Acta Pharmaceutica Sinica B 2026;16(7):4743−4745

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

COMMENTARY

Reaction-aware molecular representation learning: Toward generalizable artificial intelligence for enzymatic catalysis

KEY WORDS

Enzymatic reaction;
Molecular representation learning;
Protein language model;
Knowledge graph;
Multimodal learning;
Substrate prediction;
Binding site prediction;
Biocatalysis

Enzymes are central to life and increasingly indispensable to modern pharmaceutical research, synthetic biology, and sustainable chemical manufacturing. Their remarkable catalytic efficiency and exquisite chemo-, regio-, and stereoselectivity make them attractive alternatives to conventional chemical catalysts. However, the functional landscape of enzymes remains only sparsely annotated. Although public databases contain tens of millions of enzyme sequences, only a small fraction have been manually characterized with reliable information on catalytic function, substrate specificity, reaction scope, or binding residues[1,2]. This annotation gap has become a major bottleneck for enzyme engineering, metabolic pathway design, and biocatalyst discovery.

Machine learning has emerged as a powerful strategy to accelerate enzyme function annotation and enzymatic reaction modeling. Early approaches relied heavily on handcrafted physicochemical descriptors, structural features, or experimentally curated data, which limited their scalability and generalizability. More recently, deep learning and large pre-trained molecular models have enabled protein and small-molecule representations to be learned directly from large-scale sequence or chemical data[3,4]. These advances have substantially improved tasks such as enzyme classification, substrate prediction, and functional site annotation. Nevertheless, many existing models remain task-specific or unimodal, typically representing enzymes only through protein sequences or structures. Such representations overlook a fundamental biochemical principle: enzymes do not act in isolation, but exert their function through specific catalytic transformations linking substrates to products.

In this issue, Huang et al.[5] report an advance in enzymatic reaction modeling through the development of ERAM, an enzymatic-reaction-aware molecular representation learning framework. The study addresses a central challenge in artificial intelligence for biocatalysis: how to construct molecular representations that are not merely protein- or molecule-aware, but explicitly aware of enzymatic reaction context. By integrating pre-trained protein and small-molecule models with multimodal relational learning, ERAM provides a task-agnostic framework that can be applied to enzyme retrieval, substrate prediction, and binding site prediction without being designed exclusively for any single downstream task.

The conceptual foundation of ERAM is particularly innovative. Instead of treating enzymes, substrates, and products as independent entities, the authors formulate enzymatic reactions as triplets in a knowledge graph, where substrates and products serve as molecular entities and enzymes function as relations connecting them. This design is reminiscent of the classical TransE framework, in which relations are modeled as translations between head and tail entities in a shared embedding space[6]. In the biochemical context, this formulation carries intuitive mechanistic meaning: the enzyme representation acts as a catalytic transformation vector that maps a substrate representation toward a product representation. Such a design elegantly bridges chemical reaction logic and representation learning, providing a geometrically interpretable way to encode enzymatic catalysis.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2026.06.033

Further, ERAM strengthens this reaction-aware representation by incorporating multimodal molecular encoders. Small molecules are represented using Uni-Mol, while enzyme sequences are encoded using ESM-2-derived protein language model representations[3,4]. Importantly, the enzyme encoder includes a cross-attention mechanism that allows substrate information to modulate enzyme representations. This design reflects the biological reality that enzyme function can be substrate-dependent. Enzymes with broad specificity may adopt different functional representations depending on the substrate context, a phenomenon related to enzyme promiscuity and induced fit. By allowing the enzyme embedding to become substrate-aware, ERAM moves beyond static protein representation and toward context-dependent catalytic representation.

Another important methodological contribution is the dual-grained contrastive learning strategy. In conventional contrastive learning, negative samples are often treated uniformly. However, enzymatic reactions contain different levels of biochemical mismatch. Replacing the product disrupts the chemical transformation itself and represents a coarse-grained negative sample. In contrast, replacing or removing the enzyme may preserve the same substrate−product chemical equilibrium while altering the catalytic feasibility or rate, corresponding to a fine-grained negative sample. ERAM explicitly distinguishes these two types of negatives by assigning different margins in the contrastive objective. This design injects biochemical prior knowledge into the learning process and enables the model to better capture the functional disparity between small molecules and enzymes.

Moreover, the authors introduce prototype learning to improve contrastive learning under limited batch sizes. By maintaining enzyme-class prototypes and encouraging embeddings from the same class to move closer to their corresponding prototypes, the model indirectly increases the number of positive samples without requiring larger batches or additional model parameters. This is a practical and effective solution, especially because enzyme reaction datasets are often imbalanced and computational constraints limit the scale of contrastive learning.

The performance of ERAM across multiple tasks demonstrates the strength of this design. In enzymatic reaction retrieval, ERAM achieves outstanding product retrieval and enzyme retrieval performance across test sets with different levels of sequence identity. Notably, even in low-sequence-identity subsets, the model maintains robust retrieval performance, suggesting that ERAM does not simply memorize sequence similarity but learns more generalizable reaction-aware representations. Compared with Reactzyme[7] and CREEP[8], ERAM consistently achieves superior enzyme retrieval performance, especially in challenging subsets without EC annotations. This result is particularly important because practical enzyme discovery often involves poorly annotated or evolutionarily distant proteins.

The substrate prediction results further highlight the utility of ERAM as a general-purpose enzymatic reaction representation model. On nitrilase and aminotransferase datasets, ERAM outperforms the state-of-the-art ESP model across most metrics and splitting strategies[9]. The performance gain under substrate-based splitting is especially meaningful, as this setting evaluates whether the model can generalize to unseen chemical substrates rather than relying on familiar molecular scaffolds. ERAM's strong performance in this scenario suggests that reaction-aware relational learning can improve chemical generalization, a key requirement for discovering new enzyme−substrate pairs in biocatalysis.

