Novelty Assessment Report

Paper: PoinnCARE: Hyperbolic Multi-Modal Learning for Enzyme Classification

PDF URL: https://openreview.net/pdf?id=dGxAYNK6JU

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Enzyme Commission (EC) number prediction is vital for elucidating enzyme functions and advancing biotechnology applications. However, current methods struggle to capture the hierarchical relationships among enzymes and often overlook critical structural and active site features. To bridge this gap, we introduce PoinnCARE, a novel framework that jointly encodes and aligns multi-modal data from enzyme sequences, structures, and active sites in hyperbolic space. By integrating graph diffusion and alignment techniques, PoinnCARE mitigates data sparsity and enriches functional representations, while hyperbolic embedding preserves the intrinsic hierarchy of the EC system with theoretical guarantees in low-dimensional spaces. Extensive experiments on four datasets from the CARE benchmark demonstrate that PoinnCARE consistently and significantly outperforms state-of-the-art methods in EC number prediction.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Enzyme Commission number prediction**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Representation Learning and Embedding Methods**
- **Prediction Models and Architectures**
- **Reaction-Based and Chemical Transformation Methods**
- **Specialized Prediction Tasks and Applications**
- **Benchmarking, Evaluation, and Comparative Studies**
- **Computational Assignment and Database Methods**
- **Theoretical Foundations and Limitations**
- **Application-Driven and Discovery-Oriented Methods**
- **Binary Classification and Auxiliary Prediction Tasks**
- **Software Tools and Repositories**

Complete Taxonomy Tree

- Enzyme Commission number prediction Survey Taxonomy
- Representation Learning and Embedding Methods
  - EC Number Encoding and Embedding (1 papers)
  - [1] EC2Vec: A Machine Learning Method to Embed Enzyme Commission (EC) Numbers into Vector Representations (Mengmeng Liu, 2025) View paper
  - Protein Representation Learning (2 papers)
  - [11] Comparative Assessment of Protein Large Language Models for Enzyme Commission Number Prediction (JoÃ£o Capela, 2025) View paper
  - [44] EZpred: improving deep learning-based enzyme function prediction using unlabeled sequence homologs (Chengxin Zhang, 2025) View paper
- Prediction Models and Architectures
  - Sequence-Based Prediction Models
  - Transformer and Attention-Based Models (3 papers)
    - [9] Enhancing Enzyme Commission Number Prediction With Contrastive Learning and Agent Attention (Wen-di Zhao, 2025) View paper
    - [22] HIT-EC: Trustworthy prediction of enzyme commission numbers using a hierarchical interpretable transformer (Louis Dumontet, 2025) View paper
    - [47] naomifridman/transformer-based-enzyme-classification: Initial Release: ESM2 Enzyme Classification (Fridman, 2025) View paper
  - Recurrent and Convolutional Networks (2 papers)
    - [15] EnzymeNet: residual neural networks model for Enzyme Commission number prediction (Naoki Watanabe, 2023) View paper
    - [26] Predicting Enzyme Commission Numbers Using Recurrent Neural Networks with Amino Acid Sequence Shift and Consistency Loss (Danny Paik, 2025) View paper
  - Ensemble and Interpretable Methods (3 papers)
    - [16] ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature (Alperen DalkÄ±ran, 2018) View paper

