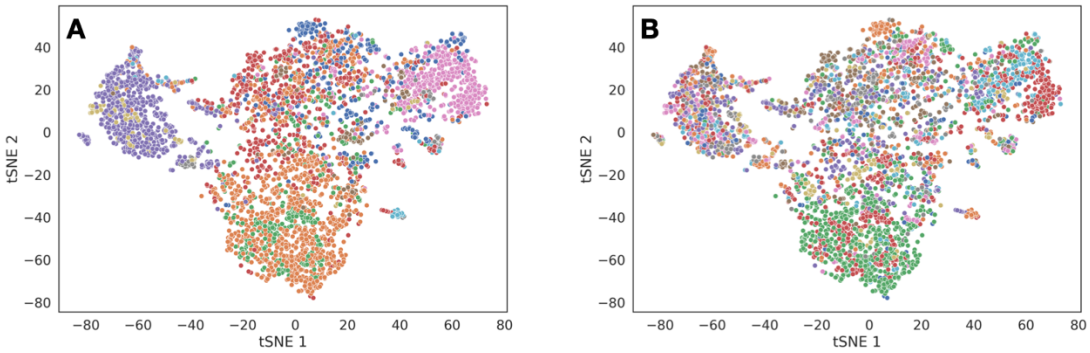

Figure S1. PCA of embeddings of single-domain proteins at lower levels of the CATH hierarchy: **(a)** architecture and **(b)** topology.

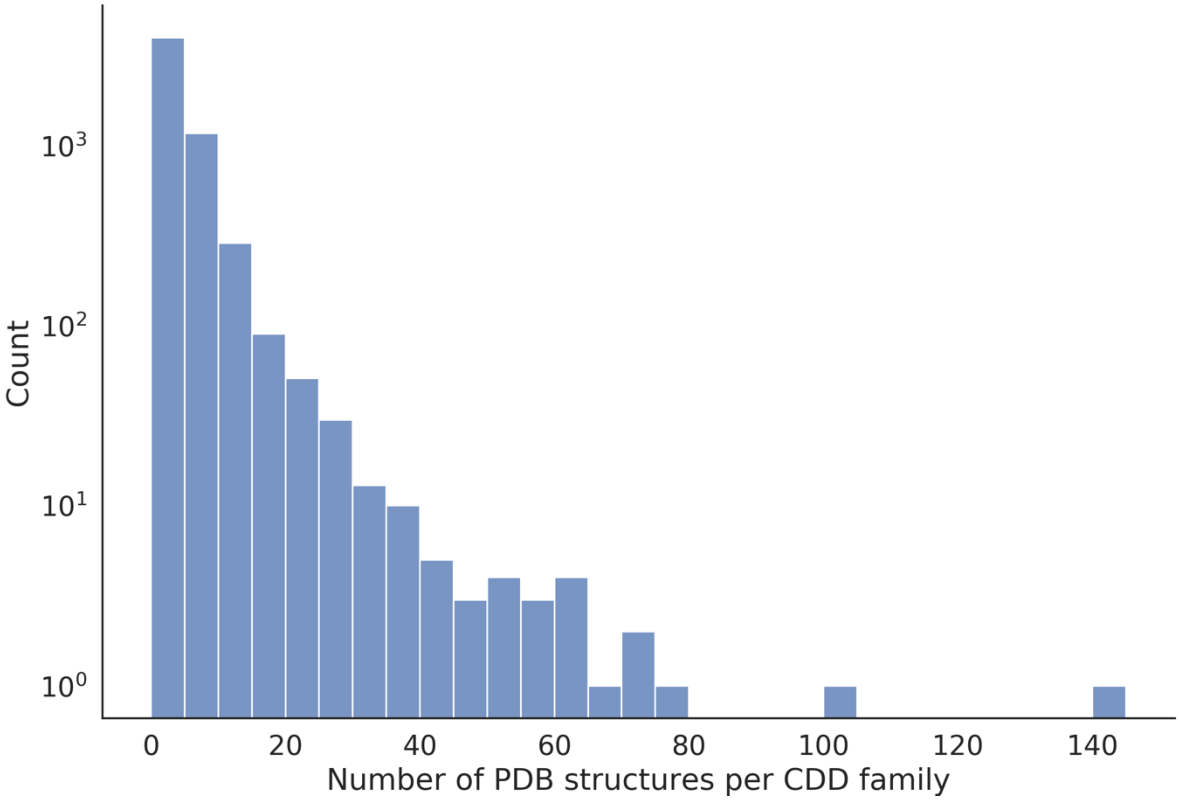

Figure S2. Histogram of number of PDB structures per CDD family in training dataset.

