Evaluating Multi-Modal Models for Enzyme-Reaction Retrieval

Annika Viswesh [1] **Jason Yang** [2] **Frances Arnold** [2] **Yisong Yue** [1]

Abstract

Identifying functional enzymes that can perform unannotated reactions is a major biotechnological bottleneck. While multi-modal machine learning models can be used to retrieve enzymes given target functions (reactions), existing methods have not been adequately compared to each other. Two key areas warrant further investigation: first, the optimal way to incorporate 3D protein structure and predicted binding pockets for enzyme retrieval, and second, the most effective learning objectives for training such multi-modal models. We examine these questions through experiments on Task 2 of Classification and Retrieval for Enzymes (CARE) benchmark, demonstrating that multi-modal representations combining protein structure with pocket information have better performance than sequence-only methods. Second, we evaluate learning objectives and found that contrastive learning generally provides superior performance for enzyme retrieval compared to a binary classification. Our work underscores the value of integrating structural and pocket information for precise enzyme-reaction matching and offers insights into effective training objectives for such retrieval models.

1. Introduction

Enzymatic catalysis underlies virtually every cellular process and powers a broad array of industrial biotechnologies. Many enzymes are useful as biocatalysts or can be engineered for industrially relevant applications (Buller et al., 2023; Reisenbauer et al., 2024), but the functional landscape of enzymes remains only partially charted (Breaker, 1997; Knowles, 1991; Arnold, 2018; Chen & Arnold, 2020). While approximately 246 million protein sequences are now available through sequencing efforts, fewer than 1% have been annotated with functional information, and only 0.23% of UniProt entries are well-studied (Consortium, 2025; Ribeiro et al., 2023). In these databases, a common formalism for categorizing enzymatic activity is the Enzyme Commission (EC) number, a four-level numerical identifier that hierarchically encodes the chemistry of the catalyzed reaction (Sanderson et al., 2023; Li et al., 2018; Dalkiran et al., 2018). Since many protein sequences are not annotated with EC numbers and many reactions do not fall under existing EC numbers, retrieving a protein sequence given a specified biochemical reaction can be a challenging task, necessitating new computational methods for this task (Yang et al., 2024a).

Recent multi-modal machine-learning efforts have tackled these problems of enzyme function prediction and retrieval. These distinct modalities can include protein sequence information, 3D protein structure, localized functional site information (such as binding pockets), and chemical reaction representations (e.g., SMILES strings or fingerprints). Contrastive Reaction-Enzyme Pretraining (CREEP) (Yang et al., 2024b) uses language models ProtT5 (Elnaggar et al., 2021) and rxnfp (Schwaller et al., 2021) to jointly align and learn sequence-only protein representations with reaction representations. CLIPZyme (Mikhael et al., 2024) contrastively learns and aligns an $\mathrm{E}(n)$-equivariant GNN over 3-D protein graphs with a *de novo* message-passing network that encodes 2-D reaction graphs. Other methods like EnzymeCAGE (Liu et al., 2024) focus directly on the catalytic pocket and are trained with a classification loss.

On “Task 2” of the CARE benchmarks, which focuses on retrieving a protein given an unannotated reaction, CREEP has demonstrated the highest performance–potentially because encoders for different modalities are finetuned but also potentially because the training process for CLIPZyme was not optimized in this case (Yang et al., 2024b). Thus, fundamental questions still remain, as CREEP, CLIPZyme, and EnzymeCAGE have not been fairly compared to each other. First, the role and impact of explicitly incorporating additional modalities like three-dimensional protein structure and predicted binding pockets on enzyme annotation performance, particularly compared to sequence-only representations utilized by CREEP, remain to be fully understood.

[1] Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena CA 91125 [2] Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena CA 91125. Correspondence to: Frances Arnold <farnold@caltech.edu>, Yisong Yue <yyue@caltech.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

Second, identifying the most effective learning objective (contrastive versus classification loss) for combining multiple modalities for retrieval has not been explored.

To fairly compare multimodal frameworks for enzyme retrieval, we modify the CREEP framework to accommodate different representations of proteins and loss objectives (Figure 1). We demonstrate that structure-based modality performs better than sequence-based modality for protein function prediction, and adding pocket information to the structure-based representations achieves even better performance. Additionally, we present a systematic evaluation of learning objectives for enzyme-reaction retrieval models, showing that contrastive learning outperforms binary classification. Overall, our findings provide valuable insights into model design choices for enzyme function retrieval.

2. Methods

Our objective is to retrieve protein sequences given target, unseen reactions and generate a ranked list of Enzyme Commission (EC) numbers that correspond to those retrieved proteins (Task 2 of CARE benchmark) (Yang et al., 2024b). Our models address this multi-modal task by leveraging protein information (from sequence, 3D protein structure, and integrating structure with pocket-level information) and chemical reaction representation.

All of our models are trained in a similar fashion to the original CREEP model, but we modify how protein representations from these different modalities are encoded within this framework. For models utilizing the protein sequence modality, protein representations are derived using fixed ESM-2 embeddings (Lin et al., 2022) processed through a non-linear multi-layer perceptron network. For models based on the 3D protein structure modality, we follow the approach used in CLIPZyme. We encode each protein as a three-dimensional graph $G = (V, E)$. Each node $v_i \in V$ has distinct coordinates $c_i \in \mathbb{R}^3$ and corresponds to residue-level information, while each edge $e_{ij} \in E$ represent bonds between residues. Initial node features are derived from ESM-2 embeddings (Lin et al., 2022) pretrained with 150 M parameters, yielding, for each residue, a 640-dimensional feature vector. The protein graph features are processed using an Equivariant Graph Neural Network (EGNN) (Satorras et al., 2021). For all models, we encoded reactions as SMILES/SMARTS strings (Weininger, 1988) and consequently finetuned using rxnfp, a BERT-style model (Devlin et al., 2019) that generates chemical fingerprints. Additional details of all models are provided in Appendix C.