        - [31] Prediction of enzyme function using an interpretable optimized ensemble learning framework (Saikat Dhibar, 2025) View paper
        - [42] Prediction of Enzyme function using interpretable optimized Ensemble learning framework (Saikat Dhibar, 2025) View paper
    - Novel Architectures and Specialized Networks (2 papers)
        - [6] Interpretable Kolmogorov-Arnold Networks for Enzyme Commission Number Prediction (Louis Dumontet, 2025) View paper
        - [13] Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers (Jae Yong Ryu, 2019) View paper
    - Structure-Based Prediction Models (5 papers)
    - [7] Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures (Yidong Song, 2024) View paper
    - [19] EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation (Afshine Amidi, 2018) View paper
    - [27] TopEC: prediction of Enzyme Commission classes by 3D graph neural networks and localized 3D protein descriptor (Karel van der Weg, 2025) View paper
    - [28] Improved Functional Classification of Hydrolases through Pairwise Structural Similarity of Reaction Cores (Sakib Ferdous, 2025) View paper
    - [50] Predicting enzyme class from protein structure without alignments (P. Dobson, 2005) View paper
    - Multimodal and Hybrid Prediction Models ★ (4 papers)
    - [0] PoinnCARE: Hyperbolic Multi-Modal Learning for Enzyme Classification (Anon et al., 2026) View paper
    - [23] Multimodal Quantum Vision Transformer for Enzyme Commission Classification from Biochemical Representations (Isik Murat, 2025) View paper
    - [38] Autoregressive enzyme function prediction with multi-scale multi-modality fusion (Dingyi Rong, 2025) View paper
    - [39] SST-ResNet: A Sequence and Structure Information Integration Model for Protein Property Prediction (Guo-Wei Zhou, 2025) View paper
    - Hierarchical and Multitask Learning Frameworks (3 papers)
    - [33] Enzyme Commission Number Prediction and Benchmarking with Hierarchical Dual-core Multitask Learning Framework (Zhenkun Shi, 2023) View paper
    - [35] Hierarchical Clustering-Based Coarse-to-Fine Classification Framework for Microbial Protein Function Prediction (Shengyang Chen, 2025) View paper
    - [36] CACLENS: A Multitask Deep Learning System for Enzyme Discovery. (Xilong Yi, 2025) View paper
    - Contrastive and Self-Supervised Learning (1 papers)
    - [5] CLAIRE: a contrastive learning-based predictor for EC number of chemical reactions (Zishuo Zeng, 2025) View paper
- Reaction-Based and Chemical Transformation Methods (5 papers)
    - [8] A general model for predicting enzyme functions based on enzymatic reactions (Wenjia Qian, 2024) View paper
    - [20] ECOH: an enzyme commission number predictor using mutual information and a support vector machine (Yoshihiko Matsuta, 2013) View paper
    - [24] Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions (Masaaki Kotera, 2004) View paper
    - [34] E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs (Yoshihiro Yamanishi, 2009) View paper
    - [37] Leveraging large language models for enzymatic reaction prediction and characterization (Lorenzo Di Fruscia, 2025) View paper
- Specialized Prediction Tasks and Applications
    - Enzyme Family and Class Classification (2 papers)
    - [4] The classification of enzymes by deep learning (Zhiyu Tao, 2020) View paper
    - [10] PredictEFC: a fast and efficient multi-label classifier for predicting enzyme family classes (Lei Chen, 2024) View paper
    - Domain-Level and Active Site Prediction (4 papers)
    - [21] Ec-psi: Associating enzyme commission numbers with pfam domains (Seyed Ziaeddin Alborzi, 2015) View paper
    - [25] Interpretable Enzyme Function Prediction via Residue-Level Detection (Yang Zhao, 2025) View paper
    - [30] ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains (Seyed Ziaeddin Alborzi, 2017) View paper
    - [40] CAPIM: Catalytic activity and site prediction and analysis tool in multimer proteins (GÃ¶khan Ã▯zsari, 2025) View paper
    - Specialized Enzyme Classes and Databases (2 papers)
    - [18] A classification of glycosyl hydrolases based on amino acid sequence similarities (B. Henrissat, 1991) View paper
    - [29] GlycoEnzDB: A database of enzymes involved in human glycosylation (Yusen Zhou, 2025) View paper
- Benchmarking, Evaluation, and Comparative Studies (2 papers)
    - [2] EC-Bench: A Benchmark for Enzyme Commission Number Prediction (Saeedeh Davoudi, 2025) View paper
    - [48] Limitations of current machine learning models in predicting enzymatic functions for uncharacterized proteins (Valerie de Crecy-Lagard, 2025) View paper
- Computational Assignment and Database Methods (1 papers)
    - [17] HiFi-NN annotates the microbial dark matter with enzyme commission numbers (G Ayres, 2023) View paper
- Theoretical Foundations and Limitations (4 papers)
    - [3] Classification of enzymes (Grunwald, 2026) View paper
    - [14] Practical limits of function prediction (Damien Devos, 2000) View paper
    - [43] A brief guide to enzyme nomenclature and classification (K Tipton, 2018) View paper
    - [49] A comparison of prediction methods for the creation of field-extent soil property maps (Thomas F. A. Bishop, 2001) View paper
- Application-Driven and Discovery-Oriented Methods (3 papers)
    - [41] A Machine Learning Approach for Plant-based Drug Discovery: High-Throughput Prediction of Biological Activities and Enzyme Commission Numbers from Phytochemicals and Amino Acid Sequences of Plants (Leoni Kim, 2025) View paper
    - [45] Machine Learning-Driven Enzyme Mining: Opportunities, Challenges, and Future Perspectives (Zhang Yanzi, 2025) View paper
    - [46] Decoding the dark proteome: Deep learning-enabled discovery of druggable enzymes in Wuchereria bancrofti (Hernandez, 2025) View paper

- Binary Classification and Auxiliary Prediction Tasks (1 papers)
  - [12] A Binary Classifier for the Prediction of EC Numbers of Enzymes (Hao Cui, 2019) View paper
- Software Tools and Repositories (1 papers)
  - [32] NickFlagleaf/GxE-FA-EC-Prediction: v1.0 (Fradgley, 2025) View paper

