Supplementary Materials

Supplementary Note 1.

The range of cosine similarities is relatively small (~0.9–1.0), 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. We have experimented with replacing the mean pooling layer with a distance-weighted averaging function and a learned attention-based pooling layer; both approaches did increase the distribution of cosine similarities, but both also resulted in inferior performance across downstream tasks. Nonetheless, 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 $p = 1 \times 10^{-4}$.

Supplementary Note 2.

There are many hyperparameters involved in the pre-training phase, and we were not able to exhaustively search all possibilities due to computational constraints. Some parameters, including embedding dimension, environment size, and number and size of GVP layers, were therefore selected based on prior knowledge. We chose 512 for our embedding dimension because it is a middle ground between the lower-dimensional embeddings typically used in graph-based models and the higher-dimensional embeddings typically used in sequence-based models. We chose an environment size of 10 Å based on previous work with FEATURE and 3D
CNNs, which has found that atomic details beyond 8–10 angstroms do not provide enough additional information to justify the additional memory and computational cost \(^1\)–\(^3\). Finally, the GVP architecture was adopted directly from Jing et al. (2021). \(^4\) There are also alternative graph featurization schemes which could be used, including those that incorporate physico-chemical features directly. We have experimented with the addition of a feature denoting whether an edge represents a covalent bond, but it didn’t have much effect on the outcome of the model. We expect that such features are inferred by the model during training based on the raw atom types and positions, as is commonly observed of deep learning algorithms in the high-data regime. Nonetheless, it is possible that the inclusion of certain additional properties as features during the pre-training phase could improve the quality of the embeddings.

**Supplementary Note 3.**

*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.

*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.

**Supplementary Note 4.**

We retrieved all PDBs and labeled surface points from the MaSIF repository (https://github.com/LPDI-EPFL/masif) for the train and test splits evaluated in Gainza et al. as well as the more difficult test set of transient interactions. For all datasets, we computed COLLAPSE embeddings for environments centered around each of the coordinates sampled on the protein surface. The prediction model consisted of the COLLAPSE encoder and a feed-forward prediction head which comprised two hidden layers of dimension 1024 with ReLU activation and dropout with probability 0.1. We report overall AUROC across all residues in each test set and compare to the corresponding metrics from Gainza et al. While our model performs slightly worse than MaSIF-site, it is important to note that our method is designed to represent sites centered around residues rather than surfaces centered around sampled surface points. This means that for any surface points that are not very close to the side chain of an amino acid, the environment is out-of-distribution for our model, making transfer learning more difficult. Pre-training a model on arbitrary surface points rather than residues may help to alleviate this issue, but it would be more difficult to define surface-to-surface correspondences between homologous pairs during training and the model would likely not generalize as well to functional sites involving residues in the core of the protein. It is also possible that further tuning of hyperparameters and model architectures would further reduce the performance.
deficit, but this is out of the scope of this work, which aims to establish the general-purpose utility of COLLAPSE embeddings given relatively simple downstream training schemes.

**Supplementary Note 5.**

We compared the performance of COLLAPSE to analogous methods built using the large language model embeddings from ESM-1b \(^6\). We computed embeddings for each site in our PROSITE dataset and implemented the same training and evaluation procedures for site-specific predictions. We find that when trained in cross-validation, ESM-1b embeddings can almost perfectly classify the TP examples. This is likely due to the fact that these examples are selected based on sequence motifs, which are easily picked up by the language model. When evaluated on the FN and FP proteins, results are more mixed; some sites perform better than COLLAPSE and some perform worse. Most notably, for IG_MHC, the ESM-1b embeddings are unable to identify any of the FN examples. We also evaluated the ability of ESM-1b embeddings to conduct local functional site searches and functional site annotation, where they generally underperform COLLAPSE embeddings in both cases. In summary, language model embeddings have strengths that largely complement structural embeddings such as COLLAPSE, and some tasks may be better suited for one than the other.

**Supplementary Note 6.**

As an additional baseline, we evaluated the ability of MMSeqs2 \(^7\), a widely used method for sequence-based searches, to reclassify PROSITE FN and FP proteins. For each sequence annotated as a FN or FP, we conducted a single search against the Swissprot database. We then
extracted the top hit from this search and checked whether this protein was a known example of that particular functional site. If so, we consider that protein successfully re-classified. The results shown in Table S4 clearly demonstrate the value of training site-specific prediction models and of the use of local structural embeddings, as very few of the FN proteins are successfully identified as members of the functional site. However, our simple MMSeqs2 baseline is very effective at ruling out FPs.

**Supplementary Note 7.**

The procedure for computing FEATURE vectors for the MSP dataset was as follows:

1. Create two PDB files for every example in the dataset, one for wild-type and one for mutated.
2. Compute structural features of each PDB using DSSP.
3. Create two .ptf files for every example in the dataset, one for wild-type and one for mutated, using the coordinates of the functional atom of the central residue.
4. Compute FEATURE vectors using `featurize` from the FEATURE package (https://simtk.org/projects/feature; see user manual for details).