To further enhance enzyme retrieval by incorporating binding pocket information as an additional modality to structure, we explored integrating active site features to the structure by leveraging a state-of-the-art pocket prediction method, P2Rank (Kriválk & Hoksza, 2018; Jendele et al., 2019), to identify potential ligand-binding sites and assign them calibrated probability and raw scores. The detailed methodology of how P2Rank generates these pocket-level scores (S_j) and probabilities ($P(S_j)$) from protein structures is provided in Appendix D.

We explored three distinct strategies for integrating these residue-level pocket features with the ESM-2 embeddings as listed below. Detailed pocket integration strategies are provided in Appendix E.

1. **Direct Concatenation (‘Structure + Pockets’):** This entails concatenating raw P2Rank-derived pocket probability and score features directly to the initial residue embeddings.

2. **Adapted Pockets (‘Structure + Adapted Pockets’):** This involves projecting the normalized residue-level pocket probability into a learnable dense vector, before concatenation with residue embeddings.

3. **Pocket-Weighted Message Passing (‘Structure + Weighted Pockets’):** This strategy modulates the message aggregation process within the EGNN layers using residue-level pocket probabilities as attention-like weights.

3. Results

We evaluated our multi-modal models on reaction-to-enzyme retrieval tasks with contrastive loss. Evaluations were conducted on easy, medium, and hard splits from a modified version of the Task 2 CARE benchmark (Yang et al., 2024b), for which protein structure representations were curated based on AlphaFold-predicted structures. Dataset details can be found in Appendix A, Figure 2, and Table 3. CREEP retrieves protein sequences for a query reaction, which are then mapped to a ranked list of EC numbers for further evaluation, details of which can be found in Appendix B. Retrieval performance is evaluated using the Top-k Success Rate at $k = 1$, where the top k EC numbers from the ranking are selected, and the accuracy of the most correct EC number is provided at four EC hierarchy levels.

3.1. Structural and Pocket Features Enhance Enzyme Retrieval

Our findings, summarized in Table 1, reveal a clear hierarchy of performance on the modified CARE Task 2 dataset. First, utilizing ‘Structure’-based representations leads to better performance than ‘Sequence’-only representations across the easy, medium, and hard data splits, particularly at the most specific EC level 4. These findings strongly suggest that structural embeddings provide a more infor-

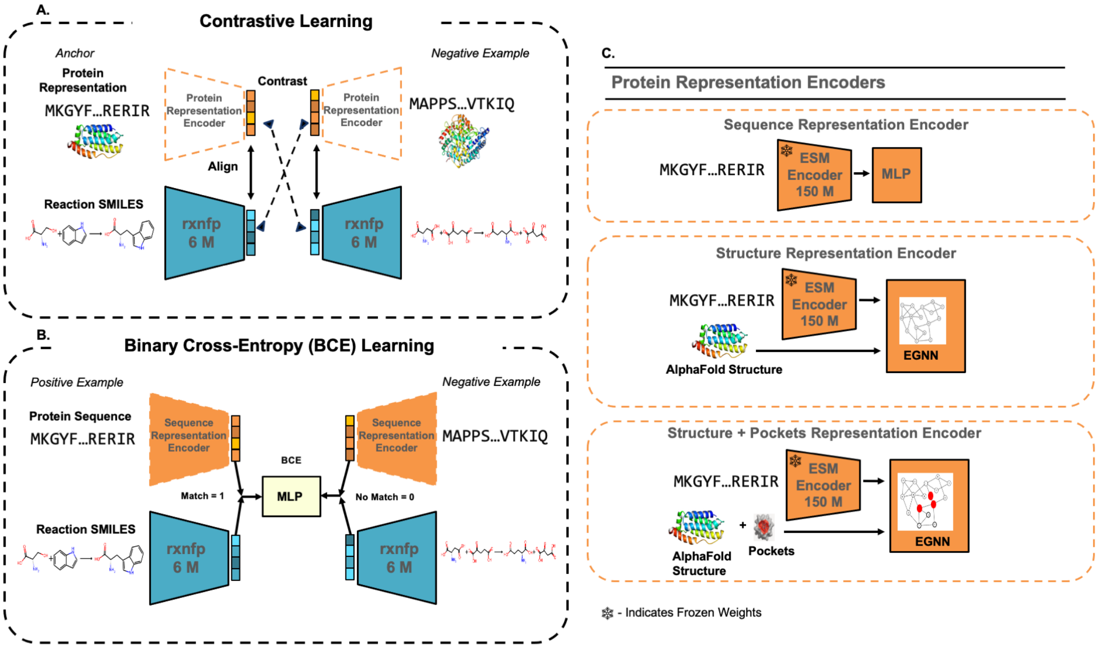

Figure 1. Overview of different multi-modal architectures for enzyme-reaction representation learning, illustrating the integration of various modalities. (A) Contrastive learning aligns embeddings from different protein and reaction modalities. (B) A Binary MLP predicts if a given protein-reaction pair matches, using embeddings derived from their respective modalities. (C) Protein representation encoders for different input modalities: (1) Protein sequence modality processed by an ESM-2 based encoder; (2). 3D Protein structure modality processed by an EGNN; (3) Integrated 3D protein structure and binding pocket modalities processed by an EGNN.

mative representation for enzyme retrieval than sequence embeddings alone.

