Towards functional annotation with latent protein language model features

Jake Silberg [*1] **Elana Simon** [*1] **James Zou** [1]

Abstract

Protein Language Models (PLMs) create high-dimensional embeddings that can be transformed into interpretable features using Sparse Autoencoders (SAEs), where each feature activates on specific protein patterns. However, scalably identifying which features are reliable enough for protein annotation remains challenging. We address this by developing a pipeline combining three complementary methods: (1) expanded database matching across 20+ annotation sources, (2) feature-guided local structural alignment to identify consistent activation regions, and (3) LLM-based feature description generation. Our annotation pipeline demonstrates three properties of SAE features that make them a useful source for functional annotation. First, they can represent more granular patterns than existing annotations, enabling the identification of sub-domains. Second, they can detect missing annotations by finding proteins that display recognizable structural motifs but lack corresponding labels. Here, we identify at least 491 missing CATH topology annotations with our pipeline. Third, they can maintain structural consistency across unseen proteins. Of our 10,240 SAE features, we find 615 that are structurally similar in unannotated metagenomic proteins, allowing us to match at least 8,077 metagenomic proteins to characterized proteins. This provides a rapid annotation pipeline with constant time search, that automatically includes structural and functional information about the feature that triggered the match.

1. Introduction

Functional annotation of uncharacterized proteins has historically relied on sequence conservation. Hidden Markov Models and Position-Specific Scoring Matrices form the foundation of major databases including CATH (Orengo et al., 1997) (via Gene3D (Buchan et al., 2002)) and Pfam (Bateman et al., 2004). These have been integrated into resources like UniProt (Consortium, 2015) and InterPro (Hunter et al., 2009), with tools like InterProScan enabling annotation of novel sequences (Jones et al., 2014). However, these methods struggle with divergent sequences, leading to specialized metagenomic databases like Novel Metagenomic Pfams (NMPfamsDB) (Baltoumas et al., 2024) that specifically curate sequences lacking Pfam annotations.

AlphaFold 2 (Jumper et al., 2021) catalyzed a shift toward 3D structural annotation. This enabled new databases of structural similarity, such as the Encyclopedia of Domains (Lau et al., 2024). However, large-scale structural search is computationally intensive. FoldSeek addresses this by converting structure into sequence matching using a 3D-informed alphabet. (van Kempen et al., 2022). Merizo-search uses embeddings trained on CATH for rapid matching (Kandathil et al., 2025). While powerful, these approaches have limitations: FoldSeek doesn't provide domain-specific functional information about the region(s) that triggered the match, and Merizo-search is constrained to identify hits based on supervised training on CATH.

Recent work has begun exploring sparse autoencoders (SAEs) for extracting interpretable features from protein language models. InterPLM (Simon & Zou, 2024) demonstrated correspondence between individual amino acid activations and UniProtKB annotations, while InterProt (Adams et al., 2025) associated protein-level SAE activations with Pfam families. However, these approaches achieved limited coverage, leaving over 75% of features unexplained when using stringent matching criteria.

Contributions

Our work extends this foundation by: (1) incorporating the full InterPro database across 20+ annotation sources including hierarchical codes, (2) developing structural validation through feature-guided local structural alignment, and (3) integrating LLM-based pattern recognition to identify features missed by existing databases. **This approach doubles annotation coverage while enabling missing annotation detection and novel protein characterization**.

[*]Equal contribution [1]Department of Biomedical Data Science, Stanford University. Correspondence to: Jake Silberg <jsilberg@stanford.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

