Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Systematic benchmarking of molecular foundation models for natural-product annotation and biosynthetic inference

Fan Lin

fanlin@fyust.edu.cn

Fuyao University of Science and Technology

Shu Xing

Jilin University

Yuqiang Han

Chinese Academy of Sciences

Dong Wei

Fuyao University of Science and Technology

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11357474/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Systematic benchmarking of molecular foundation models for natural-product annotation and biosynthetic inference

Fan Lin[1,2*], Shu Xing[4], Yuqiang Han[3*], Dong Wei[1*]

[1*]School of Life and Health Sciences, Fuyao University of Science and Technology, 104 Zhihui Avenue, Fuzhou, 350100, Fujian, China.
[2]School of Information, University of California, Berkeley, 102 South Hall Rd, Berkeley, 94720, CA, United States.
[3]Hangzhou Institute of Medicine, Chinese Academy of Sciences, 150 Fucheng Road , Hangzhou, 310018, Zhejiang, China.
[4]School of Life Science, Jilin University, 2699 Qianjin Street, Changchun, 130012, Jilin, China.

*Corresponding author(s). E-mail(s): fanlin@fyust.edu.cn; hanyuqiang@him.cas.cn; dweimj@fyust.edu.cn;

Abstract

Molecular foundation models offer a potential route to improve natural-product annotation and support biosynthetic inference, but their utility across these tasks remains insufficiently characterized. We benchmarked eight foundation-model representations against Morgan fingerprints across five tasks covering natural-product-likeness, natural product chemical class prediction, PMN and KEGG pathway association, and PlantCyc reaction-neighbor prediction. Models were evaluated under random and scaffold-based splits using both full data and sample-efficiency settings. In the full data setting, MolFormer achieved the highest mean AUROC for natural product likeness prediction, while Morgan fingerprints were strongest across other tasks. In the sample efficiency analyses, pretrained foundation model representations showed advantages when labeled data were scarce across several tasks. Even in the PMN pathway task, where foundation models did not outperform Morgan fingerprints individually, their embeddings provided complementary predictive information, as shown by error pattern and ensemble analyses. These results show that molecular foundation models are particularly useful for tasks related to global structural characteristics

and when labeled training data are limited. This benchmark extends the evaluation of molecular foundation models beyond conventional drug-property tasks and provides a practical reference for selecting and developing representations for natural-product annotation and biosynthetic inference.

Scientific Contribution
This study benchmarks molecular foundation-model representations for natural-product annotation and biosynthesis-related prediction, extending their evaluation beyond conventional molecular-property tasks. It identifies task-dependent advantages, particularly in low-data settings, and demonstrates the complementarity of pretrained representations and conventional molecular fingerprints.

Keywords: natural products, pathway association, molecular foundation models, sample efficiency, scaffold splitting

1 Introduction

Natural products (NPs) and their derivatives remain an important source of therapeutics and other biologically active products [1, 2]. A quantitative review of 1,394 small-molecule drugs approved worldwide between 1981 and September 2019 found that NPs, botanical drugs, and direct NP derivatives accounted for 441 approvals (32%) [1]. However, the practical use of some NPs is constrained by low natural abundance, variable supply, and the difficulty of producing structurally complex molecules through conventional chemical synthesis [2–4]. Elucidating the enzymes, intermediates, and reaction sequence of an NP biosynthetic pathway can address these limitations by enabling pathway reconstruction and optimization in heterologous hosts [3, 4]. Beyond enabling production, biosynthetic knowledge can support the diversification of natural products through pathway engineering, including modification of biosynthetic enzymes to generate structural analogues [4].

Despite these benefits, experimental elucidation of NP pathways remains slow. Complete biosynthetic pathways, including their intermediate reactions, remain unresolved for most known natural products [5]. Traditionally, candidate pathway genes are identified through approaches such as gene co-expression analysis, physical gene clustering, comparative genomics, and correlations between transcript and metabolite abundance [6–8]. Candidate enzymes must then be validated individually through gene perturbation, biochemical assays, or heterologous expression and pathway reconstitution [6, 7]. Natural-product pathways can contain multiple uncharacterized enzymes and undefined intermediates, and the corresponding genes are not always physically clustered [9, 10]. Computational predictions that connect molecular structures with biosynthetically meaningful classes, pathway associations, and plausible reaction neighbors could reduce the search space for candidate pathways and transformations [11–13]. Such predictions do not reconstruct pathways independently, but can prioritize hypotheses for integration with genomic, transcriptomic, enzymatic, taxonomic, and metabolomic evidence.

Molecular foundation models offer a potential means of extracting such information from molecular structure. These models are pretrained on large collections of unlabeled molecules represented as SMILES strings or molecular graphs. SMILES language models commonly use masked-token prediction. To recover a masked token, the transformer integrates information from the surrounding SMILES context, including tokens encoding atoms, bonds, branches, and ring closures, producing representations that depend on their molecular context [14, 15]. Graph-based models instead propagate information across atom–bond neighborhoods and have been pretrained using objectives such as masked-atom prediction, contrastive learning between augmented molecular graphs, and alignment between 2D topology and 3D geometry [16–18]. NaFM further adapts graph pretraining to natural products through scaffold–subgraph reconstruction and scaffold-aware contrastive learning [19]. These objectives can encode substructures and their broader molecular contexts before task-specific labels are introduced. Most original downstream evaluations have focused on drug discovery and general molecular property prediction, including bioactivity, toxicity, ADME, and physicochemical endpoints [14–18]. Whether the resulting representations are also informative for natural-product biosynthesis remains less well established.

There are reasons to expect molecular pretraining to contain information relevant to natural-product classification and pathway inference. Biosynthetic origin leaves recognizable structural signatures: broad NP classes are associated with characteristic precursors and scaffold families, pathway membership is often correlated with combinations of scaffolds and functional groups, and neighboring metabolites are connected through specific chemical transformations. Structure-based approaches have already demonstrated that NP classes and metabolic-pathway labels can be predicted to some extent from molecular representations [11, 12]. Molecular FMs may therefore encode structural correlates of biosynthetic organization even though pathway labels are not used during general-purpose pretraining.

Here, we systematically test whether pretrained molecular representations encode biosynthesis-related information and provide value beyond Morgan fingerprints. We benchmark their performance across tasks spanning different levels of biosynthetic information and examine whether pretrained embeddings provide complementary predictive information.

2 Methods

We benchmarked molecular foundation-model representations across five supervised tasks. These comprised binary natural-product-likeness prediction, seven-class natural-product chemical-class prediction, multilabel pathway-association prediction using PMN- and KEGG-derived datasets, and binary prediction of direct reaction-neighbor relationships in PlantCyc.

The primary analysis used pretrained molecular models as frozen feature extractors, with task-specific linear classifiers trained on the resulting representations. Foundation-model embeddings were compared with Morgan fingerprints using

matched data partitions and model-selection procedures. Performance was evaluated under full-data and sample-efficiency settings using repeated random and scaffold-based splits.

2.1 Datasets and data splitting

Molecular structures were parsed with RDKit. For each valid input structure, we generated a canonical isomeric SMILES string, an InChIKey, and a Bemis–Murcko scaffold [20]. Invalid or unparsable structures were excluded. Compounds were deduplicated using InChIKey as the primary structure identifier, with canonical SMILES used as a fallback. Annotations associated with duplicate records were combined according to the labeling structure of each task.

The five benchmark datasets, their prediction formulations, and final sample sizes are summarized in Table 1. For the natural-product-likeness task, structures associated with conflicting binary labels after deduplication were excluded, and the majority class was downsampled to obtain a balanced dataset of 411,448 compounds.