Second, explicitly incorporating predicted binding pocket information via direct concatenation (‘Structure + Pockets’) yielded further performance gains over the ‘Structure-only’ model. While the ‘Structure + Pockets’ model did not surpass the ‘Structure-only’ model in every single metric, showing a decrease of 0.2% in retrieval accuracy for the Medium split at EC Level 4, it demonstrated superior performance across all the other splits and EC hierarchy levels, achieving the best overall results among the tested configurations, with an average increase of 3.38% in retrieval accuracy across all splits and EC Levels. This highlights the significant value of including localized functional site information as a valuable signal for more accurate enzyme retrieval. We then explored if more sophisticated methods for integrating pocket features—specifically an “adapted” variant (projecting pocket probabilities) and a “weighted” variant (pocket-weighted message passing in the EGNN)—could further amplify these benefits. However, neither of these more complex strategies consistently outperformed the simpler direct concatenation used in ‘Structure + Pockets’. The “adapted” variant demonstrated lower performance compared to simple concatenation, while the “weighted” message-passing approach generally underperformed relative to both the standard concatenation and the adapted integration.

These results indicate that while explicit pocket information is beneficial, the simple concatenation of the raw probability and score signal proved most effective in our setup compared to the explored adaptation or weighted message passing strategies. One potential reason for the underperformance of the adapted and weighted variants could be that the transformation or weighting processes might have inadvertently obscured more important fine-grained information present in the raw pocket signal. Alternatively, performance for the adapted and weighted variants can be improved through hyperparameter tuning.

3.2. Contrastive Learning Outperforms Binary Classification

We compared our primary EBM-NCE contrastive learning (Liu et al., 2022) approach against models trained with binary classification (BCE loss) as the objective, for sequence-based representations. The contrastive setup aims to learn a joint embedding space where query reactions and their corresponding target enzymes are brought closer, while pushing non-target enzymes further away; the binary classification approach frames the task as predicting individual reaction-enzyme relationships. In both cases, we loaded the same positive and negative pairs as training data. More details on the contrastive and binary classification objectives used can be found in Appendix F and Appendix G, respectively.

Table 1. Enzyme retrieval performance by different protein modalities, split, and hierarchy levels on modified CARE Task 2 dataset under the EBM-NCE objective. Performance is measured as k=1 retrieval accuracy (%). Bolded accuracy is the best model.

Split	Protein Representation	Level 4 (X.X.X.X)	Level 3 (X.X.X.-)	Level 2 (X.X.-.-)	Level 1 (X.-.-.-)
Easy	Structure	37.3	62.0	78.7	88.7
	Structure + Pockets	**38.8**	**67.6**	**83.8**	**92.0**
	Structure + Adapted Pockets	34.4	59.9	74.3	88.7
	Structure + Weighted Pockets	31.9	63.5	76.6	90.5
	Sequence	15.7	38	58.1	79.2
Medium	Structure	**4.1**	33.4	51.4	78.1
	Structure + Pockets	3.9	**37.0**	**59.6**	**79.7**
	Structure + Adapted Pockets	2.3	35.2	56.3	77.9
	Structure + Weighted Pockets	2.3	28.5	47.3	75.6
	Sequence	2.3	18.3	39.1	66.1
Hard	Structure	1.1	6.3	17.6	50.0
	Structure + Pockets	**2.6**	**9.1**	**18.5**	**56.7**
	Structure + Adapted Pockets	1.1	7.6	16.5	51.7
	Structure + Weighted Pockets	0.9	7	17.2	49.3
	Sequence	0.4	4.6	15.9	49.8

Table 2. Enzyme retrieval performance by split, loss functions, and hierarchy Levels on modified CARE Task 2 dataset, for the sequence-based modality. Performance is measured as k=1 retrieval accuracy (%). Bolded accuracy is the best model.

Split	Loss Function	Level 4 (X.X.X.X)	Level 3 (X.X.X.-)	Level 2 (X.X.-.-)	Level 1 (X.-.-.-)
Easy	EBM-NCE	**15.7**	**38**	**58.1**	**79.2**
	BCE	3.1	15.4	37.5	69.9
Medium	EBM-NCE	**2.3**	**18.3**	**39.1**	**66.1**
	BCE	1	14.9	35.5	63.8
Hard	EBM-NCE	**0.4**	**4.6**	**15.9**	**49.8**
	BCE	**0.4**	3.9	7.2	29.3

As shown in Table 2, the EBM-NCE model generally demonstrated superior $k = 1$ retrieval accuracy compared to the BCE model, particularly on the 'Easy' and 'Medium' data splits across most EC hierarchy levels. However, on the 'Hard' split, the performance between the two loss functions was more nuanced at the finer-grained EC levels. Specifically, the BCE model achieved the same performance of in retrieval accuracy as the EBM-NCE model at Level 4 on this split. However, the EBM-NCE model remained superior for Levels 1, 2, and 3. Given the substantially weaker performance of the BCE model we did not proceed with evaluating the binary loss function on our more computationally intensive structure-based models.