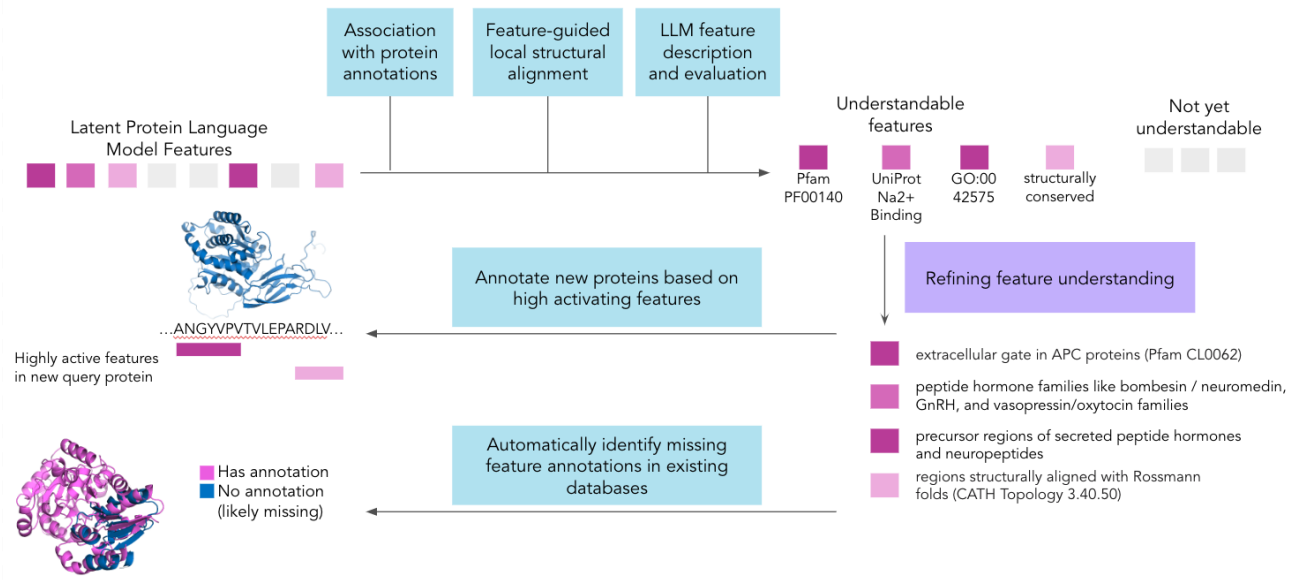


Figure 1. Workflow for protein-latent feature applications.

2. Using existing database annotations, local structural alignment, and LLMs to screen SAE features for cohesiveness

2.1. Combining protein annotation databases identifies structurally and functionally cohesive features

We expand annotation coverage by incorporating the full InterPro database (Hunter et al., 2009), which includes annotations from UniProtKB, Pfam, CATH / Gene3D (Orengo et al., 1997), (Buchan et al., 2002), and 19 other sources. We also include hierarchical codes such as Pfam clans and CATH topologies to capture broader biological concepts.

For each SAE feature, we sample up to 1100 proteins across 10 activation levels and test whether any single annotation code can predict high versus low feature activation, calculating F1 scores between predicted and actual activations. The highest-performing features are distributed relatively evenly across UniProtKB annotations, Pfam clans, Gene3D/CATH codes, and individual Pfam families, demonstrating that SAE features capture biological concepts at multiple levels of granularity. Features with moderate F1s (0.5-0.75) are more associated with UniProtKB features.

2.2. Feature-guided local structural alignment identifies structurally cohesive features that correlate with the presence of existing annotations

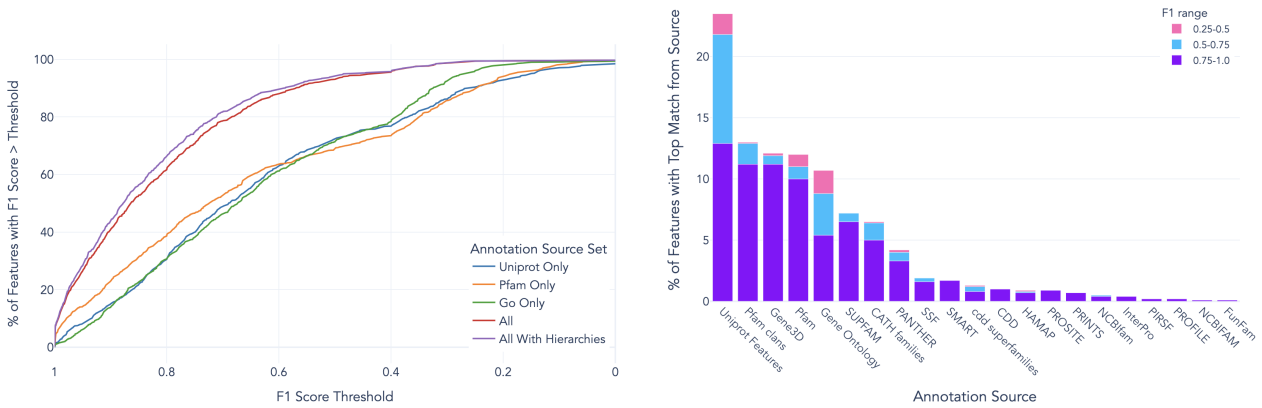
We introduce a procedure for Feature-Guided Local Structural Alignment to find "structurally cohesive" features, meaning the regions that activate highly have a consistent 3D structure. We sample 20 AlphaFold-predicted structures from the top activating proteins of a given SAE feature (activation above 0.7), after de-duplicating gene orthologs. We crop structures to 100 amino acids surrounding the peak activating amino acid, then run pairwise structural alignment, using backbone $RMSD_{100}$, a modification of Root Mean Square Deviation that allows for more lenience for longer alignments (Carugo & Pongor, 2001).