Beyond predictive accuracy, ERAM offers a degree of interpretability that is valuable for enzyme science. The authors show that attention scores from the enzyme encoder preferentially highlight experimentally annotated binding regions. In the unsupervised binding site prediction task, ERAM achieves a higher overlap score and lower false positive rate than RXNAAMapper and other baseline models[10]. Case studies further demonstrate that high-attention residues correspond to known binding motifs, conserved domains, or docking-supported binding pockets. This is an important finding because many enzyme sequences lack experimentally verified structural or functional annotations. A model capable of suggesting candidate binding regions from sequence and reaction context could provide useful hypotheses for mutagenesis, enzyme engineering, and mechanistic studies.

The attention visualization of the small-molecule encoder also provides mechanistic insight. ERAM assigns higher attention to chemically meaningful reaction sites, including bond cleavage sites, hydroxyl groups involved in elimination reactions, stereochemical transformation sites, and carbon−carbon double bonds involved in addition reactions. These observations suggest that the model does not merely learn statistical associations between proteins and molecules, but captures chemically relevant features underlying enzymatic transformations. Such interpretability is essential if artificial intelligence models are to become reliable tools for mechanistic enzymology rather than black-box predictors.

Several aspects of this study merit broader attention. First, ERAM offers a compelling example of how domain knowledge can be embedded into representation learning. The use of reaction-balanced substrate−product pairs, translation-based relational geometry, dual-grained negatives, and substrate-aware enzyme representations collectively reflects a careful integration of biochemical principles with modern deep learning. Second, the framework is task-agnostic. Rather than training separate models for enzyme retrieval, substrate prediction, and binding site prediction, ERAM learns a unified latent space that supports multiple downstream applications. Third, the work illustrates a promising direction for connecting large pretrained molecular models with structured biochemical knowledge. Protein language models and molecular foundation models provide powerful initial representations, but ERAM demonstrates that these representations can be substantially enriched when aligned with enzymatic reaction context.

The study also raises important questions for future research. Although ERAM shows strong performance across computational benchmarks, prospective experimental validation will be critical for establishing its practical value in enzyme discovery and engineering. For example, model-predicted enzyme−substrate pairs or binding residues could be tested through biochemical assays, site-directed mutagenesis, or directed evolution campaigns. In addition, ERAM currently relies primarily on protein sequence-derived representations and small-molecule embeddings. Incorporating high-quality protein structures, predicted conformational ensembles, enzyme pockets, and reaction transition-state information may further improve its ability to model catalytic mechanisms. Another important direction is uncertainty estimation, particularly for low-confidence predictions involving rare enzyme families, ambiguous RHEA reactions, or substrates outside the training distribution.

From a broader perspective, ERAM represents a step toward mechanistically informed artificial intelligence for biocatalysis. The field is moving beyond isolated prediction tasks toward integrated models that understand proteins, small molecules, reactions, and biological function within a unified framework. Such models could accelerate enzyme annotation, identify candidate catalysts for synthetic transformations, guide enzyme engineering, and support the construction of metabolic pathways. In

pharmaceutical sciences, where biocatalysis is increasingly used for greener and more selective synthesis, reaction-aware enzyme modeling may become an important component of computer-aided process development and drug discovery.

In conclusion, the study by Huang et al.[5] makes a contribution to enzymatic reaction modeling by introducing a task-agnostic, multimodal, and reaction-aware representation learning framework. By formulating enzymatic catalysis as relational learning among substrates, enzymes, and products, ERAM provides both strong predictive performance and meaningful interpretability. This work highlights the value of embedding biochemical principles directly into artificial intelligence models and points toward a future in which enzyme function, substrate scope, and catalytic mechanisms can be explored through unified molecular representation spaces. As computational predictions become increasingly coupled with experimental validation, frameworks such as ERAM may help unlock the vast uncharacterized enzyme repertoire for drug discovery, synthetic biology, and sustainable chemistry.

References

1. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. Brenda, the elixir core data resource in 2021: new developments and updates. *Nucleic Acids Res* 2021;**49**:D498−508.
2. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Res* 2025;**53**:D609−17.
3. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123−30.
4. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, et al. Uni-mol: a universal 3D molecular representation learning framework. *ChemRxiv* 2023. Available from: https://doi.org/10.26434/chemrxiv-2022-jjm0j-v4.
5. Huang Y, Li L, Qian W, Yu J, Zhao H, Wang X, et al. Accurate and task-agnostic modeling of enzymatic reactions through multimodal relational learning. *Acta Pharm Sin B* 2026;**16**:4051−66.
6. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems 26*; 2013. Available from: https://papers.nips.cc/paper_files/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.
7. Hua C, Zhong B, Luan S, Hong L, Wolf G, Precup D, et al. Reactzyme: a benchmark for enzyme-reaction prediction. *Advances in Neural Information Processing Systems 37*; 2024. Available from: https://doi.org/10.52202/079017-0832.
8. Yang J, Mora A, Liu S, Wittmann BJ, Anandkumar A, Arnold FH, et al. Care: a benchmark suite for the classification and retrieval of enzymes. *Advances in Neural Information Processing Systems 37*; 2024. Available from: https://doi.org/10.52202/079017-0101.
9. Kroll A, Ranjan S, Engqvist MK, Lercher MJ. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nat Commun* 2023;**14**:2787.
10. Teukam YGN, Dassi LK, Manica M, Probst D, Schwaller P, Laino T. Language models can identify enzymatic binding sites in protein sequences. *Comput Struct Biotechnol J* 2024;**23**:1929−37.

Kelin Xia
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore
E-mail address: xiakelin@ntu.edu.sg.