Narrative

Core task: Enzyme Commission number prediction aims to assign standardized functional labels to enzymes based on their catalytic activities, typically using sequence or structural information. The field has evolved into a rich landscape organized around several complementary directions. Representation Learning and Embedding Methods explore how to encode protein sequences and structures into informative feature spaces, often leveraging protein language models or geometric embeddings. Prediction Models and Architectures encompass a diverse array of neural network designs—ranging from convolutional and recurrent architectures to transformers and hybrid multimodal frameworks—that map these representations to EC labels. Reaction-Based and Chemical Transformation Methods focus on substrate and product information to infer enzymatic function, while Specialized Prediction Tasks address hierarchical classification, domain-level annotation, and enzyme discovery. Benchmarking, Evaluation, and Comparative Studies provide systematic assessments of model performance, and Computational Assignment and Database Methods support large-scale annotation pipelines. Theoretical Foundations examine the intrinsic limits of function prediction, and Application-Driven approaches target real-world scenarios such as metagenomics or drug discovery.

Recent work has intensified around multimodal and hybrid prediction strategies that combine sequence embeddings with structural or chemical context. For instance, PoinnCARE[0] integrates multiple data modalities to improve prediction robustness, positioning itself within the Multimodal and Hybrid Prediction Models branch alongside efforts like Autoregressive Enzyme Prediction[38] and SST-ResNet[39], which explore sequential decoding and residue-level feature extraction respectively. These approaches contrast with purely sequence-based models such as EC2Vec[1] or transformer-only designs like Transformer Enzyme Classification[47], highlighting a trade-off between model complexity and interpretability. Meanwhile, benchmarking initiatives like EC-Bench[2] and assessments of protein language models (Protein Language Models Assessment[11]) underscore ongoing questions about generalization across enzyme families and the practical limits of data-driven methods. PoinnCARE[0] thus sits at the intersection of representation fusion and architectural innovation, aiming to leverage complementary signals where single-modality methods may plateau.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Multimodal Quantum Vision Transformer for Enzyme Commission Classification from Biochemical Representations

Authors: Isik Murat, Saggi, Mandeep Kaur, Murat Isik, M. Saggi, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Accurately predicting enzyme functionality remains one of the major challenges in computational biology, particularly for enzymes with limited structural annotations or sequence homology. We present a novel multimodal Quantum Machine Learning (QML) framework that enhances Enzyme Commission (EC) classification by integrating four complementary biochemical modalities: protein sequence embeddings, quantum-derived electronic descriptors, molecular graph structures, and 2D molecular image representat...

Relationship Analysis

Both papers belong to the Multimodal and Hybrid Prediction Models category, integrating multiple data modalities (sequence, structure, and chemical information) for EC number prediction. They overlap in their use of protein sequences, structural features, and multi-modal fusion strategies to enhance enzyme function classification. However, the original paper (PoinnCARE) employs hyperbolic space embeddings with graph diffusion to preserve hierarchical EC relationships and addresses active site annotation sparsity, while the candidate paper uses a Quantum Vision Transformer framework that incorporates quantum-derived electronic descriptors and operates in quantum computing environments rather than hyperbolic geometry.

2. Autoregressive enzyme function prediction with multi-scale multi-modality fusion

Authors: Dingyi Rong, Bozitao Zhong, Wenzhuo Zheng, Liang, Hong, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Abstract Accurate prediction of enzyme function is crucial for elucidating biological mechanisms and driving innovation across various sectors. Existing deep learning methods tend to rely solely on either sequence data or structural data and predict the Enzyme Commission (EC) number as a whole, neglecting the intrinsic hierarchical structure of EC numbers. To address these limitations, we introduce Multi-scale multi-modality Autoregressive Predictor (MAPred), a novel multi-modality and multi-sca...

Relationship Analysis

Both papers belong to the Multimodal and Hybrid Prediction Models category, integrating multiple data modalities (sequence, structure) for EC number prediction. They overlap in their multi-modal approach and recognition of EC hierarchy, but differ fundamentally in their technical approaches: PoinnCARE employs hyperbolic space embeddings with graph diffusion to preserve hierarchical relationships and includes active site information, while MAPred uses an autoregressive prediction network with dual-pathway architecture to sequentially predict EC digits without hyperbolic geometry or active site modeling.

3. SST-ResNet: A Sequence and Structure Information Integration Model for Protein Property Prediction

Authors: Guo-Wei Zhou, Yanpeng Zhao, Guowei Zhou, Song He, Xiaochen Bo | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Proteins are the basic building blocks of life and perform fundamental functions in biology. Predicting protein properties based on amino acid sequences and 3D structures has become a key approach to accelerating drug development. In this study, we propose a novel sequence- and structure-based framework, SST-ResNet, which consists of the multimodal language model ProSST and a multi-scale information integration module. This framework is designed to deeply explore the latent relationships between...