4. Discussion

Our study primarily investigated two design choices for multi-modal enzyme retrieval: utilizing structural and pocket features as part of protein representations, and comparing learning objectives. Overall, the better performance of our 'Structure + Pockets' model over both sequence-only and structure-only approaches solidifies the importance of 3D conformational and functional site information. Similarly, better performance of contrastive learning over binary cross-entropy for sequence-based retrieval aligns with the broader success of contrastive methods in learning discriminative embeddings for complex data. Taken together, these findings point towards a promising direction for enzyme function prediction.

While our 'Structure + Pockets' model achieved best performance on the modified CARE benchmark, its effectiveness is tied to the quality of external tools like AlphaFold and P2Rank. The output of these tools can vary, which thereby affects how well our current models work across all protein families.

Future directions of our work include the following. First, while our study highlights the value of structural protein representations, the role of the reaction modality can be further explored. To address this, our future work will include a reaction modeling ablation study that will examine alternative reaction encodings and investigate whether structure-based graph representations of reactions can yield similar gains (Mikhael et al., 2024). Second, we aim to test ESM-2 finetuning to evaluate if it can improve performance for all models. Beyond the CARE benchmark, we plan to evaluate the generalization capabilities of our models on prominent, useful enzyme classes such as Cytochromes P450, phosphatases, and terpene synthases (Liu et al., 2024). Finally, to strengthen the functional relevance of our retrieval framework, we aim to perform wet-lab validation of top predictions, to discover real enzymes with unannotated activities. By testing a subset of high-confidence model predictions in vitro, we seek to bridge the gap between in silico retrieval and real-world enzymatic and in doing so aim confirm the biological validity of our multi-modal approach.

5. Acknowledgments

This work was supported by a U.S. Army Research Office cooperative agreement (W911NF-19-2-0026 to F.H.A.) and an Amgen Chem-Bio Engineering award. The authors thank Ariane Mora, Bryan Wang, and Tanvi Ganapathy for feedback on this work and help with understanding the CARE benchmarks.

References

Arnold, F. H. Directed evolution: Bringing new chemistry to life. *Angewandte Chemie International Edition*, 57(16): 4143–4148, April 2018. ISSN 1433-7851, 1521-3773. doi: 10.1002/anie.201708408.

Atz, K., Isert, C., Böcker, M. N., Jiménez-Luna, J., and Schneider, G. δ-quantum machine-learning for medicinal chemistry. *Physical Chemistry Chemical Physics*, 24(18): 10775–10783, 2022.

Breaker, R. R. Dna enzymes. *Nature Biotechnology*, 15(5): 427–431, 1997.

Buller, R., Lutz, S., Kazlauskas, R. J., Snajdrova, R., Moore, J. C., and Bornscheuer, U. T. From nature to industry: Harnessing enzymes for biocatalysis. *Science*, 382(6673):eadh8615, November 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adh8615. URL https://www.science.org/doi/10.1126/science.adh8615.

Chen, K. and Arnold, F. H. Engineering new catalytic activities in enzymes. *Nature Catalysis*, 3(3):203–213, January 2020. ISSN 2520-1158. doi: 10.1038/s41929-019-0385-5.

Consortium, T. U. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, January 2025. doi: 10.1093/nar/gkae1010.

Dalkiran, A. et al. Ecpred: a tool for the prediction of the enzymatic functions of protein sequences based on the ec nomenclature. *BMC Bioinformatics*, 19:1–13, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. 14(8):29, 2021.

Jendele, L., Krivak, R., Skoda, P., Novotny, M., and Hoksza, D. Prankweb: a web server for ligand binding site prediction and visualization. *Nucleic Acids Research*, 47(W1): W345–W349, 2019. doi: 10.1093/nar/gkz424. URL https://doi.org/10.1093/nar/gkz424.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Knowles, J. R. Enzyme catalysis: not different, just better. *Nature*, 350(6314):121–124, 1991.

Kriválk, R. and Hoksza, D. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(1):39, 2018. doi: 10.1186/s13321-018-0285-8. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0285-8.

Li, Y. et al. Deepre: sequence-based enzyme ec number prediction by deep learning. *Bioinformatics*, 34(5):760–769, 2018.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/10.1101/2022.07.20.500902v3.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xQUe1pOKPam.

Liu, Y., Hua, C., Zeng, T., Rao, J., Zhang, Z., Wu, R., Coley, C. W., and Zheng, S. Enzymecage: A geometric foundation model for enzyme retrieval with evolutionary insights. *bioRxiv*, 2024. doi: 10.1101/2024.12.15.628585. URL https://www.biorxiv.org/content/10.1101/2024.12.15.628585v1.

Mikhael, P. G., Chinn, I., and Barzilay, R. Clipzyme: Reaction-conditioned virtual screening of enzymes. *arXiv preprint arXiv:2402.06748*, 2024.

Reisenbauer, J. C., Sicinski, K. M., and Arnold, F. H. Catalyzing the future: recent advances in chemical synthesis using enzymes. *Current Opinion in Chemical Biology*, 83:102536, December 2024. ISSN 13675931. doi: 10.1016/j.cbpa.2024.102536. URL https://linkinghub.elsevier.com/retrieve/pii/S1367593124001121.

Ribeiro, A. J., Riziotis, I. G., Borkakoti, N., and Thornton, J. M. Enzyme function and evolution through the lens of bioinformatics. *Biochemical Journal*, 480(22):1845–1863, 2023.