We find that better structural alignment is correlated (pearson r=0.53) with higher annotation code F1 scores. Specifically, 92% of features with $RMSD_{100} < 5$ have a code-based F1 $> .8$, while only 51% of features with an $RMSD_{100} > 5$ have a code-based F1 $> .8$. Thus, for pairwise alignments with low $RMSD_{100}$, we would expect they share a structural feature, even if one is missing an annotation, or is an uncharacterized novel protein.

2.3. Large Language Models identify additional cohesive features missed by other methods

While many features fire on a single existing database annotation code, others appear to fire on traits across annotations. Thus, asking LLMs to reason over protein data is a natural step in improving feature descriptions that can be used for protein annotation. We adapt the automated pipeline from InterPLM (Simon & Zou, 2024), using Claude-3.5 Sonnet (New) to generate descriptions using protein metadata from our annotation sources for 40 proteins at varying levels of maximum feature activation.

As validation, we test if these descriptions can predict fea-



(a) Percentage of features that match existing database annotations, by sources of annotations.

(b) Distribution of the common sources of the top annotation match per feature (highest feature F1 score).

Figure 2. Additional databases expand our ability to find cohesive features, and SAEs learn features at different levels of specificity.

ture activations on held-out proteins. We evaluate the ability of both the generated descriptions and the expanded annotation metadata to classify proteins as high or low activating for each feature, calculating F1 scores on a separate test set.

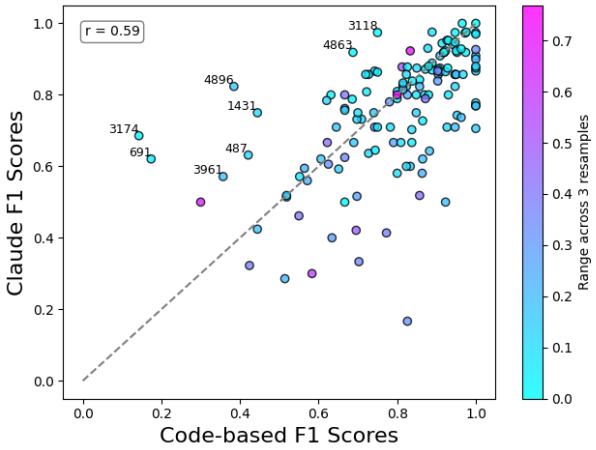


Figure 3. LLM-based F1 score vs. code-based F1 score on the same train/val split.

We find that the LLM's ability to describe a feature is highly correlated with the F1 score of the single best code, as shown in Figure 3.

Still, we find interesting cases where the LLM quantitatively improves performance. For example, for f/3174, the LLM identifies "this feature activates on transmembrane domains in multi-pass membrane proteins, particularly those involved in protein complex assembly and ion transport across membranes." This description, which focuses on transmembrane domains across a broader range of proteins than would be covered by a single Pfam family or clan, allows the description to outperform existing annotation codes.

We also analyze the 21 cases from our sampled features where the code-based F1 is at least .2 or more higher than the LLM-generated description's F1, that is, the LLM underperformed. In 20 of these cases, we find that precision was higher than recall. In fact, in 17 of the 21, precision was higher than .8 while for only 1 of the 21 cases was recall higher than .8. This indicates that the LLM is sometimes writing overly specific descriptions, so it is missing a broader view of what causes the feature to activate. For example, for f/657, the description is "This feature activates on catalytic domains of intradiol ring-cleavage dioxygenases and carboxypeptidases," which achieves an F1 of 0.5, but a precision of .875. The description is missing other elements besides these that also cause the feature to fire. In fact, this feature is highly associated with the Pfam clan CL0287, 7 stranded beta sandwiches. This does potentially point a way forward for the LLMs. Their focus on specificity is likely driven by having too few examples in order to fit into the context window. This suggests additional activating examples might help them identify the broadest possible applicable description.

3. Using PLM SAE features to enhance protein annotation

Having identified cohesive features using our approaches, we now show three benefits of utilizing these features.

3.1. SAE features capture granular subdomains interpretable through targeted literature search

While our annotation matching successfully identifies coherent features, it also reveals an important limitation: we

can identify the **types** of proteins where features activate, but not always what **specific elements within** those proteins cause the activation. This limitation, however, highlights a key advantage of SAE features—their ability to capture functional granularity beyond existing annotation schemes.