Relationship Analysis

Both papers belong to the Multimodal and Hybrid Prediction Models category, integrating sequence and structure information for EC number prediction. They overlap in their use of multi-modal data (sequence and structure) and their application to enzyme function prediction tasks. However, PoinnCARE uniquely incorporates active site information and employs hyperbolic space embeddings with graph diffusion to preserve hierarchical EC relationships, while SST-ResNet focuses on a multimodal language model (ProSST) with multi-scale information integration in Euclidean space without explicit hierarchical modeling or active site features.

Contributions Analysis

Overall novelty summary. The paper introduces PoinnCARE, a framework that jointly encodes enzyme sequences, structures, and active sites in hyperbolic space for EC number prediction. It resides in the 'Multimodal and Hybrid Prediction Models' leaf, which contains four papers total (including PoinnCARE). This leaf sits within the broader 'Prediction Models and Architectures' branch, indicating a moderately populated research direction focused on integrating multiple data modalities. The taxonomy shows that multimodal approaches represent one of several parallel strategies, alongside sequence-only, structure-only, and hierarchical learning frameworks.

The taxonomy reveals neighboring leaves such as 'Hierarchical and Multitask Learning Frameworks' (three papers) and 'Structure-Based Prediction Models' (five papers), suggesting that PoinnCARE bridges structural modeling with multimodal integration. The 'Representation Learning and Embedding Methods' branch (two leaves, three papers) addresses complementary questions about encoding EC numbers and proteins, while 'Reaction-Based and Chemical Transformation Methods' (five papers) explores an orthogonal direction using substrate-product information. PoinnCARE's hyperbolic embedding approach diverges from standard Euclidean representations common in sibling papers, positioning it at the intersection of geometric representation learning and multimodal fusion.

Among 30 candidates examined, the analysis identifies limited prior work overlap. The core hyperbolic framework contribution (Contribution 1) examined 10 candidates with zero refutations, suggesting relative novelty in applying hyperbolic geometry to enzyme prediction. However, multi-modal dataset augmentation (Contribution 2) and graph diffusion for sparsity (Contribution 3) each found one refutable candidate among 10 examined, indicating that structural and active site integration, as well as graph-based augmentation techniques, have precedents in the limited search scope. The statistics reflect a focused semantic search rather than exhaustive coverage.

Based on the limited search scope of 30 semantically similar papers, PoinnCARE appears to occupy a relatively novel position by combining hyperbolic embeddings with multi-modal enzyme data. The taxonomy context shows a moderately crowded multimodal prediction space but sparse exploration of non-Euclidean geometries. The analysis does not capture potential overlaps outside the top-30 semantic matches or in adjacent fields like graph representation learning, leaving open questions about broader precedents for hyperbolic enzyme embeddings.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: PoinnCARE framework for hyperbolic multi-modal enzyme learning

Description: The authors propose PoinnCARE, a framework that integrates sequence, structure, and active site information of enzymes and represents them in hyperbolic space. This approach preserves the hierarchical EC taxonomy structure while capturing comprehensive enzyme characteristics through multi-modal learning and alignment.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Bidirectional Hierarchical Protein Multi-Modal Representation Learning

URL: View paper

Brief Assessment

Bidirectional Hierarchical Protein[67] focuses on bidirectional hierarchical fusion of sequence and structure modalities using attention and gating mechanisms in Euclidean space, not hyperbolic space-based multi-modal alignment with active site information as in PoinnCARE.

2. OneProt: Towards multi-modal protein foundation models via latent space alignment of sequence, structure, binding sites and text encoders

URL: View paper

Brief Assessment

OneProt[61] focuses on multi-modal protein foundation models using ImageBind-style alignment in Euclidean space, while the original paper proposes hyperbolic space embeddings specifically for enzyme EC number prediction with hierarchical structure preservation.

3. Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites

URL: View paper

Brief Assessment

Multi-modal Active Sites[64] focuses on enzyme active site annotation using PLM-structure fusion and reaction information, not on hyperbolic space representations or EC taxonomy preservation.

4. A multimodal Transformer Network for protein-small molecule interactions enhances predictions of kinase inhibition and enzyme-substrate relationships

URL: View paper

Brief Assessment

Multimodal Transformer Network[68] focuses on protein-small molecule interactions using a multimodal transformer that processes amino acid sequences and SMILES strings together, but does not address enzyme classification, EC number prediction, or hyperbolic space representations. The candidate's approach is fundamentally different from PoinnCARE's hyperbolic multi-modal learning framework for enzyme function prediction.

5. OneProt: Towards multi-modal protein foundation models

URL: View paper

Brief Assessment

OneProt Multimodal[66] focuses on general protein foundation models integrating sequence, structure, text, and binding sites across diverse tasks. The original paper specifically addresses enzyme EC number prediction with hierarchical taxonomy preservation in hyperbolic space, which is not the primary focus of this candidate.