Sanderson, T., Bileschi, M. L., Belanger, D., and Colwell, L. J. Proteinfer, deep neural networks for protein functional inference. *eLife*, 12:e80942, 2023.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, July 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Schwaller, P., Probst, D., Vaucher, A. C., Nair, V. H., Kreutter, D., Laino, T., and Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152, January 2021. ISSN 2522-5839. doi: 10.1038/s42256-020-00284-w. URL https://www.nature.com/articles/s42256-020-00284-w.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017. doi: 10.1038/nbt.3988.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, pp. 1–12, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

Yabukarski, F., Hu, H., Lounnas, V., and Herschlag, D. Assessment of enzyme active site positioning and tests of catalytic mechanisms through x-ray–derived conformational ensembles. *Proceedings of the National Academy of Sciences*, 117(51):32378–32386, 2020. doi: 10.1073/pnas.2011350117.

Yang, J., Li, F.-Z., and Arnold, F. H. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Central Science*, pp. acscentsci.3c01275, February 2024a. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.3c01275. URL https://pubs.acs.org/doi/10.1021/acscentsci.3c01275.

Yang, J., Liu, S., Annadunkar, A., Munez, A., Wittmann, B. J., Arnold, F. H., and Yue, Y. Care: A benchmark suite for the classification and retrieval of enzymes. *arXiv preprint arXiv:2406.15669*, 2024b.

A. Dataset Details

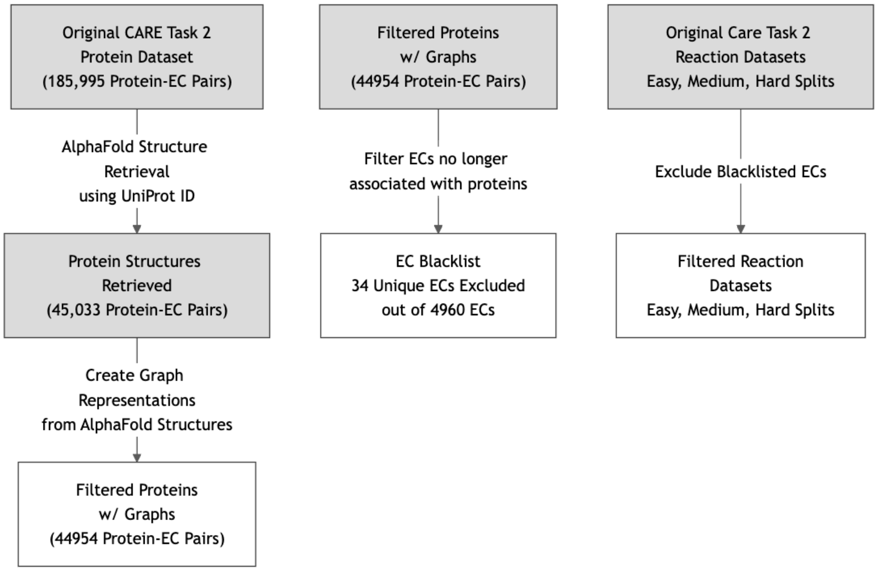

Figure 2. Workflow used to process the datasets containing EC numbers, protein sequences, and reactions used in this study. This process starts with the original CARE Task 2 datasets and applies filtering based on AlphaFold structure availability and successful graph creation to produce the modified datasets used for our experiments.

Our experiments utilized the pre-defined protein dataset, EC list, and Easy, Medium, and Hard out-of-distribution splits for Task 2 from the CARE benchmark (Yang et al., 2024b). These datasets were subsequently modified for our structure-based approach (Figure 2), which addresses the problem of reaction-enzyme mapping.

The Task 2 splits of the CARE Benchmark contain 185,995 protein-EC data pairs in total. For our structure-based approach, we sought to obtain corresponding AlphaFold structures from AlphaFold DB (Varadi et al., 2022) for the proteins in our Task 2 protein dataset based on their UniProt IDs. We also clustered the sequences at 50% identity with mmseqs2 (Steinegger & Söding, 2017) to reduce the number of sequences used during training and inference, while preserving diversity. Some structures could not be successfully retrieved for all proteins using the UniProt IDs from our protein dataset, potentially due to missing entries in AlphaFold database, or because the corresponding protein structure was listed under a different UniProt ID in AlphaFold database than the one provided in our dataset. Consequently, the initial set of protein-EC pairs for which AlphaFold structures were successfully retrieved contained 45,033 pairs.

From the proteins corresponding to these 45,033 successfully retrieved AlphaFold structures, we proceeded to create graph representations. Only the protein entries for which this graph creation process was successful were retained, forming the protein entries included in our final modified protein dataset. 79 protein entries were lost due to errors encountered during graph creation, such as structural inconsistencies in the AlphaFold data that prevented graph formation. This filtering, based on the successful creation of graph representations from the retrieved AlphaFold structures, directly determined the final size of the protein dataset used in our study, a total of 44,954 protein-EC data pairs.

This filtering of protein entries also impacted the associated EC numbers. We compiled an EC blacklist, which consists of 34 unique EC numbers that no longer had any associated protein entries after the protein filtering step. Consistent with this protein and EC filtering, reaction entries in the separate reaction datasets prepared for each of the easy, medium, and hard splits were also filtered to remove those associated with the blacklisted EC numbers. Detailed statistics regarding the number of unique ECs and sample counts for each split in the original and modified datasets are provided in Table 3.

Despite this reduction in the size of the datasets and the exclusion of a small number of ECs, the modified dataset retains the vast majority of the functional space covered by the original EC numbers. We note that due to the filtering criteria tied to structural data availability and successful graph creation, the distribution of protein instances per specific EC may differ when compared to the original dataset.