The PMN pathway-association benchmark was constructed from two compound sets reported by Bao et al. [12]. The larger set was used for model development, including training and validation, whereas the second set was reserved as an external test set. Although the external test set was originally described as containing newly added compounds absent from the development set, structure-level comparison identified 160 compounds shared between the two sets. These compounds were excluded to ensure compound-level independence of the external test set.

The PlantCyc reaction-neighbor task was constructed from curated substrate–product relationships in PlantCyc version 17.0.0 [21]. Following structure filtering, pair deduplication, and final task construction, 5,222 unique positive pairs remained. After positive-pair partitioning, three balanced negative-set variants were generated independently within each split. Easy negatives were non-neighbor pairs with no known shared PlantCyc pathway, whereas hard negatives were preferentially sampled from same-pathway non-neighbors, with same-scaffold non-neighbors used as a fallback. Mixed sets contained approximately equal numbers of easy and hard negatives. One negative was sampled per positive pair, yielding 10,444 pairs in each balanced variant.

For the natural-product-likeness and natural-product class tasks, random training, validation, and test partitions were generated in an approximately 80:10:10 ratio using class-stratified splitting. For the KEGG pathway task, the corresponding partitions were generated using iterative multilabel stratification. For the PMN task, iterative multilabel stratification was used to divide the development set into training and validation partitions under random splitting.

For scaffold-based partitions, compounds sharing a Bemis–Murcko scaffold were assigned to the same partition. Because acyclic compounds have an empty Murcko scaffold, molecule-specific split keys were used for these structures rather than treating all acyclic compounds as one scaffold group. Scaffold overlap among partitions generated from each development dataset was prohibited. For the PMN task, random and scaffold-based splitting were applied only to the development set to generate training

and validation partitions, while the external test set remained fixed for both evaluation schemes. Compound overlap between the development and external test sets was removed, but scaffold disjointness between these sets was not enforced.

2.2 Molecular representations

Morgan circular fingerprints were generated with RDKit using radii of 2 and 3 and a fixed length of 2,048 bits [22]. The radius-2 fingerprint was used as the primary conventional baseline, while the radius-3 fingerprint was included as an additional fingerprint representation.

Eight pretrained molecular representations were evaluated: ChemBERTa-77M-MLM, ChemBERTa-100M-MLM, MolFormer, MolCLR-GCN, MolCLR-GIN, GraphMVP-GIN, Mole-BERT-GIN, and NaFM [14–19]. All pretrained parameters were kept fixed during embedding extraction.

For ChemBERTa and MolFormer, canonical SMILES were tokenized using the corresponding model-specific tokenizers, and final-layer token representations were mean-pooled over non-padding positions. For MolCLR, the graph-level encoder output before the contrastive projection head was used. For GraphMVP and Mole-BERT, node representations from the pretrained GIN encoders were aggregated using mean pooling. For NaFM, node representations from the pretrained encoder were mean-pooled to obtain a graph-level embedding. The resulting embeddings were aligned to task compounds using InChIKey.

2.3 Downstream modeling and model selection

Frozen molecular representations were evaluated using an L2-regularized linear classification head. Dense foundation-model embeddings were standardized using parameters estimated from the training partition, whereas Morgan fingerprints were used without scaling.

Natural-product-likeness and PlantCyc reaction-neighbor prediction were formulated as binary classification. Natural-product class prediction used multiclass logistic regression, with the class having the highest predicted probability selected as the final prediction. PMN and KEGG pathway association were modeled using one-versus-rest logistic regression, with one binary classifier fitted for each pathway label.

For PlantCyc reaction-neighbor prediction, order-invariant pair features were constructed from the representations of the two compounds. Dense embeddings were represented by concatenating their element-wise sum and absolute difference. Morgan-fingerprint pairs were represented by concatenating their bitwise intersection and exclusive difference.

The logistic-regression regularization parameter was selected using validation AUROC for binary tasks and validation macro-AUROC for multiclass and multilabel tasks. Test data were not used for feature scaling and hyperparameter selection.

2.4 Sample-efficiency analysis

Sample efficiency was evaluated by subsampling only the training partition, while keeping the corresponding validation and test partitions fixed. Within each data split and

sampling replicate, training subsets were nested, and the same sampled compounds were used across all molecular representations.

For the binary and multiclass tasks, subsets were sampled to approximately preserve the class distribution. For the multilabel PMN and KEGG tasks, training level was defined by the minimum number of positive examples requested per pathway. Multiple independent sampling replicates were evaluated at each training level.

2.5 Probability ensembles

To assess whether foundation-model representations contain information complementary to Morgan fingerprints, probability-level ensembles were evaluated for the PMN pathway tasks. Pairwise ensembles were formed by equally averaging the predicted probabilities from Morgan r2 and one foundation-model classifier for each pathway label. An additional all-representation ensemble averaged probabilities from all evaluated classifiers.

2.6 Parameter-efficient adaptation

ChemBERTa-77M-MLM and MolFormer were selected as representative Transformer-based models, allowing the same LoRA adaptation protocol to be applied consistently to both. LoRA adaptation was evaluated on the natural-product-likeness and natural-product class tasks. LoRA adapters were applied to the query and value projection modules with rank 8, scaling factor 16, and dropout 0.1. The adapter parameters and a newly initialized task-specific classification head were trained while the remaining pretrained-model parameters were kept fixed.

Models were trained separately for each random and scaffold split for up to three epochs, by which point the training loss had largely plateaued. The checkpoint with the highest validation AUROC was retained for natural-product-likeness prediction, whereas validation macro-AUROC was used for natural-product class prediction. Test data were evaluated only after checkpoint selection.

2.7 Evaluation and statistical analysis

Performance was evaluated using AUROC for the binary natural-product-likeness and PlantCyc reaction-neighbor tasks and macro-averaged one-versus-rest AUROC for the multiclass natural-product class and multilabel pathway-association tasks.

Full-data results were summarized across five split seeds. Means and 95% confidence intervals were calculated across seed-level estimates. For sample-efficiency experiments, repeated training subsets were first averaged within each split seed, and uncertainty was then calculated across the split-level means.

For selected comparisons, performance differences were calculated within matched split seeds. For sample-efficiency analyses, sampling replicates were first averaged within each split seed. Mean paired differences and 95% confidence intervals across split-level differences are reported.

3 Results

3.1 Overview of benchmark tasks and evaluation

We evaluated eight molecular foundation-model representations and compared them with Morgan fingerprints across five downstream tasks spanning natural-product annotation and biosynthetic context (Table 1). Natural-product-likeness was used to assess whether molecular representations capture structural information associated with biological versus synthetic origins. NP chemical-class prediction assessed whether molecular representations capture broad structural signatures associated with major natural-product biosynthetic origins. Fine-grained pathway association was evaluated using a plant-focused dataset derived from the Plant Metabolic Network (PMN) and a broader dataset derived from KEGG, which is not restricted to plant metabolism. Finally, PlantCyc reaction-neighbor prediction assessed whether molecular representations could recover local metabolic connectivity by distinguishing curated reaction-linked compound pairs from non-neighbor pairs, a capability relevant to prioritizing candidate missing pathway connections.

Performance was evaluated under two settings: full-data training and sample-efficiency analysis. Performance was evaluated using five different splits under both random and scaffold-based partitions. Sample-efficiency experiments evaluated how performance changed as the amount of labeled training data increased.

Table 1 Overview of the datasets used for downstream tasks.