6. Atom level enzyme active site scaffolding using RFdiffusion2

URL: View paper

Brief Assessment

RFdiffusion2 Active Site[62] focuses on de novo enzyme design from active site descriptions using generative AI, not on multi-modal learning that integrates sequence, structure, and active site information for enzyme classification in hyperbolic space.

7. A Highly Sensitive Model Based on Graph Neural Networks for Enzyme Key Catalytic Residue Prediction

URL: View paper

Brief Assessment

Graph Neural Networks Catalytic[69] focuses on predicting enzyme catalytic residue sites using graph neural networks for structural feature characterization, not on multi-modal enzyme learning with hyperbolic space representations or EC number classification.

8. TUNA: A Target-aware Unified Network for Protein-Ligand Binding Affinity Prediction via Multi-Modal Feature Integration.

URL: View paper

Brief Assessment

TUNA[70] focuses on protein-ligand binding affinity prediction using pocket-level features and ligand representations, not enzyme classification with EC number hierarchy in hyperbolic space.

9. A center-anchored adaptive hierarchical graph neural network with application in structure-aware recognition of enzyme catalytic specificity

URL: View paper

Brief Assessment

Center-anchored Hierarchical Graph[63] focuses on hierarchical graph neural networks for enzyme catalytic specificity recognition, but does not demonstrate prior work on hyperbolic multi-modal learning that integrates sequence, structure, and active site information in hyperbolic space. The candidate's limited context does not provide sufficient detail to assess overlap with the original paper's specific approach.

10. MMSite: A Multi-modal Framework for the Identification of Active Sites in Proteins

URL: View paper

Brief Assessment

MMSite[65] focuses on active site identification using protein-attribute text alignment with biomedical language models, not hyperbolic space representations or EC taxonomy preservation. The multi-modal approaches differ fundamentally in their modalities and objectives.

Contribution 2: Multi-modal dataset augmentation with structural and active site information

Description: The authors extend the existing CARE benchmark by adding structural information from PDB and AlphaFold2/ESMFold predictions, along with active site annotations from UniProt. This augmentation transforms the single-modality benchmark into a multi-modal dataset for enzyme classification.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites

URL: View paper

Prior Art Analysis

Multi-modal Active Sites[64] demonstrates prior work in augmenting enzyme datasets with structural information from PDB and AlphaFold2, along with active site annotations from UniProt. The candidate paper explicitly describes constructing a multi-modal dataset by supplementing sequence-only benchmarks with structure information from PDB/AlphaFold2/ESMFold and active site annotations from UniProt, which directly parallels the original paper's claimed contribution of extending CARE benchmark with similar augmentations.

Evidence

Evidence 1 - **Rationale**: Both papers describe integrating protein language models with 3D structural information for enzyme analysis, establishing prior work in multi-modal enzyme representation. - **Original**: we augment the care benchmark with comprehensive structural and active site annotations. while most enzymes can be assigned structures via experimental data or alphafold2/esmfold predictions, experimentally validated active site annotations are available for only a small subset - **Candidate**: we introduce easifa, an enzyme active site annotation algorithm that fuses latentenzyme representations from the protein language model and 3d structural encoder, and then aligns protein-level information with the knowledge of enzymatic reactions using a multi-modal cross-attention framework

2. Enzyme active sites: Identification and prediction of function using computational chemistry

URL: View paper

Brief Assessment

Computational Chemistry Active Sites[76] focuses on computational chemistry methods for predicting active sites in protein structures for enzyme engineering, not on creating multi-modal datasets for enzyme classification benchmarks.

3. SEFP: Structure-Based Enzyme Function Prediction

URL: View paper

Brief Assessment

SEFP[77] focuses on enzyme structure point cloud analysis without dataset augmentation. The candidate does not describe creating or augmenting datasets with structural and active site annotations from PDB, AlphaFold2, ESMFold, or UniProt.

4. EnzyMine: a comprehensive database for enzyme function annotation with enzymatic reaction chemical feature

URL: View paper

Brief Assessment

EnzyMine[74] focuses on mining enzymatic reaction chemical features and linking them to enzyme annotations, rather than augmenting existing benchmarks with structural and active site data for multi-modal classification tasks.

5. Predicting enzymatic function of protein sequences with attention

URL: View paper

Brief Assessment

Attention Enzymatic Function[72] focuses exclusively on sequence-based enzyme prediction using transformer models, without incorporating structural or active site information into their dataset or methodology.