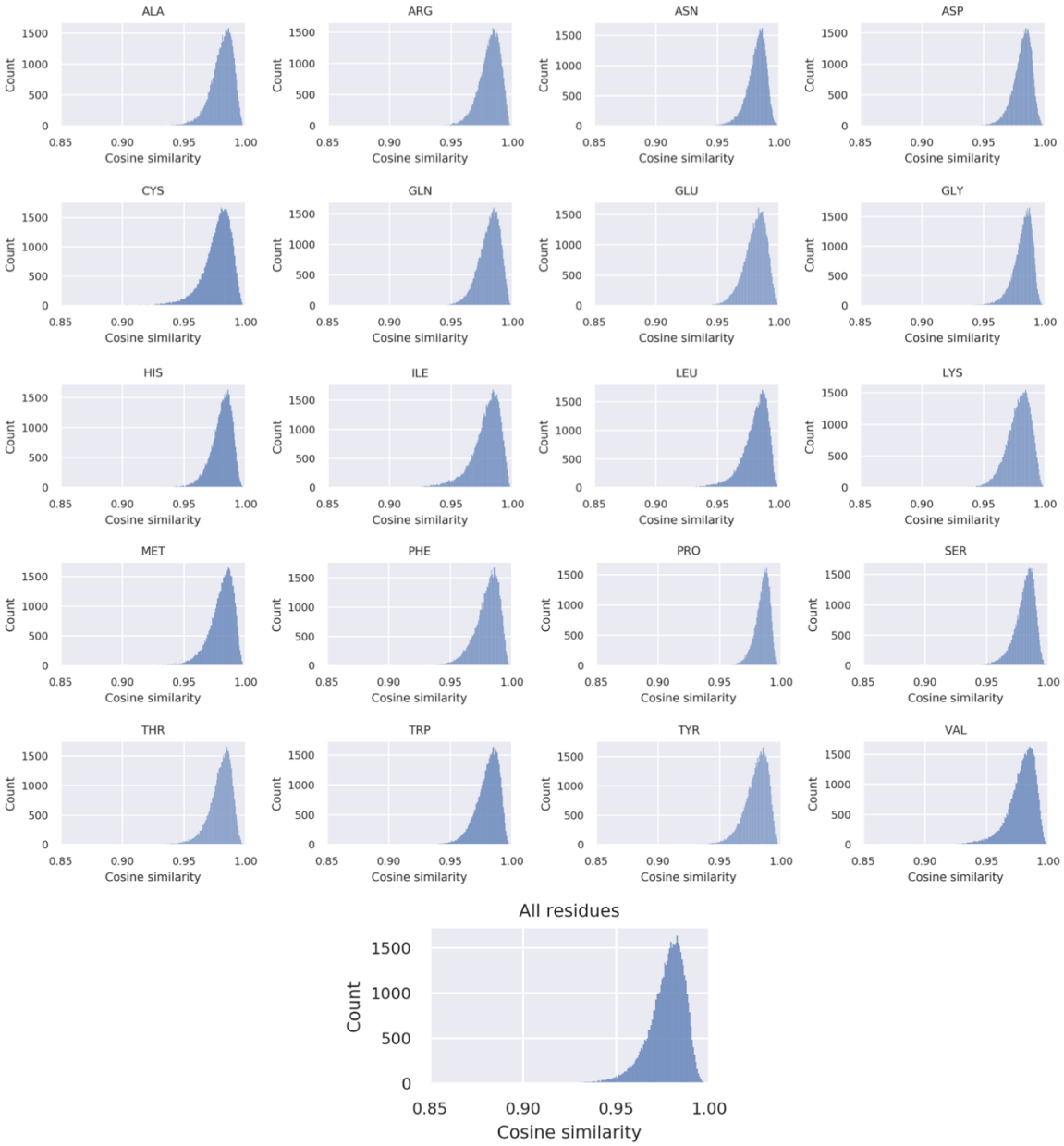

Figure S3. Empirical cosine similarity distributions computed for each amino acid and the combined dataset.