Task	Label structure	Data size[a]	Label source
NP-likeness	Binary	411,448 compounds	Menke et al. [23]
NP class	Seven-class single-label	78,331 compounds	Kim et al. [11]
PMN pathway	18-label multilabel	2,431 compounds	Bao et al. [12]
KEGG pathway	11-label multilabel	5,288 compounds	Bao et al. [12]
Reaction neighbor	Binary compound pair	10,444 pairs	PlantCyc [21]

[a] Counts refer to the final processed benchmark datasets and may differ from source-reported counts because of structure validation, deduplication, overlap removal, class balancing, and negative-pair construction. See Methods for task-specific preprocessing and dataset construction.

3.2 Predicting natural-product likeness

The natural-product-likeness prediction task evaluates whether a molecular representation captures broad structural characteristics associated with biosynthetic origin. Under the full-data random split, MolFormer achieved the highest mean AUROC, with Morgan r2 performing similarly (Fig. 1A). The difference became more pronounced under scaffold splitting, where MolFormer retained similar performance while both Morgan fingerprint baselines declined. MolFormer achieved the highest mean AUROC

under the scaffold split and showed little performance reduction relative to the random split. Under scaffold splitting, MolFormer achieved higher mean AUROC than Morgan r2 (ΔAUROC = 0.0118, 95% CI: 0.0008 to 0.0227).

The sample-efficiency analysis showed that several pretrained representations supported strong prediction performance with limited labeled data (Fig. 1B–C). At small training-set sizes, multiple foundation-model embeddings achieved higher mean AUROC than the Morgan fingerprint baselines. These differences narrowed as the number of training examples increased, and the performance of most representations converged at larger training-set sizes.

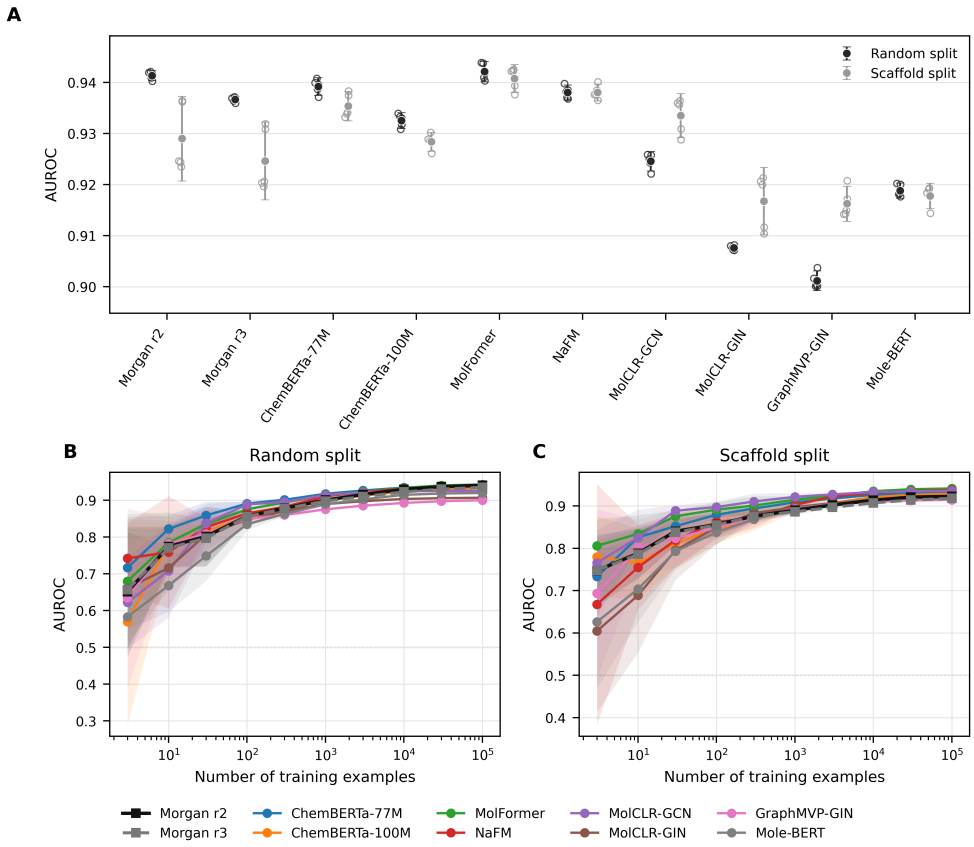


Fig. 1 Performance of molecular representations for natural-product-likeness prediction. (A) Full-data AUROC of molecular fingerprint baselines and frozen foundation-model representations under random and scaffold splits. Open symbols represent individual splits, while filled symbols and error bars represent the mean and 95% confidence interval across split seeds. (B–C) Sample-efficiencyperformance on the test sets under random and scaffold splits, respectively. Lines represent mean AUROC and shaded regions represent 95% confidence intervals across split-level means, after averaging sampling replicates within each split.

3.3 Predicting natural-product chemical classes

With the full training set, Morgan fingerprints achieved the highest macro AUROC under both random and scaffold splits (Fig. 2A). Their performance was close to 1.0, whereas all foundation-model embeddings showed lower mean performance. ChemBERTa-77M, MolFormer, and Mole-BERT were the strongest foundation-model representations. Scaffold splitting reduced performance for most representations, although the magnitude of the reduction was small.

The sample-efficiency results showed a different pattern (Fig. 2B–C). At 10 training compounds, ChemBERTa-77M achieved the highest mean macro-AUROC under both split schemes. Relative to Morgan r2, its mean advantage was 0.040 under the random split (95% CI: 0.014 to 0.065) and 0.044 under the scaffold split (95% CI: 0.029 to 0.060). As the number of training examples increased, the Morgan baselines caught up with or surpassed the frozen embeddings, and the differences between representations narrowed. Thus, for this task, Morgan fingerprints were strongest with abundant labeled data, whereas the clearest advantage of a pretrained representation was observed for ChemBERTa-77M in the lowest-data settings.

3.4 Predicting metabolic pathway associations

Fine-grained pathway prediction tasks test whether molecular representations capture structural features associated with the metabolic contexts in which compounds occur. We evaluated this using two datasets with different biological scopes. The PMN-derived task contains relatively fine-grained, plant-focused pathway labels, whereas the KEGG-derived task uses metabolic categories from a broad range of organisms. Together, these datasets allowed us to examine whether representation performance was consistent across different pathway definitions and scopes.

In the PMN-derived 18-label task, Morgan fingerprints achieved the highest full-data macro-AUROC on the external test set under both random and scaffold splits (Fig. 3A). MolCLR-GCN was the strongest individual foundation-model representation under both split schemes, although it remained below the Morgan baselines.

In the sample-efficiency analysis, GraphMVP-GIN achieved the highest mean macro-AUROC with one training example per pathway under both random and scaffold splitting (Fig. 3B–C). Under the random split, Morgan r2 matched or surpassed GraphMVP-GIN as the training level increased. Under scaffold splitting, GraphMVP-GIN retained the highest mean performance across the evaluated training levels, with a mean advantage of 0.0182 in macro-AUROC over Morgan r2 at two training examples per pathway (95% CI: 0.0001 to 0.0363).