Figure 4 illustrates both this limitation and opportunity. Eight different features (f/253, f/515, f/1505, f/1579, f/1712, f/1731, f/2768, and f/3288) all achieve high F1 scores for the same Pfam clan annotation (APC clan, CL0062), yet each activates on distinct protein regions—different transmembrane alpha helices, cytoplasmic domains (f/253), and extracellular domains (f/515). While our annotation-based screening identifies these as coherent features, it cannot explain their specific functional roles, demonstrating both the power of SAE features to decompose protein families and the need for methods to interpret this finer granularity.

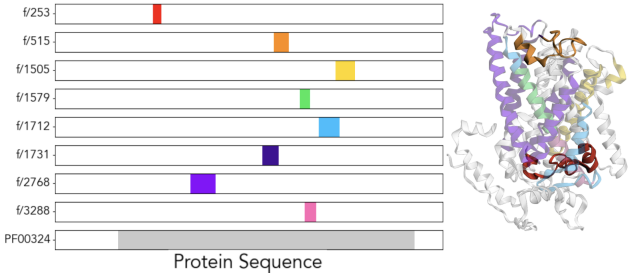


Figure 4. Eight SAE features correspond to the same Pfam clan (APC, CL0062) but activate on distinct structural components within GAP1_YEAST, including different transmembrane helices, cytoplasmic domains, and extracellular loops. Left: Each row shows feature activation along the length of the protein sequence with highly activated (> 0.8) residues highlighted in color, compared to the single Pfam domain (gray). Right: Feature activations on AlphaFold predicted structure (AFDB: P19145) showing each feature highlighting distinct structural components.

To bridge this gap, we provided OpenAI's o4-mini-high with the 10 highest-activating proteins for selected features, their gene descriptions from UniProtKB, and precise amino acid positions of peak activation, asking it to search for literature describing these specific regions. In a pilot evaluation of 6 features, this approach identified papers that discussed the exact regions highlighted by our features for 4 cases. For example, for f/515, the model retrieved literature on extracellular gating loops in APC proteins (Raba et al., 2014)—functional detail entirely absent from the broader clan annotation. An example prompt for this pilot, and more detail on the retrieved papers is in Appendix B.

3.2. SAE features can identify missingness in existing databases

Our local alignment procedure can also analyze features to find missing annotations. Specifically, we examine proteins that all activate the same SAE feature, but only some have the expected database annotation while others do not. When we measure $RMSD_{100}$ for alignments between annotated and unannotated proteins, they can be as good as RMSD scores for alignments between two proteins with the annotation, as shown in Figure 5. This suggests the unannotated proteins may just be missing labels.

To analyze this at scale, we reviewed the 20 proteins per feature randomly sampled in Section 2.2, looking at features with a code-based F1 of .8 or above for a CATH code or CATH topology. Across 221 features, we identify 1,055 of those top activating proteins in Swiss-Prot that are not tagged with a CATH code by Gene3D, but have strong local structural similarity to one of the proteins that does have a Gene3D annotation.

As external validation, we then compare the CATH tags from Gene3D, a sequence-based model, to TED (Lau et al., 2024), which uses a deep learning structure-based approach. We find that 491 of these proteins indeed do have a hit for that same CATH topology in TED (see Appendix Table 2).

While we can verify these seemingly missing annotations for CATH by using TED, there are SAE features that align with annotations from other databases like Pfam or UniProtKB. Thus, this combination of screening for existing annotations and local structural alignment can help identify potential missing annotations beyond just CATH.

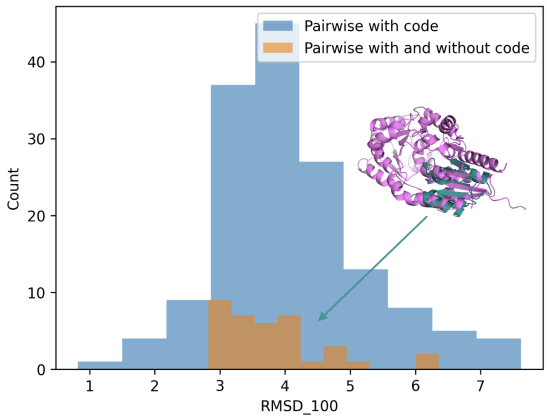


Figure 5. Histogram for f/401 of $RMSD_{100}$ for pairwise alignments. Aligning structures with the top existing annotation to structures without the annotation reveals features like f/401 with strong structural similarity despite different annotations. We show an example of an alignment for two proteins that activate on f/401 where one protein (Q0P9A8) has and one protein (Q5WSK6) lacks the CATH 3.40.50 annotation, the topology code for the Rossmann fold. For clarity, only the area around the feature is shown for Q5WSK6.