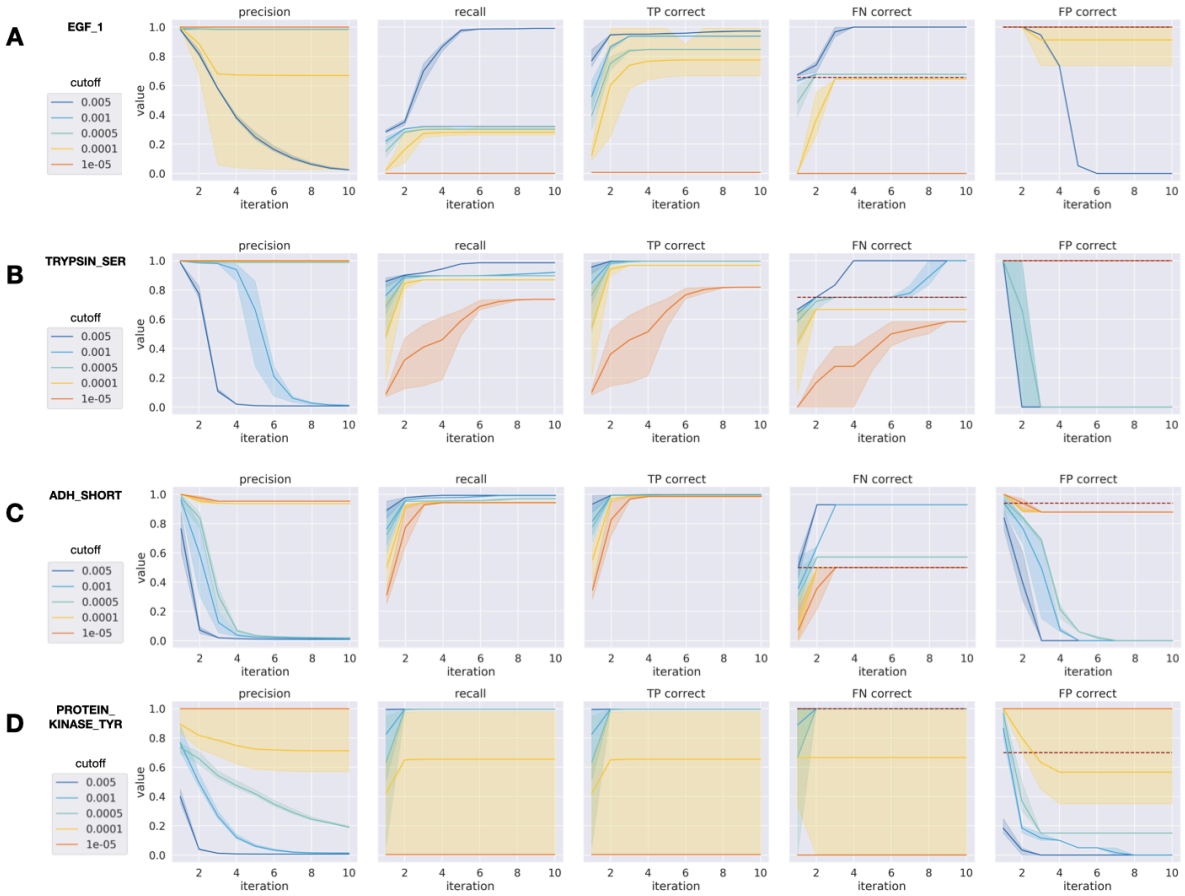

Figure S4. Iterated functional site search performance per iteration for remaining PROSITE families with FP and FN annotations not shown in Figure 5: **(a)** EGF_1, **(b)** TRYPSIN_SER, **(c)** ADH_SHORT, and **(d)** PROTEIN_KINASE_TYR.

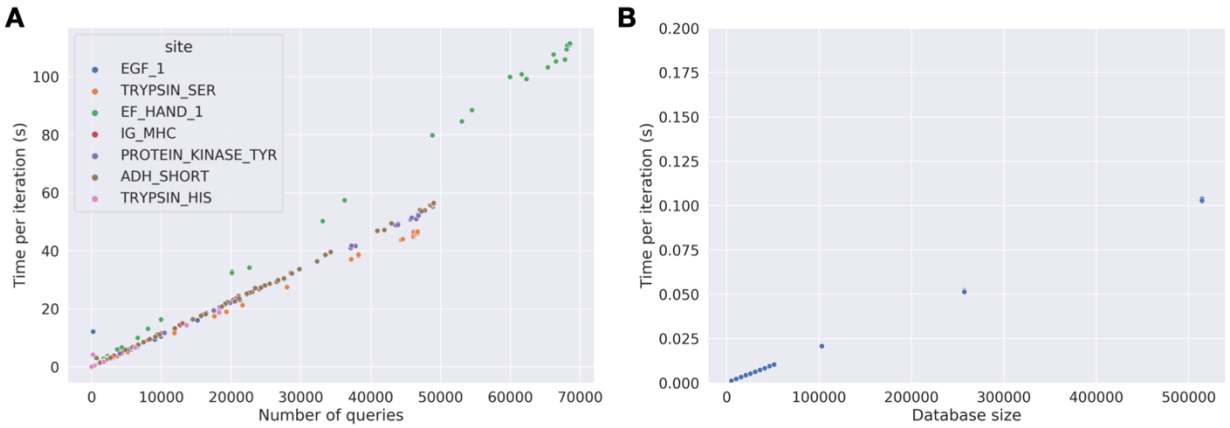

Figure S5. Runtime analysis for functional site search tool. Time per iteration as a function of **(a)** number of queries at the start of the iteration, colored by functional site, and **(b)** the size of the database searched against.