Pathway-level AUROC revealed substantial heterogeneity that was not apparent from the macro-AUROC alone (Fig. 4). Morgan fingerprints were particularly strong for several pathway such as terpenophenolic and polyketide biosynthesis, whereas most foundation-model representations showed lower performance for these labels. Conversely, selected foundation-model representations achieved higher pathway-specific AUROC than Morgan fingerprints for terpenoid degradation, flavonoid degradation, and aroma-compound biosynthesis, with similar patterns under random and scaffold

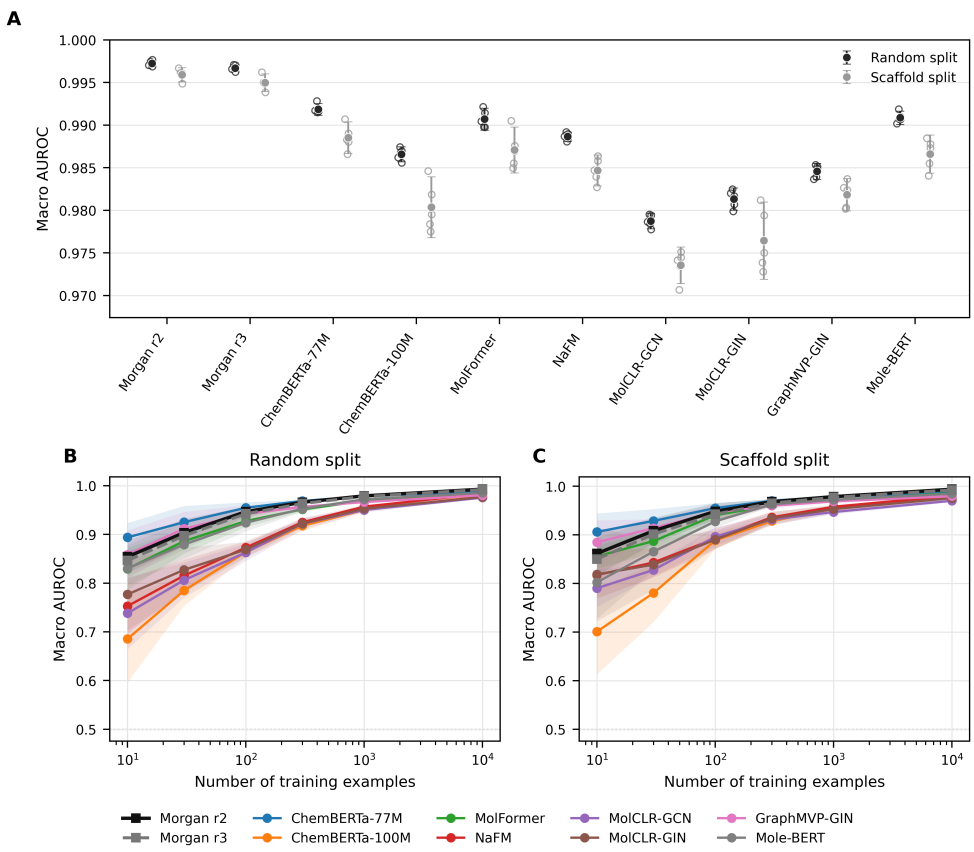


Fig. 2 Performance of molecular representations for broad natural-product class prediction. (A) Full-data macro-AUROC comparison between molecular fingerprint baselines and frozen foundation-model representations under random and scaffold splits. Open symbols represent individual splits, while filled symbols and error bars represent the mean and 95% confidence interval across split seeds. (B–C) Sample-efficiency performance under random and scaffold splits, respectively. Lines represent mean macro-AUROC and shaded regions represent 95% confidence intervals across repeated training-set samples

splitting. Hormone biosynthesis remained comparatively difficult for nearly all representations, indicating that neither fingerprints nor the evaluated foundation-model embeddings captured this label consistently well. Several pathways, including sulfur-containing secondary-compound biosynthesis and sugar-derivative biosynthesis were predicted accurately by nearly all representations.

We also evaluated the representations on the KEGG-derived 11-label task, which covers broader metabolic pathway categories (Fig. 5A). Morgan fingerprints achieved the highest mean macro-AUROC under both random and scaffold splits. MolFormer was the strongest individual foundation-model representation under the random split, whereas ChemBERTa-77M was strongest under the scaffold split.

Morgan fingerprints were strongest across most of the KEGG sample-efficiency experiment, although ChemBERTa-77M showed a small mean advantage at the smallest training level under the random split (Fig. 5B–C).

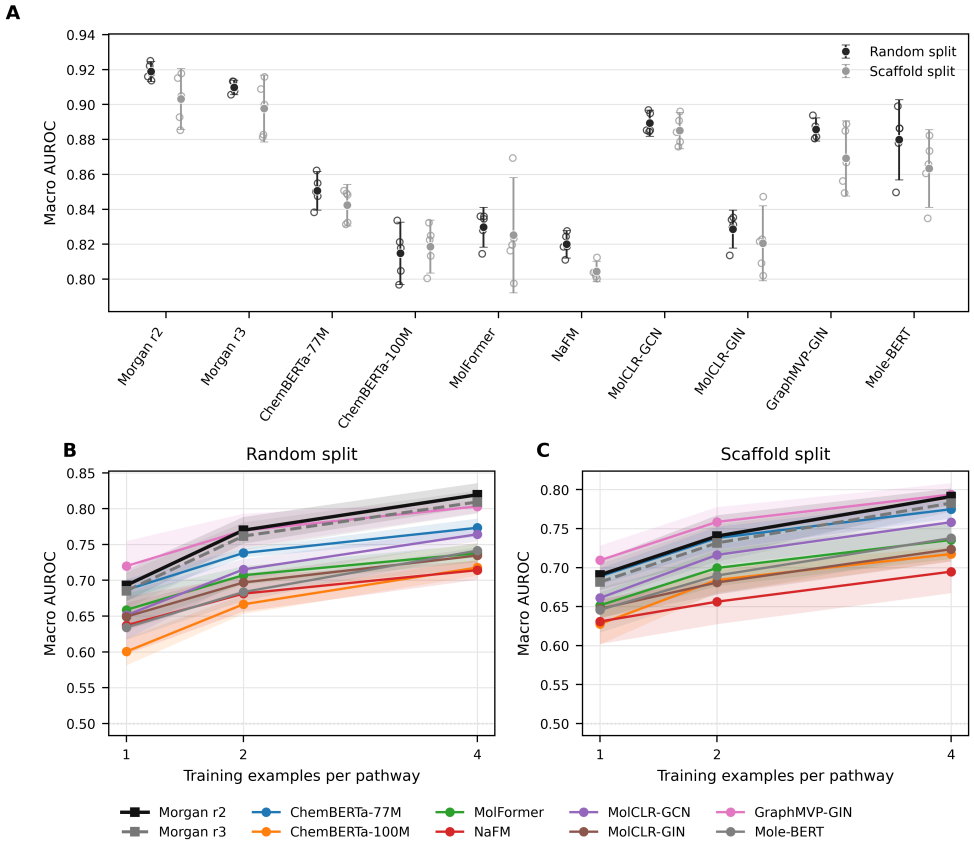


Fig. 3 Performance of molecular representations for PMN-derived secondary-pathway prediction. (A) Full-data macro-AUROC on the PMN test set. Open symbols represent individual splits, while filled symbols and error bars represent the mean and 95% confidence interval across split seeds. (B–C) Sample-efficiency performance on the external test set. The x-axis denotes the minimum number of positive training examples sampled for each pathway label. Lines represent mean macro-AUROC and shaded regions represent 95% confidence intervals.