3.3. Features can rapidly detect structural matches in unannotated metagenomic proteins

A key advantage of feature-based annotation is that granular features can detect conserved domains even when full proteins show no sequence similarity to known families. This enables annotation of highly divergent metagenomic sequences that lack Pfam matches, which we test using NMPfamsDB (Baltoumas et al., 2024), a collection of metagenomic proteins that do not have any Pfam matches.

We find that many of our features activate highly in proteins the SAE was not trained on. Specifically, for over 50% of our features, there is at least one NMPfamsDB protein that activates that feature with a value ≥ 0.7. Then, by applying our local structural alignment procedure, we can identify 615 features with strong median local structural alignments ($RMSD_{100} < 5$ for pairwise alignments between one Swiss-Prot protein and one metagenomic protein) and 181 features with median $RMSD_{100} < 4$. This is strong evidence these features activate on the same structural element in both Swiss-Prot proteins and metagenomic proteins, as seen in Figure 6.

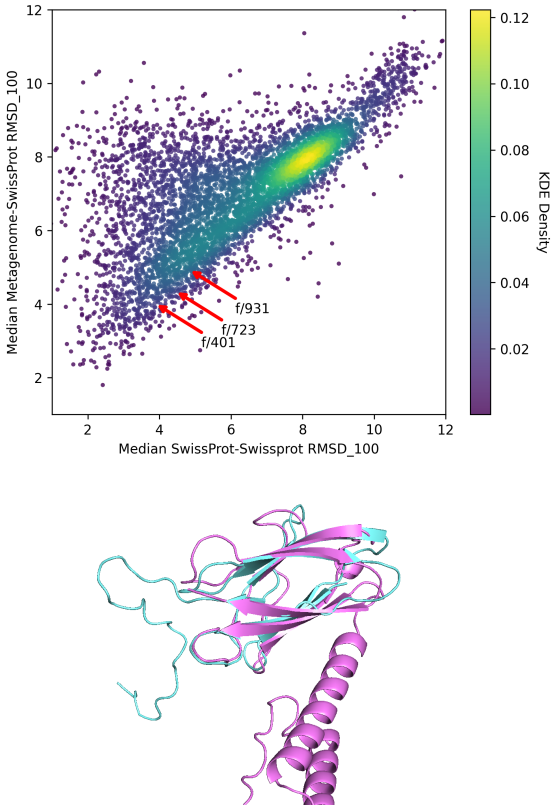


Figure 6. Median $RMSD_{100}$ per feature for comparisons within Swiss-Prot or between Swiss-Prot and metagenomic proteins (top), with local structural alignment example for f/931 between Swiss-Prot protein Q9M9H7 and metagenomic protein F014369 (bottom).

Using just these 615 features allows us to find matches between 12,526 metagenomic proteins in NMPfamsDB and Swiss-Prot proteins with an $RMSD_{100} \leq 5$ (14.9% of the 83,878 metagenomic proteins in NMPfamsDB we analyzed). At an $RMSD_{100}$ threshold of 4 or better, we can find matches for 8,077 metagenomic proteins (9.6%).

4. Conclusion

In this work, we demonstrate the potential benefits of using latent SAE features from protein language models for protein annotation. We find features that consistently activate on local elements of proteins. We also find features that identify missing database annotations at scale, and features that allow us to characterize unannotated metagenomic proteins. This annotation workflow automatically returns not only a structurally similar characterized protein, but also structural or functional information about exactly the aligned region.

We note several limitations of the work: first, while LLMs can help, identifying where within proteins these features fire remains primarily a manual task due to hallucinations and the need for verification. Second, our structural approach requires conservation within 100 amino acids of peak activation, which likely misses features that span distant regions or involve flexible motifs. We discuss one example flexible motif in Appendix A. Third, for now we rely on a single layer (Layer 18) of ESM-2 650M and focus our annotation validation on Swiss-Prot proteins, which increases the density of annotations within structural matches but limits the space of proteins we can match.

Future work should expand this analysis across multiple layers within PLMs and evaluate additional protein embedding models. Additionally, other methods for clustering PLM features should be considered, with this work showing valuable tasks they may be able to perform. We also expect advances in LLM capabilities and more advanced protein representations may improve further on these tasks, and hope this framework can provide a useful proof-of-concept as this field develops further.

Code availability

Code is available at `https://github.com/ElanaPearl/interp-agents`

Impact statement

This paper presents work whose goal is to advance the field of protein annotation. There are many potential societal consequences of our work, such as helping biologists better understand and annotate novel features, or discover new functional domains in proteins. We do not see significant or specific ethical risks with this work, though we suppose it is possible that better understanding (and eventually designing) proteins could be used maliciously in rare circumstances.