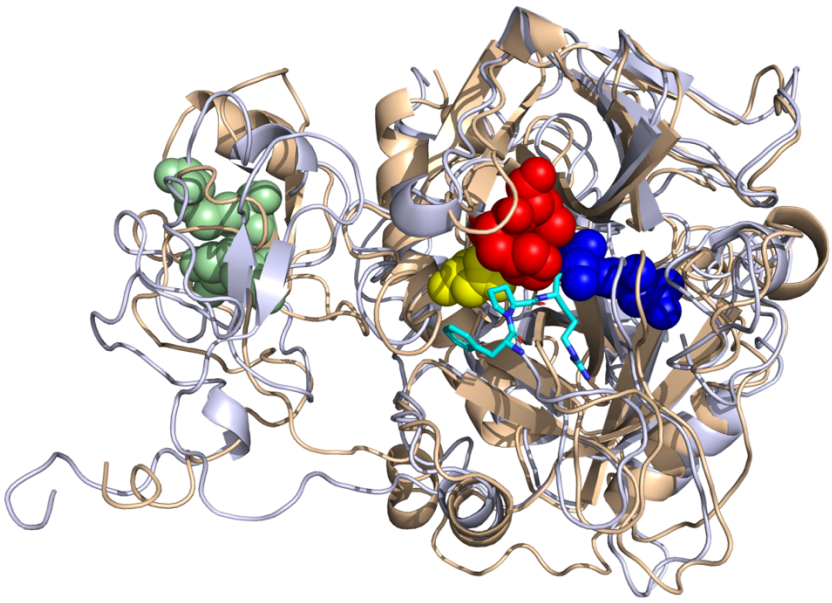

Figure S6. Annotated structure of meizothrombin structure predicted by AlphaFold (gold; Uniprot ID P00735) superimposed on crystal structure (light blue; PDB ID 1A0H). Colors correspond to the predicted functional site, using the same colors as Fig. 5a.

Table S1. Pfam families selected for held-out validation set and corresponding sequence identity to nearest protein in CDD training set.

Pfam family	Average sequence identity to closest training set protein
pfam02445	0.0949
pfam04122	0.0790
pfam00297	0.0865
pfam01278	0.0774
pfam18981	0.0789
pfam07676	0.1115
pfam01395	0.1435
pfam09477	0.1652
pfam13739	0.1053
pfam10862	0.1411
pfam04175	0.2349
pfam01455	0.2550
pfam00706	0.2787
pfam05188	0.2837
pfam09392	0.2749
pfam03497	0.3162
pfam00766	0.3052
pfam01808	0.3068
pfam04726	0.3846
pfam08799	0.3117
pfam14204	0.4270
pfam01396	0.4638
pfam00754	0.4271
pfam08501	0.4856
pfam14821	0.4063
pfam03950	0.5293
pfam08674	0.5456
pfam03104	0.5529
pfam13720	0.5515
pfam19034	0.5873
pfam02811	0.6217
pfam02927	0.6709
pfam09092	0.6313
pfam00173	0.6495
pfam00814	0.6327
pfam03366	0.7891

pfam01017	0.7480
pfam00654	0.7142
pfam17855	0.7609
pfam06628	0.7867
pfam17136	0.8088
pfam03931	0.8750
pfam02511	0.8538
pfam10431	0.8496
pfam05001	0.8714
pfam01412	1.0
pfam07161	1.0
pfam14324	1.0
pfam00567	1.0
pfam12124	1.0

Table S2. Description of 10 PROSITE functional sites and definition of functional centers for each.

Site name	*Description*	*Target residue index in pattern*	*Amino Acid*	*Functional Atom*
EGF_1	EGF-like domain signature 1	10	CYS	SG
TRYPSIN_SER	Serine proteases, trypsin family, serine active site	6	SER	OG
RNASE_ PANCREATIC	Pancreatic ribonuclease family signature	2	LYS	NZ
EF_HAND_1	EF-hand calcium-binding domain	1	ASP	OD1
IG_MHC	Immunoglobulins and major histocompatibility complex proteins signature	3	CYS	SG
PROTEIN_ KINASE_TYR	Tyrosine protein kinases specific active-site signature	5	ASP	OD2
TRYPSIN_HIS	Serine proteases, trypsin family, histidine active site	5	HIS	NE2
INSULIN	Insulin family signature	2	CYS	SG
PROTEIN_ KINASE_ST	Serine/Threonine protein kinases active-site signature	5	ASP	OD2
ADH_SHORT	Short-chain dehydrogenases/reductases family signature	5	TYR	OH