3.5 Predicting direct reaction neighbors in PlantCyc

Reaction-neighbor prediction evaluates whether molecular representations can distinguish compound pairs connected by a direct metabolic reaction from synthetically generated non-neighbor pairs. Unlike the preceding tasks, reaction-neighbor prediction tests whether molecular representations capture more specific, local structural

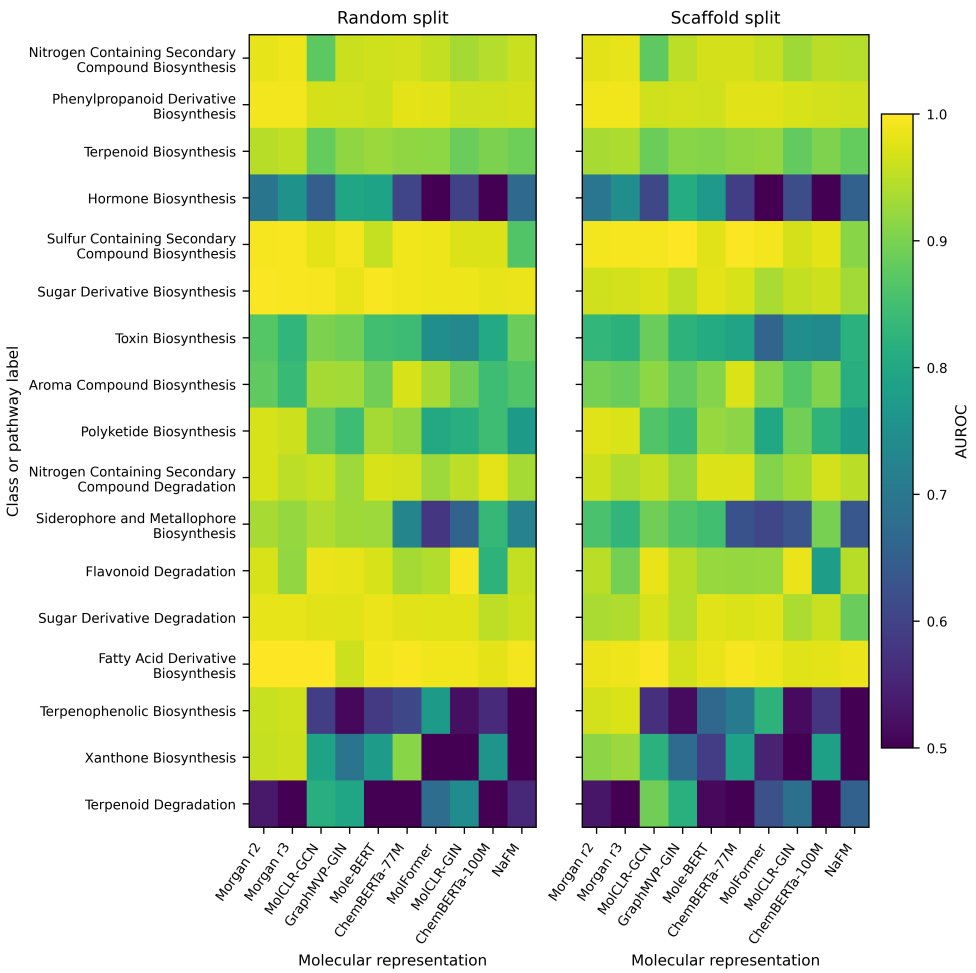


Fig. 4 Pathway-level performance of molecular representations for PMN-derived secondary-pathway prediction. Rows correspond to the 18 pathway labels and columns correspond to the evaluated Morgan fingerprint and all foundation-model representations. Cell color indicates pathway-specific AUROC, averaged across splits. Results are shown separately for models developed using random and scaffold-based partitions of the PMN development set and evaluated on the test set.

patterns associated with individual biochemical transformations within a pathway. Synthetic negative pairs were generated at three difficulty levels: easy negatives were randomly sampled non-neighbor pairs, hard negatives were non-neighbor compounds from the same pathway, and mixed negatives combined both types.

Performance decreased progressively as the negative-pair construction became more challenging (Fig. 6). All representations achieved high AUROC on the easy negative set, with Morgan fingerprints showing the highest mean performance under both random and scaffold-pair splits. The differences between representations became

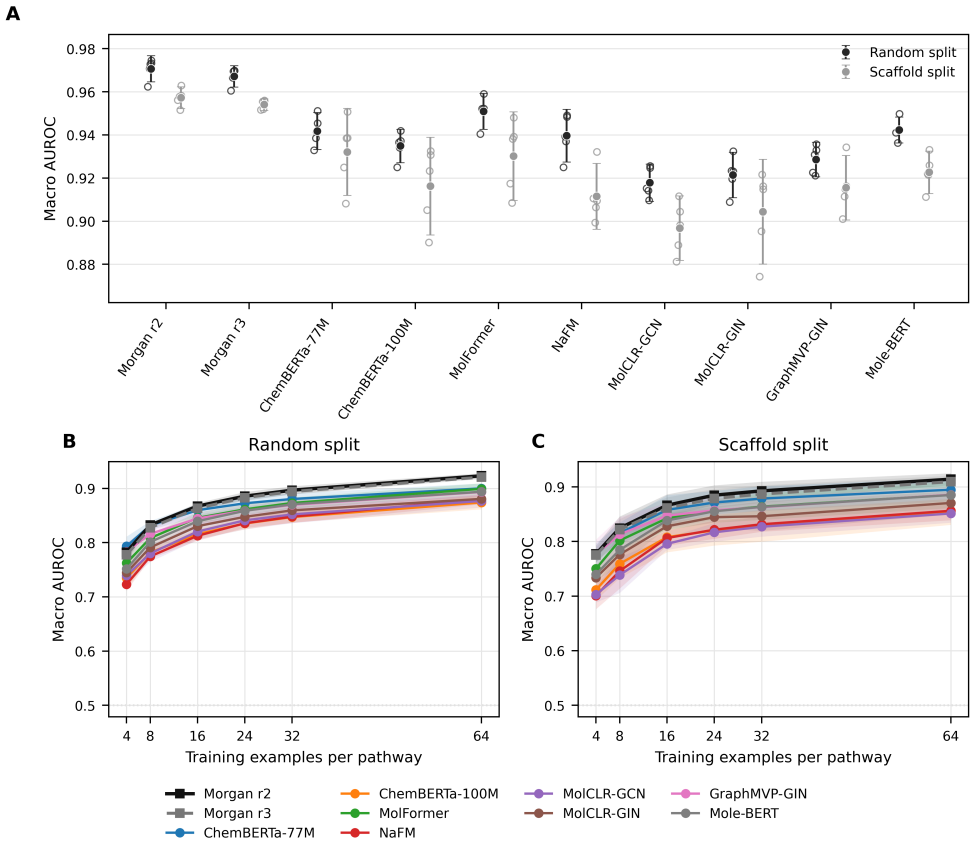


Fig. 5 Performance of molecular representations for KEGG pathway prediction. (A) Full-data macro-AUROC comparison between molecular fingerprint baselines and frozen foundation-model representations under random and scaffold splits. Open symbols represent individual split seeds, while filled symbols and error bars represent the mean and 95% confidence interval across split seeds. (B–C) Sample-efficiency performance under random and scaffold splits, respectively. The x-axis denotes the minimum number of positive training examples sampled for each pathway label. Lines represent mean macro-AUROC and shaded regions represent 95% confidence intervals.

more pronounced for the mixed and hard negative sets. Morgan fingerprints remained the strongest representations, whereas most foundation-model embeddings showed substantially lower AUROC, particularly under the scaffold-pair split. Scaffold-pair splitting generally reduced performance relative to random splitting, with the largest reductions observed for the mixed and hard negative sets. These results show that distinguishing direct reaction neighbors is considerably more difficult than separating structurally dissimilar compound pairs.