6. A Highly Sensitive Model Based on Graph Neural Networks for Enzyme Key Catalytic Residue Prediction

URL: View paper

Brief Assessment

Graph Neural Networks Catalytic[69] uses structural features and active site information for catalytic residue prediction, not for augmenting enzyme classification datasets with multi-modal annotations from PDB, AlphaFold2/ESMFold, and UniProt.

7. Structure-based activity prediction for an enzyme of unknown function

URL: View paper

Brief Assessment

Structure-based Activity Prediction[75] focuses on predicting enzymatic activity for unknown functions using active site features, not on creating multi-modal datasets for enzyme classification benchmarks.

8. Autoregressive enzyme function prediction with multi-scale multi-modality fusion

URL: View paper

Brief Assessment

Autoregressive Enzyme Prediction[38] integrates sequence and 3D structural tokens but does not mention active site annotations from UniProt or augmentation of the CARE benchmark with such information. The candidate focuses on autoregressive EC number prediction rather than dataset curation with active site features.

9. Protein functional site annotation using local structure embeddings

URL: View paper

Brief Assessment

Local Structure Embeddings[71] focuses on local structural environment embeddings for residue-level functional annotation, not on creating a multi-modal benchmark dataset with structural and active site annotations for enzyme classification.

10. The Computerâ□□Assisted Sequence Annotation (CASA) workflow for enzyme discovery

URL: View paper

Brief Assessment

CASA Workflow[73] focuses on annotation workflow for individual enzyme characterization from sequence data, not on creating multi-modal benchmark datasets for machine learning. The original paper augments the CARE benchmark with structural and active site data for training classification models, while CASA provides tools for annotating novel proteins.

Contribution 3: Graph diffusion mechanism for addressing annotation sparsity

Description: The authors develop pairwise similarity graphs for structure and active site modalities, then apply graph diffusion operations to mitigate data sparsity by incorporating both direct and indirect connections. This approach enriches functional representations despite incomplete modality information.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. LapDDPM: A Conditional Graph Diffusion Model for scRNA-seq Generation with Spectral Adversarial Perturbations

URL: View paper

Brief Assessment

LapDDPM[58] applies graph diffusion to single-cell RNA-seq data for generative modeling, not for addressing annotation sparsity in enzyme functional data with incomplete modality information.

2. Label Diffusion Graph Learning network for semi-supervised breast histological image recognition

URL: View paper

Brief Assessment

Label Diffusion Graph Learning[55] focuses on semi-supervised learning for breast histological image recognition, not biological enzyme data with multi-modal structure/active site annotations. The application domains and data modalities are fundamentally different.

3. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion

URL: View paper

Prior Art Analysis

Label Diffusion Homology[52] demonstrates prior work using graph diffusion to address annotation sparsity in biological data. The candidate paper explicitly describes applying 'label diffusion algorithm' to exploit homology information and account for overlapping communities of proteins with related functions. This directly shows that graph diffusion techniques for handling sparse annotations in biological contexts existed before the original paper's submission. Both papers construct similarity relationships and apply diffusion operations to propagate information and mitigate data sparsity issues.

Evidence

Evidence 1 - **Rationale**: Both papers use graph-based diffusion mechanisms to address sparsity. The candidate explicitly mentions 'label diffusion algorithm' for exploiting relationships between proteins, while the original uses 'graph diffusion' for similar purposes of alleviating annotation sparsity. - **Original**: we construct pairwise similarity graphs for structure and active site modalities, leveraging intra-modality graph diffusion and inter-modality dual-graph alignment to alleviate annotation sparsity and bridge modality gaps - **Candidate**: the prediction is further advanced by exploiting the homology information and accounting for the overlapping communities of proteins with related functions through the label diffusion algorithm

Evidence 2 - **Rationale**: This pair demonstrates that both methods construct graphs and apply diffusion operations to handle sparse annotations. The candidate's 'label diffusion to exploit the homology information' serves the same purpose as the original's 'graph diffusion to address annotation sparsity.' - **Original**: we construct similarity graphs and use graph diffusion to address annotation sparsity and enhance functional representation - **Candidate**: sprof-go applies hierarchical learning strategy to produce consistent predictions and label diffusion to exploit the homology information

Evidence 3 - **Rationale**: Both approaches construct similarity-based graphs and apply diffusion to handle incomplete data. The candidate's label diffusion for 'overlapping communities' parallels the original's graph diffusion for 'alleviating annotation sparsity.' - **Original**: to overcome this, poinncare constructs pairwise similarity graphs for structure and active site modalities, leveraging intra-modality graph diffusion and inter-modality dual-graph alignment to alleviate annotation sparsity and bridge modality gaps - **Candidate**: the prediction is further advanced by exploiting the homology information and accounting for the overlapping communities of proteins with related functions through the label diffusion algorithm

4. BGMSDDA: A bipartite graph diffusion algorithm with multiple similarity integration for drugâ□□disease association prediction

URL: View paper

Brief Assessment

BGMSDDA[56] applies graph diffusion to drug-disease association matrices in a bipartite graph setting, not to multi-modal biological data with incomplete modality annotations. The technical context and application domain differ fundamentally from the original paper's enzyme classification task.