Acknowledgments

The authors wish to thank Wei Deng, Jessica Karaguesian, Brian Trippe, and Ben Viggiano for helpful conversations that improved this work. JS is supported by the Arc Institute. ES is supported by NSF GRFP grant DGE-2146755.

References

Adams, E., Bai, L., Lee, M., Yu, Y., and AlQuraishi, M. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, pp. 2025–02, 2025.

Baltoumas, F. A., Karatzas, E., Liu, S., Ovchinnikov, S., Sofianatos, Y., Chen, I.-M., Kyrpides, N. C., and Pavlopoulos, G. A. Nmpfamsdb: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Research*, 52(D1):D502–D512, 2024.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., et al. The pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004.

Buchan, D. W., Shepherd, A. J., Lee, D., Pearl, F. M., Rison, S. C., Thornton, J. M., and Orengo, C. A. Gene3d: structural assignment for whole genes and genomes using the cath domain structure database. *Genome research*, 12(3): 503–514, 2002.

Carugo, O. and Pongor, S. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein science*, 10(7):1470–1473, 2001.

Consortium, U. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.

Cosgriff, A. J. and Pittard, A. A topological model for the general aromatic amino acid permease, arop, of escherichia coli. *Journal of bacteriology*, 179(10):3317–3323, 1997.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. Interpro: the integrative protein signature database. *Nucleic acids research*, 37(suppl_1): D211–D215, 2009.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kandathil, S. M., Lau, A. M., Buchan, D. W., and Jones, D. T. Foldclass and merizo-search: Scalable structural similarity search for single-and multi-domain proteins using geometric learning. *Bioinformatics*, pp. btaf277, 2025.

Lau, A. M., Bordin, N., Kandathil, S. M., Sillitoe, I., Waman, V. P., Wells, J., Orengo, C. A., and Jones, D. T. Exploring structural diversity across the protein universe with the encyclopedia of domains. *Science*, 386(6721):eadq4946, 2024.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Raba, M., Dunkel, S., Hilger, D., Lipiszko, K., Polyhach, Y., Jeschke, G., Bracher, S., Klare, J. P., Quick, M., Jung, H., et al. Extracellular loop 4 of the proline transporter putp controls the periplasmic entrance to ligand binding sites. *Structure*, 22(5):769–780, 2014.

Simon, E. and Zou, J. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, pp. 2024–11, 2024.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Gilchrist, C. L., Söding, J., and Steinegger, M. Foldseek: fast and accurate protein structure search. *Biorxiv*, pp. 2022–02, 2022.

A. Flexible regions

We note that many features still have a high code-based F1 even though they also have high $RMSD_{100}$. Some of these are regions that are structurally flexible, for example f/73 below. We show three proteins, each with a homeodomain-like region highlighted in orange and blue, and the highly activated region for f/73 in pink.

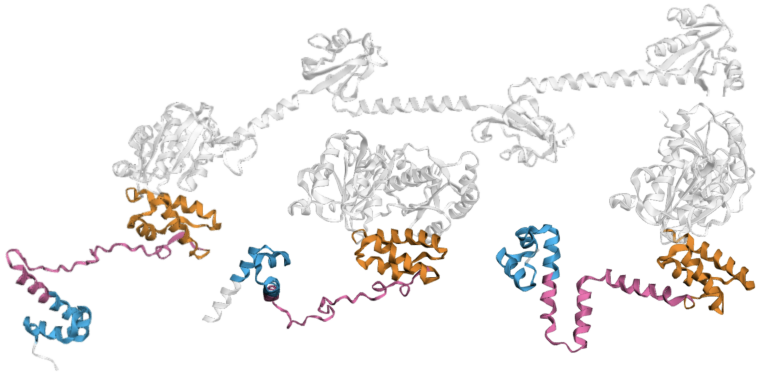


Figure 1. f/73 captures a variety of structures that all link two structurally consistent domains. Here, three proteins have homeolike-domains are in orange and blue, while the highly active region for f/73 is in pink

B. Understanding granular features

For six features, we used o4-mini-high to attempt to find citations that discussed the specific regions of interest. An example prompt is given below, followed by the best citation (where applicable) found by the model. Each citation returned by the model was reviewed manually, as the model could sometimes hallucinate specific quotes or mutations that were not found in the underlying papers. Where no citation was retrieved, the feature was analyzed manually to determine its specific function. Sometimes very promising literature about the specific region exists, but was not returned by o4-mini-high, perhaps because of only asking about up 10 gene-species combinations per feature. For example, manually searching for additional papers revealed that f/253 corresponds to the cytoplasmic loop between transmembrane domain 2 and 3 in amino acid permeases (Cosgriff & Pittard, 1997).