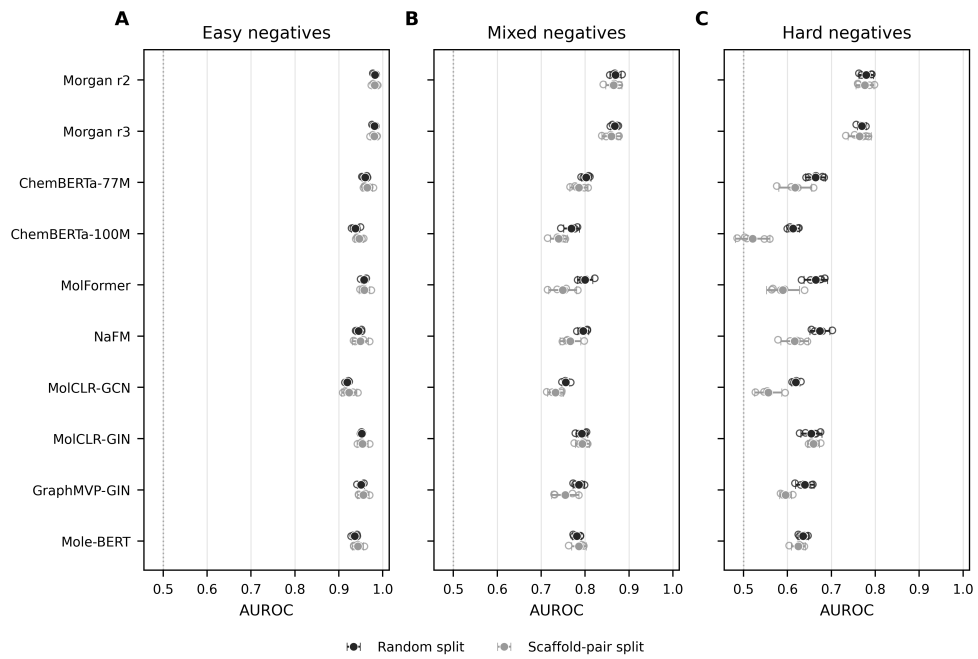


Fig. 6 Performance of molecular representations for PlantCyc reaction-neighbor prediction. Performance is reported as AUROC for easy, mixed, and hard negative-set variants. Within each panel, open symbols represent individual splits, while filled symbols and horizontal error bars represent the mean and 95% confidence interval across split seeds. Results are shown under random and scaffold-pair evaluation schemes.

3.6 Cross-task comparison of molecular representations

Across the downstream tasks, the relative performance of molecular representations depended on both the prediction problem and the amount of labeled training data (Table 2). Morgan fingerprints were generally the strongest full-data representations, particularly for chemical-class, pathway-association, and reaction-neighbor prediction. The main exception was NP-likeness prediction, for which MolFormer achieved the highest mean performance under both random and scaffold splits. Among the foundation-model representations, the strongest-performing model varied across downstream tasks. Foundation-model representations showed their clearest advantages in the sample-efficiency experiments: selected models exceeded the Morgan baselines for NP-likeness, NP chemical-class prediction, and the PMN pathway task when the training set was small.

As an extension of the main analysis based on frozen representations, we conducted a narrower secondary evaluation of task-specific adaptation for natural-product-likeness and broad NP chemical-class prediction. We applied LoRA to ChemBERTa-77M and MolFormer, two of the strongest-performing foundation models on these tasks when evaluated using frozen embeddings. For natural-product-likeness prediction,

Table 2 Summary of representation performance across downstream tasks.

Task	Split	Best FM/FP	Best FM	FM sample efficiency
NP-likeness	Random	MolFormer	MolFormer	Yes
NP-likeness	Scaffold	MolFormer	MolFormer	Yes
NP chemical class	Random	Morgan r2	ChemBERTa-77M	Yes
NP chemical class	Scaffold	Morgan r2	ChemBERTa-77M	Yes
PMN pathway	Random	Morgan r2	MolCLR-GCN	Yes
PMN pathway	Scaffold	Morgan r2	MolCLR-GCN	Yes
KEGG pathway	Random	Morgan r2	MolFormer	Yes
KEGG pathway	Scaffold	Morgan r2	ChemBERTa-77M	No
Reaction neighbor	Random	Morgan r2	easy: ChemBERTa-77M mixed: ChemBERTa-77M hard:NaFM	No
Reaction neighbor	Scaffold	Morgan r2	easy: ChemBERTa-77M mixed: MolCLR-GIN hard: MolCLR-GIN	No

Best-performing representations were identified from the highest mean AUROC or macro AUROC. FM sample efficiency indicates that at least one foundation-model representation had a higher mean AUROC than the strongest Morgan baseline.

both adapted models outperformed their corresponding frozen representations, showing that task-specific adaptation could provide additional predictive benefit beyond the frozen embeddings. In contrast, the effects on chemical-class prediction were smaller and model dependent, with modest improvement for MolFormer but not ChemBERTa-77M. These results indicate that the benefit of task-specific adaptation depends on both the downstream task and the pretrained model (Appendix A.2).

3.7 Foundation-model predictions complement molecular fingerprints in pathway prediction

In both pathway-prediction tasks, Morgan fingerprints achieved higher overall performance than all individual frozen foundation-model representations. To examine whether a foundation-model representation nevertheless captured information complementary to the morgan fingerprint, we compared Morgan r2 with GraphMVP-GIN in the PMN pathway task. GraphMVP-GIN was selected because it showed the strongest foundation-model performance in the PMN sample-efficiency analysis.

For this comparison, false-positive and false-negative predictions were partitioned into errors unique to each classifier and errors shared by both classifiers. Morgan r2 and GraphMVP-GIN showed only partially overlapping error patterns across many pathway labels under both random and scaffold splitting (Fig. 7). Substantial numbers of false-positive and false-negative errors were unique to each classifier, indicating that GraphMVP-GIN has a different error pattern from Morgan fingerprints.

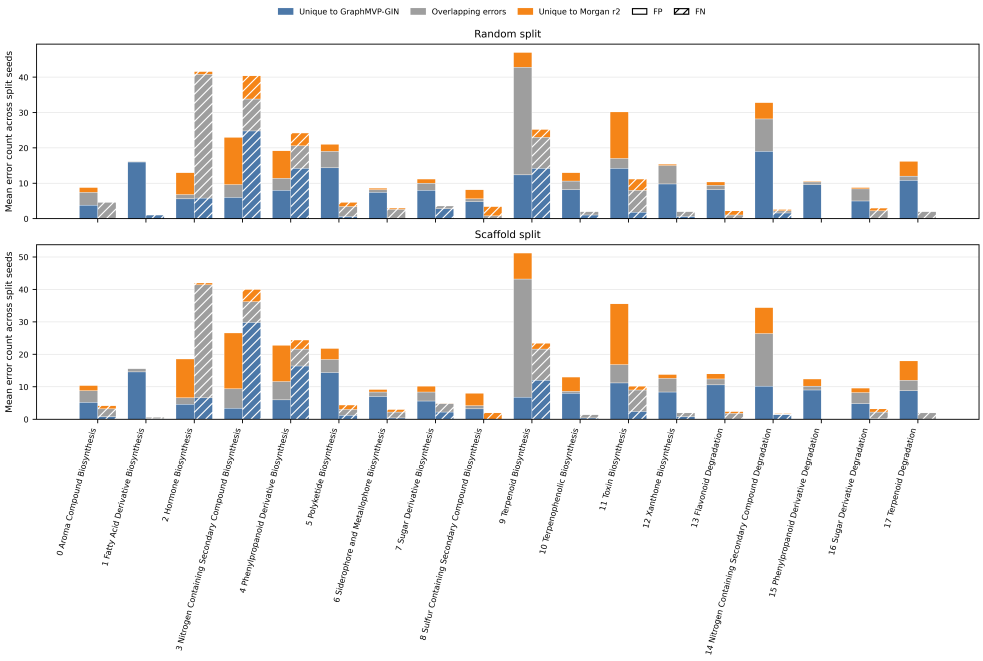


Fig. 7 Morgan r2 and GraphMVP-GIN show only partially overlapping false PMN pathway predictions. For each pathway, the paired bars represent false-positive (FP) and false-negative (FN) predictions, with hatched bars denoting FN errors. Each bar is partitioned into errors unique to GraphMVP-GIN, errors shared by both models, and errors unique to Morgan r2. Error counts were averaged across independently generated splits of the PMN development set and evaluated on the external test set.