The training procedure is identical to that used for the COLLAPSE benchmark. While FEATURE performs quite well relative to the ATOM3D benchmarks on this difficult task, COLLAPSE still outperforms it when fine-tuned. We note that the inability to fine-tune FEATURE on task-specific datasets is a major limitation which is addressed by our self-supervised learning framework. Unfortunately, we were not able to fully train a FEATURE benchmark for PIP, since the computation of FEATURE vectors is not very efficient for very large datasets. Given the
relative difficulty of using FEATURE relative to COLLAPSE, its inability to be fine-tuned, and its inferior performance across many benchmarks, we are confident that COLLAPSE provides an improved general-purpose representation of protein sites.

Supplementary Note 8.

We obtained all PDBs and site data for each dataset from the following URL https://www.ccs.neu.edu/home/radivojac/data/xin_cppps_supplementary.zip. We then embedded all residues in each dataset and split the embeddings randomly into 10 folds, as in Xin et al. The model and training protocol was exactly the same as for the PROSITE models (Section 4.4): we train an SVM with radial basis function kernel and evaluate total AUROC over all held-out folds. We compare against the three methods presented in Xin et al., without retraining. Note that the exact splits for each fold are different between our model and the baseline models.

Supplementary Note 9.

We obtained the train, validation, and test splits for both remote homology and enzyme classification tasks from Hermosilla et al., downloaded the corresponding PDB chains and domains, and embedded all residues in each. All ligands and heteroatoms were excluded. We then trained a bidirectional two-layer gated recurrent unit (GRU) model to aggregate over the full sequence of embeddings for each protein. The GRU output was computed as the concatenation of the final hidden state of the forward and backward networks in the last layer, which was fed into a two-layer feed-forward network with ReLU activation and dropout with
probability 0.5 to predict the final classification. Basic hyperparameter tuning was performed, and the best model was selected based on the validation accuracy. The final fold prediction model used a learning rate of $1 \times 10^{-3}$, batch size of 64, hidden unit size of 256, and an AdamW optimizer with weight decay of 0.5. The enzyme class prediction model used a learning rate of $1 \times 10^{-4}$, batch size of 32, hidden unit size of 1024, and an AdamW optimizer with weight decay of 0.5. We compare to all baseline methods presented in Hermosilla et al., including the hidden Markov model (HMM) based method HHSuite \textsuperscript{11} as well as sequence- and structure-based machine learning models, many of which were fine-tuned specifically for each task. Our model, on the other hand, only aggregates over fixed COLLAPSE embeddings. Notably, HHSuite is one of the top performers for all tasks except the most difficult homology split, where we achieve much higher accuracy. The performance of the other methods varies across tasks, and even without fine-tuning our COLLAPSE-RNN model consistently ranks in the top five methods for all tasks and test sets. We expect that this performance could be improved by fine-tuning or by improving the aggregation method from the simple RNNs used here.
Supplementary figures

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.
Figure S7. Performance of COLLAPSE on protein-level prediction tasks using a GRU aggregator. The tasks are fold prediction (homology) and enzyme class prediction using datasets and baseline methods from Hermosilla et al. (2021). We report results for three held-out test sets for the homology task, defined by the stringency of the structural overlap with the training set, with “fold split” being the most difficult and “family split” being the easiest. Our COLLAPSE-RNN method is shown in blue, and the HMM-based HHsuite is shown in dark gray. For the ML-based baselines, we note which operate primarily on sequences and which operate primarily on structure using diagonal lines and dots, respectively.
Figure S8. Performance of COLLAPSE on protein-protein interaction binding site identification.

We compare to MaSIF-site as well as two baseline methods, as reported in Gainza et al. (2020). The metric is AUROC over all residues in all proteins, and we evaluate on both the full test set (left) and the more difficult subset of transiently interacting proteins (right). See Supplementary Note 4 for details.
Figure S9. Performance of ESM-1b language model embeddings in functional site search (see Supplementary Note 5 for details). Sequence embeddings are not able to match the recall of COLLAPSE over most sites, even after many iterations, although precision is often high.