5. Graph Diffusion Network for Drug-Gene Prediction

URL: View paper

Brief Assessment

Graph Diffusion Drug-Gene[51] applies graph diffusion to drug-gene networks for generating negative samples in contrastive learning, not for addressing annotation sparsity in multi-modal biological data like enzyme structures and active sites.

6. Semantically Consistent Discrete Diffusion for 3D Biological Graph Modeling

URL: View paper

Brief Assessment

Discrete Diffusion Biological Graph[59] focuses on discrete diffusion for 3D biological graph generation with semantic consistency constraints, not on graph diffusion operations to address annotation sparsity in multi-modal enzyme data.

7. Single-cell RNA sequencing data imputation using bi-level feature propagation

URL: View paper

Brief Assessment

Bi-level Feature Propagation[54] addresses sparsity in single-cell RNA sequencing data through graph-based feature propagation, not enzyme annotation sparsity. The candidate focuses on technical noise and dropout events in gene expression data, which is a fundamentally different biological domain and data type than enzyme functional annotations.

8. Normalized Laplacian Diffusion for Robust Cancer Pathway Extension and Critical Gene Identification from Limited Data

URL: View paper

Brief Assessment

Laplacian Diffusion Pathway[60] applies normalized Laplacian diffusion to protein interaction networks for pathway extension from limited gene sets, not for addressing incomplete multi-modal enzyme annotations. The technical contexts differ fundamentally: cancer pathway analysis versus enzyme function prediction with missing modality data.

9. Exploiting ontology graph for predicting sparsely annotated gene function

URL: View paper

Brief Assessment

Ontology Graph Prediction[57] addresses annotation sparsity in Gene Ontology (GO) functional labels through information transfer between similar labels, not through graph diffusion operations on pairwise similarity graphs of protein structure and active site modalities as in the original paper.

10. DRGAT: Predicting Drug Responses Via Diffusion-Based Graph Attention Network

URL: View paper

Brief Assessment

DRGAT[53] uses diffusion models for data augmentation in drug response prediction with gene expression data, not for addressing annotation sparsity in multi-modal biological data through graph diffusion operations on similarity graphs.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] PoinnCARE: Hyperbolic Multi-Modal Learning for Enzyme Classification View paper
- [1] EC2Vec: A Machine Learning Method to Embed Enzyme Commission (EC) Numbers into Vector Representations View paper
- [2] EC-Bench: A Benchmark for Enzyme Commission Number Prediction View paper
- [3] Classification of enzymes View paper
- [4] The classification of enzymes by deep learning View paper
- [5] CLAIRE: a contrastive learning-based predictor for EC number of chemical reactions View paper
- [6] Interpretable Kolmogorov-Arnold Networks for Enzyme Commission Number Prediction View paper
- [7] Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures View paper
- [8] A general model for predicting enzyme functions based on enzymatic reactions View paper
- [9] Enhancing Enzyme Commission Number Prediction With Contrastive Learning and Agent Attention View paper
- [10] PredictEFC: a fast and efficient multi-label classifier for predicting enzyme family classes View paper
- [11] Comparative Assessment of Protein Large Language Models for Enzyme Commission Number Prediction View paper
- [12] A Binary Classifier for the Prediction of EC Numbers of Enzymes View paper
- [13] Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers View paper
- [14] Practical limits of function prediction View paper
- [15] EnzymeNet: residual neural networks model for Enzyme Commission number prediction View paper
- [16] ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature View paper
- [17] HiFi-NN annotates the microbial dark matter with enzyme commission numbers View paper
- [18] A classification of glycosyl hydrolases based on amino acid sequence similarities View paper
- [19] EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation View paper
- [20] ECOH: an enzyme commission number predictor using mutual information and a support vector machine View paper
- [21] Ec-psi: Associating enzyme commission numbers with pfam domains View paper
- [22] HIT-EC: Trustworthy prediction of enzyme commission numbers using a hierarchical interpretable transformer View paper