Table 3. Summary of train-test splits used, comparing the original Task 2 CARE benchmark counts to the modified version. ”Samples” refers to the number of reaction-EC pairs.

Split name	**Description**	**Unique ECs**		**Reaction Train Samples**		**Reaction Test Samples**	
		Original	Modified	Original	Modified	Original	Modified
Easy	Certain reactions are held out, sampled uniformly across ECs, but no EC numbers are held out. The test set is the same as the holdout set.	4,960	4,926	61,373	61,147	393	393
Medium	All reactions corresponding to certain ECs are held out, at EC level 4 (X.X.X.X). Test set reactions are sampled uniformly across ECs from the holdout set.	4,748	4,716	57,691	57,480	393	393
Hard	All reactions corresponding to certain ECs are held out, at EC level 3 (X.X.X.-). Test set reactions are sampled uniformly across ECs from the holdout set.	3,052	3,037	35,252	35,086	460	460

B. Downstream Retrieval

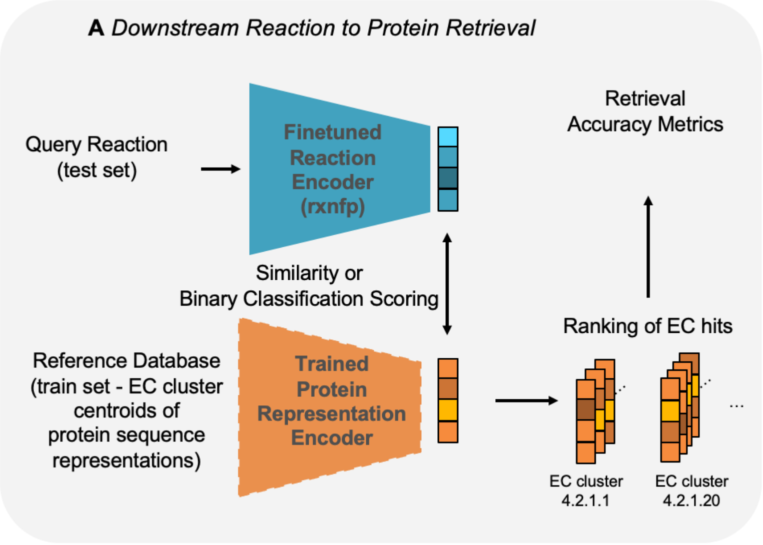
A Downstream Reaction to Protein Retrieval
Query Reaction (test set)
Finetuned Reaction Encoder (rxnfp)
Retrieval Accuracy Metrics
Similarity or Binary Classification Scoring
Ranking of EC hits
Reference Database (train set - EC cluster centroids of protein sequence representations)
Trained Protein Representation Encoder
EC cluster 4.2.1.1
EC cluster 4.2.1.20
...

Figure 3. Downstream retrieval for all pretrained models in this work. Modified with permission from Yang et al. (2024b).

Downstream retrieval for any pretrained model in this study is illustrated in Figure 3. Our approach is similar to the methodology used in CREEP, but we additionally incorporate binary classification scoring for the binary classification task. The process, briefly described here, retrieves Enzyme Commission (EC) numbers relevant to a query chemical reaction. The query reaction is encoded into an embedding using a finetuned reaction encoder, rxnfp. This embedding is then compared against a reference database of protein representations. To reduce inference time, first, this database is constructed by clustering protein sequences at 50% identity. Subsequently, the trained protein representation encoder (e.g., structure representation encoder) is used to calculate the centroid representation for each EC cluster. The comparison between the query embedding and the reference database yields a ranked list of ECs, which is then used to evaluate retrieval accuracy. Ranking depends on the learning objective. For contrastive learning, inference is done by comparing the reaction representation to the cluster centroids of protein sequence representations for each EC number, and subsequently ranking ECs based on their proximity to the reaction. For binary classification, inference is done by predicting the probability of a reaction belonging to each EC number, and subsequently ranking ECs based on their predicted probabilities.

C. Core Model Implementation Details

C.1. Structure Representation Encoder

Our Equivariant Graph Neural Network (EGNN) architecture for processing 3D protein graphs is identical to the structural encoder described in CLIPZyme (Mikhael et al., 2024), including aspects such as the number of layers, hidden dimensions, and activation functions. Contrary to CLIPZyme, which utilizes 1280-dimesional embeddings from the ESM-2 language model pretrained with 650M parameters, due to computational resource contraints, we used 640-dimensional embeddings derived from ESM-2 language model (Lin et al., 2022) pretrained with 150M parameters.

C.2. Sequence Representation Encoder

For our 'Sequence-only' baseline model, we use the ESM-2 language model (Lin et al., 2022) pretrained with 150M parameters with frozen weights, which produces 640-dimensional embeddings, whereas CREEP utilized embeddings from the ProtT5 language model. We then use an MLP to introduce non-linear transformations to these fixed embeddings. These input embeddings from the protein language model are first passed through a linear layer to 512-dimensional hidden space, followed by a `ReLU` activation. A final linear layer then maps this hidden representation to a 256-dimension feature vector.

C.3. Reaction Encoder

The encoding of chemical reactions from SMILES/SMARTS strings and the subsequent fine-tuning of the rxnfp model follows the exact methodology as CREEP (Yang et al., 2024b). This includes the end-to-end fine-tuning of rxnfp as part of the overall model training.