We next tested whether these complementary error patterns translated into improved ensemble performance. Pairwise ensembles were generated by equally averaging the predicted probabilities from Morgan r2 and each foundation-model classifier. An additional ensemble applied the same probability-averaging procedure across all individual fingerprint- and foundation-model-based classifiers (Fig. 8).

Several pairwise ensembles achieved higher mean macro-AUROC than Morgan r2 alone in the PMN task under both random and scaffold splits. The largest increase was observed for the Morgan r2–GraphMVP-GIN combination, consistent with the partially distinct error patterns of the two classifiers. Relative to Morgan r2 alone, the mean paired improvement was 0.0149 in macro-AUROC under random partitioning of the PMN development set (95% CI: 0.0067 to 0.0232) and 0.0142 under scaffold-based partitioning (95% CI: 0.0068 to 0.0216). Other pairwise combinations produced smaller gains. These results indicate that selected foundation-model predictors provided complementary information to Morgan fingerprints, but that this complementarity was not shared uniformly across all representations.

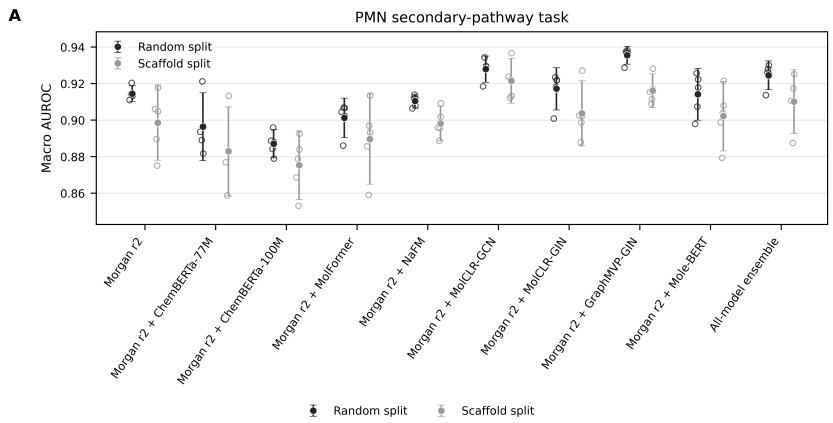


Fig. 8 Performance of ensembles combining Morgan fingerprints with foundation-model predictions for PMN secondary-pathway prediction. Open symbols denote individual split seeds, whereas filled symbols and error bars denote the mean and 95% confidence interval across split seeds.

4 Discussion

This benchmark shows that molecular foundation-model representations contain biosynthesis-related predictive information, but their utility depends strongly on the downstream task. Morgan fingerprints remained strong full-data baselines, whereas selected foundation-model embeddings provided advantages in low-sample settings. In addition, some foundation-model predictors that were weaker as standalone models improved when ensembled with Morgan fingerprints, indicating that representation utility cannot be assessed solely from individual model rankings.

4.1 Molecular pretraining benefits tasks related to global structural characteristics and low-data settings

The differences in performance across tasks suggests that the benefit of pretrained FMs depends on how well their pretrained representations align with the downstream target. Pretrained embeddings were relatively more effective for tasks associated with global structure such as natural-product-likeness, whereas molecular fingerprints showed greater advantages for pathway- and reaction-level predictions.

The relative strength of pretrained embeddings on NP-likeness and broad NP class prediction is consistent with both tasks being strongly associated with global molecular structure and core molecular scaffolds [24, 25]. Recent analyses of chemical language models found that tasks emphasizing global molecular features can benefit from representations that integrate information across the molecule [26]. This may be a reason why pretrained embeddings were relatively effective for these tasks, particularly when labeled data were limited.

Morgan fingerprints achieved the strongest full-data performance in both the plant-focused PMN task and the broader KEGG task, which is not specific to natural products. Many pathway-associated compounds differ in specific structural motifs and

local chemical modifications, including those introduced by tailoring reactions such as oxidation, methylation, glycosylation, and acylation [27]. Because Morgan fingerprints explicitly retain local atomic environments, they may be particularly effective at capturing these structural differences [22]. One exception was the lowest-data regime of the PMN task, where GraphMVP-GIN outperformed Morgan fingerprints. Its 3D-informed pretraining may retain geometric information beyond that represented by local 2D substructures [17]. Such information may be particularly relevant to natural products, which are often rich in stereochemistry and three-dimensional structural complexity [28, 29].

Reaction-neighbor prediction places an even greater emphasis on local and relational information, and Morgan fingerprints outperformed the foundation-model representations. Determining whether two metabolites are directly connected requires identifying structural differences caused by a biochemical transformation rather than recognizing their overall molecular characteristics. Compounds can share the same scaffold or pathway but differ in whether they can be directly connected by a plausible biochemical transformation [30]. This was reflected in the substantially greater difficulty of same-pathway hard negatives in the PlantCyc task. Although the pair representations used here captured shared and differing structural information, they did not identify atom correspondences or localize reaction centers. This may explain why prediction remained difficult for hard examples even with Morgan fingerprints. Reaction-aware representation learning provides one possible direction for addressing this limitation. For example, RXGL uses known reactions to construct a reaction-aware molecular graph and learns latent relations between reactant and product representations, allowing reaction-derived transformation information to be captured [31].

4.2 Architecture and pretraining strategy influence the transferability of pretrained representations

The evaluated molecular FMs can be broadly divided into two classes: transformer based masked language models and graph neural network models. Transformer models such as ChemBERTa and MolFormer process molecules as token sequences and use self-attention to integrate information across the full SMILES representation [14, 15], whereas graph models such as MolCLR, GraphMVP, and Mole-BERT propagate information through neighboring atoms [16, 17]. These different inductive biases may partly explain the task-dependent performance observed here. ChemBERTa-77M and MolFormer were among the strongest foundation-model representations for the broader molecule-level tasks, where global scaffold organization and long-range structural context are informative. In contrast, graph-based models performed competitively or best in several PMN settings, where local topology and pathway-specific structural modifications may be more important. The pretraining objectives may further contribute to these differences. Masked-token prediction in ChemBERTa and MolFormer requires integrating information across the molecular sequence, which may favor broader molecule-level representations [14, 15]. In contrast, MolCLR learns from graph perturbations that alter atoms, bonds, and subgraphs, potentially emphasizing molecular

topology and local structural patterns [16], while GraphMVP additionally incorporates information from three-dimensional molecular geometry [17]. Differences in architecture and pretraining objectives provide plausible explanations for the observed performance patterns, though other factors such as the scale and scope of pretraining data may also play a role.

Domain-specific pretraining did not necessarily produce stronger frozen representations. For NP-likeness and NP class prediction, NaFM did not outperform the strongest foundation models, even though those models were not specifically pretrained on natural products. NaFM uses scaffold-aware contrastive learning and masked graph reconstruction, but these objectives do not necessarily make downstream-relevant information directly accessible from fixed embeddings [19]. In addition, NaFM was originally evaluated with end-to-end fine-tuning, which differs from our evaluation using fixed embeddings. Thus, natural-product-specific pretraining may still be useful for task-specific adaptation, but does not necessarily yield stronger task-ready frozen representations.