- [23] Multimodal Quantum Vision Transformer for Enzyme Commission Classification from Biochemical Representations View paper
- [24] Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions View paper
- [25] Interpretable Enzyme Function Prediction via Residue-Level Detection View paper
- [26] Predicting Enzyme Commission Numbers Using Recurrent Neural Networks with Amino Acid Sequence Shift and Consistency Loss View paper
- [27] TopEC: prediction of Enzyme Commission classes by 3D graph neural networks and localized 3D protein descriptor View paper
- [28] Improved Functional Classification of Hydrolases through Pairwise Structural Similarity of Reaction Cores View paper
- [29] GlycoEnzDB: A database of enzymes involved in human glycosylation View paper
- [30] ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains View paper
- [31] Prediction of enzyme function using an interpretable optimized ensemble learning framework View paper
- [32] NickFlagleaf/GxE-FA-EC-Prediction: v1.0 View paper
- [33] Enzyme Commission Number Prediction and Benchmarking with Hierarchical Dual-core Multitask Learning Framework View paper
- [34] E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs View paper
- [35] Hierarchical Clustering-Based Coarse-to-Fine Classification Framework for Microbial Protein Function Prediction View paper
- [36] CACLENS: A Multitask Deep Learning System for Enzyme Discovery. View paper
- [37] Leveraging large language models for enzymatic reaction prediction and characterization View paper
- [38] Autoregressive enzyme function prediction with multi-scale multi-modality fusion View paper
- [39] SST-ResNet: A Sequence and Structure Information Integration Model for Protein Property Prediction View paper
- [40] CAPIM: Catalytic activity and site prediction and analysis tool in multimer proteins View paper
- [41] A Machine Learning Approach for Plant-based Drug Discovery: High-Throughput Prediction of Biological Activities and Enzyme Commission Numbers from Phytochemicals and Amino Acid Sequences of Plants View paper
- [42] Prediction of Enzyme function using interpretable optimized Ensemble learning framework View paper
- [43] A brief guide to enzyme nomenclature and classification View paper
- [44] EZpred: improving deep learning-based enzyme function prediction using unlabeled sequence homologs View paper
- [45] Machine Learning-Driven Enzyme Mining: Opportunities, Challenges, and Future Perspectives View paper
- [46] Decoding the dark proteome: Deep learning-enabled discovery of druggable enzymes in Wuchereria bancrofti View paper
- [47] naomifridman/transformer-based-enzyme-classification: Initial Release: ESM2 Enzyme Classification View paper
- [48] Limitations of current machine learning models in predicting enzymatic functions for uncharacterized proteins View paper
- [49] A comparison of prediction methods for the creation of field-extent soil property maps View paper
- [50] Predicting enzyme class from protein structure without alignments View paper
- [51] Graph Diffusion Network for Drug-Gene Prediction View paper
- [52] Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion View paper
- [53] DRGAT: Predicting Drug Responses Via Diffusion-Based Graph Attention Network View paper
- [54] Single-cell RNA sequencing data imputation using bi-level feature propagation View paper
- [55] Label Diffusion Graph Learning network for semi-supervised breast histological image recognition View paper
- [56] BGMSDDA: A bipartite graph diffusion algorithm with multiple similarity integration for drugâ□□disease association prediction View paper
- [57] Exploiting ontology graph for predicting sparsely annotated gene function View paper
- [58] LapDDPM: A Conditional Graph Diffusion Model for scRNA-seq Generation with Spectral Adversarial Perturbations View paper
- [59] Semantically Consistent Discrete Diffusion for 3D Biological Graph Modeling View paper
- [60] Normalized Laplacian Diffusion for Robust Cancer Pathway Extension and Critical Gene Identification from Limited Data View paper
- [61] OneProt: Towards multi-modal protein foundation models via latent space alignment of sequence, structure, binding sites and text encoders View paper
- [62] Atom level enzyme active site scaffolding using RFdiffusion2 View paper
- [63] A center-anchored adaptive hierarchical graph neural network with application in structure-aware recognition of enzyme catalytic specificity View paper
- [64] Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites View paper
- [65] MMSite: A Multi-modal Framework for the Identification of Active Sites in Proteins View paper
- [66] OneProt: Towards multi-modal protein foundation models View paper
- [67] Bidirectional Hierarchical Protein Multi-Modal Representation Learning View paper
- [68] A multimodal Transformer Network for protein-small molecule interactions enhances predictions of kinase inhibition and enzyme-substrate relationships View paper
- [69] A Highly Sensitive Model Based on Graph Neural Networks for Enzyme Key Catalytic Residue Prediction View paper
- [70] TUNA: A Target-aware Unified Network for Protein-Ligand Binding Affinity Prediction via Multi-Modal Feature Integration. View paper
- [71] Protein functional site annotation using local structure embeddings View paper
- [72] Predicting enzymatic function of protein sequences with attention View paper
- [73] The Computerâ□□Assisted Sequence Annotation (CASA) workflow for enzyme discovery View paper
- [74] EnzyMine: a comprehensive database for enzyme function annotation with enzymatic reaction chemical feature View paper
- [75] Structure-based activity prediction for an enzyme of unknown function View paper
- [76] Enzyme active sites: Identification and prediction of function using computational chemistry View paper
- [77] SEFP: Structure-Based Enzyme Function Prediction View paper