D. P2Rank

P2Rank (Kriválk & Hoksza, 2018; Jendele et al., 2019) operates on predicted protein structures to identify potential ligand-binding sites. It analyzes local surface properties and employs a trained Random Forest model to assign a preliminary ligandability score to each surface accessible (SAS) point x_i on the protein surface. A pre-trained Random Forest model RF maps each descriptor vector $\mathbf{f}_i$ characterizing the local surface environment of point x_i to a ligandability score:

$$s_i = \mathrm{RF}(\mathbf{f}_i),$$

where $s_i \in [0, 1]$ quantifies the likelihood that the SAS point x_i is part of a ligand-binding region.

Subsequently, SAS points with high ligandability scores are grouped into discrete binding pockets $\{P_j\}_{j=1}^{M}$ through single-linkage clustering, employing a spatial threshold of 3Å. Each identified pocket P_j is assigned a raw score S_j by aggregating the ligandability scores of its constituent SAS points:

$$S_j = \frac{1}{|P_j|} \sum_{x_i \in P_j} s_i.$$

These raw scores S_j are then calibrated to probability scores $P(S_j)$ representing the empirical likelihood that a pocket with score S_j is a true ligand-binding site. This calibration is performed using empirical priors derived from annotated datasets:

$$P(S_j) = \frac{T_{S_j}}{T_{S_j} + F_{S_j}},$$

where T_{S_j} and F_{S_j} denote the cumulative counts of true positive and false positive pockets, respectively, in a calibration dataset for pockets with raw scores less than or equal to S_j.

We utilized all pockets and their associated scores (S_j) and probabilities ($P(S_j)$) as output by P2Rank using its default parameters, without applying any further filtering criteria, before deriving our residue-level features as described in Section 2.

E. Integrating Pockets into our Core Model

Ligand binding and catalytic activity in enzymes are typically associated with specific three-dimensional regions known as binding pockets (active sites in enzymes). Incorporating information about these predicted pockets can provide valuable functional priors to the residue representations learned by our graph neural network. To achieve this, we first leverage a state-of-the-art pocket prediction method, P2Rank (Krivák & Hoksza, 2018; Jendele et al., 2019), to identify potential ligand-binding sites and assign them calibrated probability and raw scores. The detailed methodology of how P2Rank generates these pocket-level scores (S_j) and probabilities ($P(S_j)$) from protein structures is provided in Appendix D.

For each residue r_k, we define two scalar features summarizing the pocket information:

$$s_k = \max_{\substack{j \\ r_k \in P_j}} S_j$$

$$p_k = \max_{\substack{j \\ r_k \in P_j}} P(S_j)$$

Here, s_k is the maximum raw score among all predicted pockets P_j (outputted by P2Rank) that contain residue r_k, and p_k is the corresponding maximum probability-calibrated score. By taking the maximum, we assign the strongest pocket signal to a residue. This is especially important for residues that are located at the interface of multiple predicted pockets. For residues r_k that do not satisfy the distance criterion for inclusion in any predicted pocket P_j, we assign default values of $(s_k, p_k) = (0, 0)$, indicating the absence of a significant pocket signal at that location.

It is important to note that the probability $P(S_j)$ used to derive p_k is an aggregated value calibrated at the pocket level by P2Rank, representing the overall likelihood of the *entire pocket* P_j being a true binding site based on its total score S_j. Our residue feature p_k leverages this pre-computed, pocket-level confidence measure directly.

We explored three distinct strategies for integrating these residue-level pocket features (s_k, p_k) with the initial ESM-2 embeddings h_k:

E.1. Direct Concatenation ('Structure + Pockets'):

The two pocket-derived features, s_k and p_k, are directly concatenated to the original ESM-2 node feature vector h_k for each residue r_k, yielding an enhanced feature vector $h'_k \in \mathbb{R}^{642}$.

E.2. Adapted Pockets ('Structure + Adapted Pockets')

While the direct incorporation of raw pocket scores s_k and probabilities p_k as described in Appendix E, provides a valuable initial signal to the model, we observe two main limitations within this straightforward approach (described in Section 3.1). First, the residue-level pocket probability p_k, derived from P2Rank, is inherently sparse. Approximately 95% of residues in typical proteins do not fall within predicted pocket regions and are assigned a p_k value of 0. Second, representing potentially complex pocket-related information using only two scalar values (s_k, p_k) might be insufficient to fully complement the high-dimensional structural context captured the 640-dimensional node embeddings per residue.

We propose an adapted pocket feature representation. Our approach transforms the sparse, scalar pocket probability p_k into a dense, learnable vector representation.

First, we focus exclusively on residue-level pocket probabilities p_k, as they directly encode the chance a residue is part of a potential binding site. To provide a more standardized input for subsequent processing, we apply a standard normalization across distribution of $\{p_k\}_{k=1}^{L}$ values. We denote these normalized probabilities for residue r_k as p_k^{norm}. We then project p_k^{norm} into a higher-dimensional feature space using an MLP, which produces a feature vector we denote as $\tilde{f}_k \in \mathbb{R}^{16}$, for each residue r_k.

Furthermore, to allow the model to adaptively control the influence of this engineered pocket feature relative to the pre-computed structural features h_k, we introduce a learnable scalar scaling factor $\gamma \in \mathbb{R}$. This factor modulates the magnitude of the projected pocket feature vector:

$$f_k = \gamma \cdot \tilde{f}_k,$$

where $f_k \in \mathbb{R}^{16}$ is the final adapted pocket feature vector for residue r_k. The scaling factor γ is initialized to 1.0 and is trained jointly with the rest of the network parameters. The resulting n-dimensional adapted pocket feature vector f_k is then concatenated with the original ESM-2 feature vector h_k to form the enhanced node feature vector $\tilde{h}_k$ that is used as input to the subsequent EGNN layers:

$$\tilde{h}_k = \text{concat}(h_k, f_k).$$

where $\tilde{h}_k \in \mathbb{R}^{656}$.