Figure S10. Performance of ESM-1b in functional site annotation (see Supplementary Note 5 for details) for (a) Meizothrombin (PDB ID: 1A0H) at \( p = 5 \times 10^{-5} \) and (b) beta-glucuronidase (PDB ID: 3HN3) at \( p = 1 \times 10^{-4} \). In both cases, ESM-1b sequence embeddings result in lower sensitivity and more false positives than COLLAPSE embeddings. For meizothrombin, there are no correct predictions at a p-value of \( 5 \times 10^{-5} \). When the threshold is lowered to \( 1 \times 10^{-4} \), the serine and histidine active sites and the kringle domain are recognized, but at the cost of 10 false positive annotations. The catalytic aspartic acid is not identified. For beta-glucuronidase, one of the three active sites is correctly identified at \( p = 1 \times 10^{-4} \), but there are also three false positive annotations at a residue far from the active site.
### Supplementary tables

**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 |
| Pfam ID | Score |
|---------|-------|
| 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 | 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                                       | 5              | TYR        | OH              |
Table S3. Dataset statistics for PROSITE binary classification task.

| Dataset                  | Positive Examples | Negative Examples |
|--------------------------|-------------------|-------------------|
| EGF_1                    | 144               | 50,023            |
| TRYPSIN_SER              | 315               | 48,032            |
| RNASE_PANCREATIC         | 393               | 49,996            |
| EF_HAND_1                | 1,971             | 48,141            |
| IG_MHC                   | 2,188             | 49,073            |
| PROTEIN_KINASE_TYR      | 301               | 50,004            |
| TRYPSIN_HIS              | 445               | 47,476            |
| INSULIN                  | 438               | 49,074            |
| PROTEIN_KINASE_ST        | 1,199             | 49,999            |
| ADH_SHORT                | 383               | 50,128            |

Table S4. Performance on PROSITE FN/FP proteins compared to ESM-1b language model embeddings \(^6\) and MMseqs2. \(^7\) See Supplementary Notes 5–6 for details.

| Site   | PROSITE | COLLAPSE | ESM-1b | MMseqs2 | PROSITE total |
|--------|---------|----------|--------|---------|---------------|
| label  |         |          |        |         |               |

21
| Protein          | FN  | 8  | 5  | 14 |   |
|------------------|-----|----|----|----|---|
| ADH_SHORT        | 11  | 33 | 33 | 33 |   |
| EF_HAND_1        | 40  | 48 | 10 | 48 |   |
| EGF_1            | 60  | 61 | 1  | 90 |   |
| IG_MHC           | 21  | 0  | 2  | 47 |   |
| PROTEIN_KINASE_ST| 269 | 270| 0  | 271|   |
| PROTEIN_KINASE_TYR| 3   | 3  | 0  | 3  |   |
| TRYPSIN_HIS      | 10  | 9  | 1  | 16 |   |
| TRYPSIN_SER      | 9   | 4  | 4  | 4  |   |

**Table S5.** Performance of COLLAPSE on functional site identification benchmarks from Xin et al. (2011). Metric is AUROC, and highest-performing method is in bold. Performance of baseline methods is taken directly from the paper, so the exact splits used in each fold are not identical. See Supplementary Note 8 for details.
Table S6. ATOM3D benchmarking results compared to FEATURE for MSP dataset. The metric is area under the receiver operator characteristic curve (AUROC), and we report mean and standard deviation across three training runs. Numbers in bold indicate best performance on each task (within one standard deviation). See Supplementary Note 7 for details.

| Task (metric) | COLLAPSE (fixed) | COLLAPSE (fine-tuned) | FEATURE (fixed) |
|---------------|------------------|-----------------------|-----------------|
| MSP (AUROC)   | 0.616 ± 0.006    | 0.668 ± 0.018         | 0.617 ± 0.012   |

| COLAPSE       | 0.884            | 0.812                | 0.889           | 0.647 |
|---------------|------------------|----------------------|-----------------|-------|
| FEATURE       | 0.767            | 0.824                | 0.754           | 0.620 |
| GBT           | 0.713            | 0.713                | 0.814           | 0.559 |
| Graphlet kernel| 0.758          | 0.619                | 0.732           | 0.688 |
| Structure kernel | 0.808         | 0.800                | 0.839           | 0.711 |
References

1. Torng, W. & Altman, R. B. High precision protein functional site detection using 3D convolutional neural networks. *Bioinformatics* **35**, 1503–1512 (2019).

2. Anand, N. *et al.* Protein sequence design with a learned potential. *Nat. Commun.* **13**, 1–11 (2022).

3. Torng, W. & Altman, R. B. 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics* **18**, 302 (2017).

4. Jing, B., Eismann, S., Soni, P. N. & Dror, R. O. Equivariant Graph Neural Networks for 3D Macromolecular Structure. *arXiv [cs.LG]* (2021).

5. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).

6. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

7. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

8. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* (1983) doi:10.1002/bip.360221211.

9. Xin & Radivojac. Computational Methods for Identification of Functional Residues in Protein Structures. *Curr. Protein Pept. Sci.* **12**, (2011).

10. Hermosilla, P. *et al.* Intrinsic-Extrinsic Convolution and Pooling for Learning on 3D Protein Structures. *arXiv [cs.LG]* (2020).
11. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).