Prompt for f/515

What do we know about the following proteins in these amino acid regions mentioned for each? Cite papers that look specifically at or very near these regions
ACTP_PECCP Cation/acetate symporter ActP (Acetate permease) (Acetate transporter ActP) in Pectobacterium carotovorum subsp. carotovorum (strain PC1) at 332
MNTH_AGRFC Divalent metal cation transporter MntH in Agrobacterium fabrum (strain C58 / ATCC 33970) (Agrobacterium tumefaciens (strain C58)) at 309
...
KUP_CYTH3 Probable potassium transport system protein Kup in Cytophaga hutchinsonii (strain ATCC 33406 / DSM 1761 / CIP 103989 / NBRC 15051 / NCIMB 9469 / D465) at 271
KUP_STRA1 Probable potassium transport system protein Kup in Streptococcus agalactiae serotype Ia (strain ATCC 27591 / A909 / CDC SS700) at 277
PUTP_STAAN Sodium/proline symporter (Proline permease) in Staphylococcus aureus (strain N315) at 310
In total (across all proteins) provide me only with 2-4 of the best matching citations and what we know about the region from each paper

Table 1. o4-retrieved feature citations for specific sub-domains.

Feature	o4 description of best citation	Paper link
73	No citation retrieved discussed this specific subdomain (a flexible linker region sandwiched between HTH domains).	—
253	No citation retrieved discussed this specific subdomain (in the cytoplasmic loop between TM2 and TM3).	—
401	"Schmidt *et al.* solved crystal structures of the *Aquifex aeolicus* Kdo transferase (WaaA), a GT–B family homolog, revealing that the loop encompassing residues 98–102 (equivalent to *E. coli* position ∼101) shapes the acceptor–substrate binding site…"	https://pubmed.ncbi.nlm.nih.gov/22474366/
515	"Site-directed mutagenesis targeting extracellular loop 4 of *S. aureus* PutP (residues 310, 314, 318) showed that altering the amino acid at position 310 completely abolishes proline uptake…"	https://www.sciencedirect.com/science/article/pii/S0969212614000835?utm_source=chatgpt.com
711	"In vivo cysteine cross-linking between TM2 and TM8—including sites around residue 316—demonstrated that MurJ adopts both inward- and outward-facing states during transport. Disruption of membrane potential selectively destabilized the inward-facing conformation..."	https://pubmed.ncbi.nlm.nih.gov/30482840/
931	"Residue 88 (human numbering) lies in the β-strand F of the TTR fold, forming part of a critical hydrogen-bond network that stabilizes the tetramer…"	https://pmc.ncbi.nlm.nih.gov/articles/PMC8122960/

C. Missing CATH annotations

Below we show the first 100 missing CATH annotations identified by our workflow, of the 491 that match with TED annotations. We considered a match if TED contained a code matching the top CATH code for a given feature that was within the same topology. Best code represents the top CATH code for that feature, while TED label is the exact TED label for that protein (which may either be a CATH homologous superfamily or topology).

Table 2. Rows 1–50

Protein	Feature	Best code	TED label
Q01473	4	1.20.120.160	1.20.120
Q7Y0V9	52	3.30.530.20	3.30.530.20
Q0WV12	52	3.30.530.20	3.30.530.20
Q9FVI6	52	3.30.530.20	3.30.530.20
Q8XA02	57	3.10.105.10	3.10.105.10
P46890	57	3.10.105.10	3.10.105.10
P77172	81	3.30.70.270	3.30.70.270
Q58121	112	1.20.58.340	1.20.58
O28044	112	1.20.58.340	1.20.58
Q01473	119	1.20.120.160	1.20.120
Q49430	121	1.20.1560.10	1.20.1560.10
Q0VFX2	129	1.20.5.500	1.20.5
Q5BL57	129	1.20.5.500	1.20.5
Q499U4	129	1.20.5.500	1.20.5
Q25C79	129	1.20.5.500	1.20.5
Q4UMJ9	159	1.20.1250.20	1.20.1250.20
P28246	159	1.20.1250.20	1.20.1250.20
Q9JXM5	169	1.25.40.10	1.25.40.10
O51072	180	1.25.40.10	1.25.40.10
Q8K4P7	180	1.25.40.10	1.25.40.10
Q8RWN0	194	3.40.395.10	3.40.395.10
Q0WKV8	194	3.40.395.10	3.40.395.10
Q09275	194	3.40.395.10	3.40.395.10
O13769	194	3.40.395.10	3.40.395.10
Q8L7S0	194	3.40.395.10	3.40.395.10
Q2PS26	194	3.40.395.10	3.40.395.10
Q54KW6	196	1.25.40.10	1.25.40
P34511	218	1.25.40.20	1.25.40.20
P18540	251	3.40.50.2300	3.40.50.2300
O22232	281	3.80.10.10	3.80.10
O04615	310	2.60.210.10	2.60.210.10
Q9FKD7	310	2.60.210.10	2.60.210.10
Q9XHZ8	310	2.60.210.10	2.60.210.10
Q8C008	371	2.60.40.10	2.60.40
Q9STL8	399	3.40.140.10	3.40.140.10
Q9FG71	399	3.40.140.10	3.40.140.10
O82264	399	3.40.140.10	3.40.140.10
Q9LYC2	399	3.40.140.10	3.40.140.10
Q04368	399	3.40.140.10	3.40.140.10
P76349	401	3.40.50.2000	3.40.50.2000
D3DJ42	401	3.40.50.2000	3.40.50.20
Q5WSK6	401	3.40.50.2000	3.40.50.2000
Q9SH31	419	3.40.50.2000	3.40.50
Q3E9A4	419	3.40.50.2000	3.40.50.2000
Q9LFP3	434	3.40.50.2000	3.40.50.2000
P75207	442	1.20.1560.10	1.20.1560.10
Q49430	443	1.20.1560.10	1.20.1560.10
Q9ZD06	449	3.30.2350.10	3.30.2350.10
Q8FB47	449	3.30.2350.10	3.30.2350.10
Q8ZGM2	449	3.30.2350.10	3.30.2350.10