E.3. Pocket-Weighted Message Passing (‘Structure + Weighted Pockets’)

Building upon the residue representations $\tilde{h}_k$ which are enriched with local structural and pocket-derived features, we introduce a mechanism to guide the message aggregation process within the EGNN, prioritizing information flow from residues likely involved in catalysis. Standard Graph Neural Networks often aggregate information from neighbors either uniformly or through learned attention mechanisms, where the weighting is determined solely by the network’s parameters based on features (Kipf & Welling, 2017; Veličković et al., 2018; Wu et al., 2020). However, the catalytic activity of an enzyme is predominantly localized within specific binding pocket regions (Yabukarski et al., 2020).

To leverage this prior explicitly, we propose a *pocket-weighted message passing* scheme. Instead of learning context-dependent attention weights or applying uniform weights, we directly utilize the previously computed residue-level pocket probability p_k (derived from P2Rank’s pocket prediction confidence) to modulate the contribution of neighbor messages.

Let $h_i^{(l)}$ denote the node embedding of residue i at layer l. The set of neighbors of residue i in the graph is denoted by $\mathcal{N}(i)$. Like CLIPZyme (Mikhael et al., 2024), the message $m_{ij}^{(l)}$ generated from neighbor j to node i is produced by an Edge MLP, ϕ_e, as follows.

$$m_{ij}^{(l)} = \phi_e(h_i^{(l)}, h_j^{(l)}, e_{ij}^{(l)})$$

where $e_{ij}^{(l)}$ refers to the relative distance embeddings between residues i and j at layer l, which are encoded sinusoidally (Vaswani et al., 2017; Atz et al., 2022). For our pocket-weighted message passing, we modify the message passing method of CLIPzyme to obtain $h_i^{(l+1)}$ at layer $l+1$ as follows:

$$h_i^{(l+1)} = h_i^{(l)} + \phi_h \left(\text{concat} \left(h_i^{(l)}, \sum_{j \in \mathcal{N}(i)} \left(m_{ij}^{(l)} \odot p_i^{(l)} \right) \right) \right)$$

where p_i is the pocket probability feature (maximum pocket probability) of node i.

F. Contrastive Objective

For training the sequence and structure models, we follow a similar methodology to CREEP, which is briefly described in the following paragraphs.

In the contrastive pre-training phase, protein representations x_p (either sequence or structure-based) and reaction representations x_r (derived from rxnfp) from the same enzyme family x form positive pairs (x_p, x_r). We create negative samples in the following manner. We randomly sample from a Gaussian distribution, which is taken as an approximation of empirical data distribution, to obtain representations x'_p and x'_r. For each positive sample, we sample sixteen, negative samples in our implementation. The protein and reaction representations are projected to 256-dimensional vectors in a shared latent space.

Our approach takes inspiration from GraphMVP and employs the Energy-Based Model Noise-Contrastive Estimation (EBM-NCE) objective function to estimate mutual information between our modalities (Liu et al., 2022). This objective aligns protein-reaction pairs from the same enzyme family while contrasting them against pairs from different families. The EBM-NCE loss is written as

$$\begin{aligned} L_{EBM\text{-}NCE} = -\frac{1}{2} [&\mathbb{E}_{x_p,x_r} \log \sigma(E(x_p, x_r)) + \mathbb{E}_{x_p,x'_r} \log(1 - \sigma(E(x_p, x'_r))) \\ &+ \mathbb{E}_{x_p,x_r} \log \sigma(E(x_p, x_r)) + \mathbb{E}_{x'_p,x_r} \log(1 - \sigma(E(x'_p, x_r)))] \end{aligned} \tag{1}$$

where $E(\cdot)$ and $\sigma(\cdot)$ are the energy and sigmoid functions, respectively. The dot product is used to compute the similarity of the aforementioned representations within the latent space.

G. BCE Objective with Sequence-Based CREEP

For binary classification, the protein (x_p) and reaction (x_r) representations associated with an enzyme family x are combined via concatenation, yielding a feature vector that is then input to a shallow neural network classifier. The classifier yields a predictive probability $\hat{y} \in [0, 1]$, signifying the likelihood that the specific protein and reaction corresponding to x_p and x_r engage in the interaction of interest. The classifier is trained by minimizing the Binary Cross-Entropy (BCE) logits loss function, which is formulated as:

$$L_{BCEWithLogits}(z(x_p, x_r), y) = -[y \log(\sigma(z(x_p, x_r))) + (1 - y) \log(1 - \sigma(z(x_p, x_r)))]$$

where $z(x_p, x_r)$ explicitly denotes the model's predicted logits conditioned on the input protein and reaction representations.

We employed a Multi-Layer Perceptron (MLP). This network processes concatenated 256-dimensional protein from the sequence representation encoder and 256-dimensional reaction embeddings, resulting in a 512-dimensional input. This input is passed through a linear layer to a 128-dimensional hidden space, followed by `BatchNorm1d` for stabilization and a `ReLU` activation. A final linear layer then maps this hidden representation to a single output logit, which is used with a `BCEWithLogitsLoss` function.