Table 3. Rows 51–100

Protein	Feature	Best code	TED label
Q92HG4	449	3.30.2350.10	3.30.2350.10
P77607	544	3.40.50.1390	3.40.50.1390
Q8XA02	607	3.10.105.10	3.10.105.10
P46890	607	3.10.105.10	3.10.105.10
A9NCA7	713	3.40.50.2000	3.40.50.2000
O67214	713	3.40.50.2000	3.40.50.2000
P37597	751	1.20.1250.20	1.20.1250.20
O04292	789	3.30.450.20	3.30.450.20
Q9Z7F1	789	3.30.450.20	3.30.450.20
A2XBL9	789	3.30.450.20	3.30.450.20
Q39123	789	3.30.450.20	3.30.450.20
B0K165	843	3.30.479.30	3.30.479.30
P32233	848	3.40.50.300	3.40.50.300
Q2NL82	848	3.40.50.300	3.40.50.300
A0A0H2URH2	849	3.40.50.2000	3.40.50.2000
A0A0H2URJ6	849	3.40.50.2000	3.40.50.2000
P33694	849	3.40.50.2000	3.40.50.2000
A1JSF2	873	3.40.50.300	3.40.50.300
P50837	889	3.30.420.10	3.30.420.10
Q60953	889	3.30.420.10	3.30.420.10
P53296	891	3.30.559.10	3.30.559.30
I1S097	894	3.90.550.10	3.90.550.10
D3ZZN9	936	2.60.40.150	2.60.40.150
Q9C8E6	936	2.60.40.150	2.60.40.150
P54739	972	3.30.200.20	3.30.200.20
P54735	972	3.30.200.20	3.30.200.20
C4JDF8	1003	1.10.1200.10	1.10.1200.10
P39404	1013	3.40.50.2300	3.40.50.2300
O83933	1016	1.20.1600.10	1.20.1600
P63400	1039	1.10.1760.20	1.10.1760
O67248	1039	1.10.1760.20	1.10.1760
Q58299	1053	3.60.40.10	3.60.40.10
O29259	1119	1.10.443.10	1.10.443.10
P07261	1119	1.10.443.10	1.10.443.20
O83202	1205	3.30.70.270	3.30.70.270
P77172	1205	3.30.70.270	3.30.70.270
Q2NKC0	1205	3.30.70.270	3.30.70.270
Q10419	1233	2.40.30.170	2.40.30.170
P55501	1239	3.30.420.10	3.30.420.10
Q44493	1320	2.150.10.10	2.150.10.10
P75800	1329	3.30.70.1230	3.30.70.270
B5XZP2	1347	1.20.1250.20	1.20.1250.20
D0CCT2	1347	1.20.1250.20	1.20.1250.20
E0T2N0	1347	1.20.1250.20	1.20.1250.20
O34353	1360	2.120.10.80	2.120.10.30
Q10412	1380	1.20.920.10	1.20.920
Q6YRK2	1416	3.30.70.270	3.30.70.270
B0XZV4	1447	1.20.1250.20	1.20.1250.20
O83837	1466	1.25.40.10	1.25.40
Q8IAR5	1482	2.30.42.10	2.30.42.10