Scaffold splitting provided an additional view of representation transfer. Such splits are intended to evaluate prediction on molecules whose Bemis–Murcko frameworks were not observed during training and are generally more challenging than random partitions [32, 33]. Previous molecular-property benchmarks have similarly shown that pretrained molecular models can perform strongly under scaffold-based evaluation [15, 34]. Consistent with these findings, selected foundation-model representations, particularly MolFormer in the natural-product-likeness task, retained strong performance under scaffold splitting and showed smaller decreases than Morgan fingerprints. This is consistent with pretrained models capturing structural features that remain informative beyond the exact scaffolds encountered during downstream training. The strength of this effect nevertheless depends on the task, particularly when class or pathway labels are themselves closely associated with specific molecular scaffolds.

4.3 Foundation-model embeddings provide complementary predictive information

Performance of individual representations did not fully reflect the complementary predictive information they contained. In PMN pathway prediction, GraphMVP-GIN was weaker than Morgan fingerprints as a standalone model, but the two models showed partially different error patterns and their probability-level ensemble improved over Morgan alone. The ensemble improvement provides evidence that GraphMVP-GIN retained predictive information that was not fully represented by the fingerprint model, even though this information was insufficient to make it the stronger individual predictor.

The pathway-level performance provide a concrete example of how this complementarity may arise. Morgan fingerprints were markedly stronger for terpenophenolic biosynthesis, whereas several graph-based embeddings performed better for terpenoid degradation under both random and scaffold splits. This pattern suggests that the relative value of each representation may depend on the structural characteristics associated with individual pathway labels.

This complementarity may arise because the representations organize molecular structure differently. Morgan fingerprints explicitly encode local atom environments, whereas pretrained embeddings compress broader structural contexts learned through their pretraining objectives. Similar benefits have been reported for hybrid molecular-property models that combine fingerprints with graph-derived representations, with ablation analyses showing that both components contribute predictive information [35]. Our results extend this observation to pathway association and show that a weaker standalone embedding can still improve fingerprint-based prediction when the two models capture partially nonredundant information.

Future work could extend this complementarity beyond multiple molecular representations by incorporating biochemical context. Reaction-aware models such as BioNavi-NP have shown that learned molecular transformations can support multi-step natural-product pathway reconstruction [5], while ESP demonstrated that joint representations of enzyme sequences and substrate structures can predict enzyme–substrate compatibility across diverse enzyme families [36]. Combining metabolite representations, candidate transformations, and enzyme information could therefore improve pathway reconstruction by prioritizing connections that are both chemically plausible and supported by compatible enzymes.

5 Conclusion

Our study shows that molecular foundation model representations provide advantages over Morgan fingerprints, particularly for tasks associated with global structural characteristics and in low-data settings. For NP-likeness and broad NP class prediction, masked-language models such as ChemBERTa and MolFormer were especially competitive when few labeled examples were available. For plant pathway prediction, graph-based models such as GraphMVP-GIN performed particularly well in the low-data regime. In contrast, Morgan fingerprints were often stronger when larger labeled datasets were available. They also performed better on tasks requiring finer local structural discrimination, such as pathway association and reaction-neighbor prediction.

Foundation-model embeddings also contributed predictive information that was not fully captured by fingerprints, as shown by improved pathway prediction from selected probability-level ensembles. Their value should therefore not be judged only by standalone rankings. At the same time, the reduced performance on hard reaction-neighbor pairs shows that representations of individual molecular structures are insufficient for some biosynthetic inference tasks. Future models that integrate molecular structure with explicit transformations, reaction centers, and enzyme information may provide a stronger basis for metabolic pathway reconstruction. Overall, molecular foundation models are most useful as task-dependent and potentially complementary representations, while Morgan fingerprints remain an essential baseline for natural-product prediction.

Declarations

Funding

This research received no external funding

Acknowledgements

Not applicable.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The original datasets analyzed in this study are available from the sources cited in the manuscript. Processed benchmark datasets and data split assignments are available at https://github.com/linf545/pathway-fm-benchmark. The code for dataset processing, molecular representation extraction, model training, evaluation, and figure generation is available in the same repository.

Code availability

Code used for dataset processing, representation extraction, model training, evaluation, and figure generation is available at https://github.com/linf545/pathway-fm-benchmark.

Author contributions

F.L., D.W., and Y.H. wrote the manuscript. F.L. designed the benchmark, curated and processed the data, and implemented the analyses. Y.H., D.W., and S.X. provided critical feedback. All authors read and approved the final manuscript.

Appendix A Supplementary methods and results

A.1 Sample efficiency for PlantCyc reaction-neighbor prediction

We additionally evaluated sample efficiency for PlantCyc reaction-neighbor prediction using the mixed-negative variant, which combines non-neighbor pairs with and without

a shared pathway. Class-balanced training subsets ranging from 10 to 3,000 labeled pairs were sampled while the corresponding validation and test sets were kept fixed.

Performance generally increased with the number of labeled training pairs under both evaluation schemes (Fig. A1). Morgan fingerprints achieved the highest mean AUROC across most or all evaluated training levels under both random and scaffold-pair splitting. The frozen foundation-model representations generally remained below the Morgan baselines, although the differences among several representations were modest at the smallest training sizes. Overall, the sample-efficiency analysis did not alter the full-data conclusion that direct reaction-neighbor prediction is challenging and that the evaluated frozen foundation-model representations did not consistently outperform Morgan fingerprints for this task.

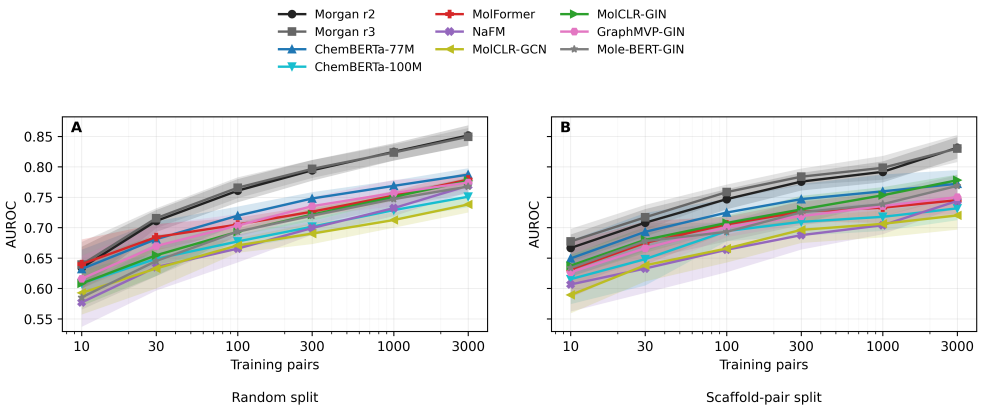


Fig. A1 Sample efficiency for PlantCyc reaction-neighbor prediction using mixed negatives. AUROC is shown as a function of the total number of labeled training pairs under random and scaffold-pair splitting. Lines represent mean AUROC and shaded regions represent 95% confidence intervals across split seeds.

A.2 LoRA adaptation

LoRA adaptation was evaluated for ChemBERTa-77M and MolFormer on the natural-product-likeness and natural-product chemical-class tasks. The adapted models were compared with their corresponding frozen representations and with Morgan r2 using the same random and scaffold partitions.

For natural-product-likeness prediction, both adapted models consistently outperformed their frozen representations (Fig. A2A–B). The gains were observed under both split schemes, although the relative ranking of ChemBERTa-77M and MolFormer became similar after adaptation. In contrast, the chemical-class results were mixed (Fig. A2C–D). LoRA modestly improved MolFormer but slightly reduced the performance of ChemBERTa-77M, and neither adapted model surpassed Morgan r2. These results show that the benefit of parameter-efficient adaptation depends on both the downstream task and the pretrained model.

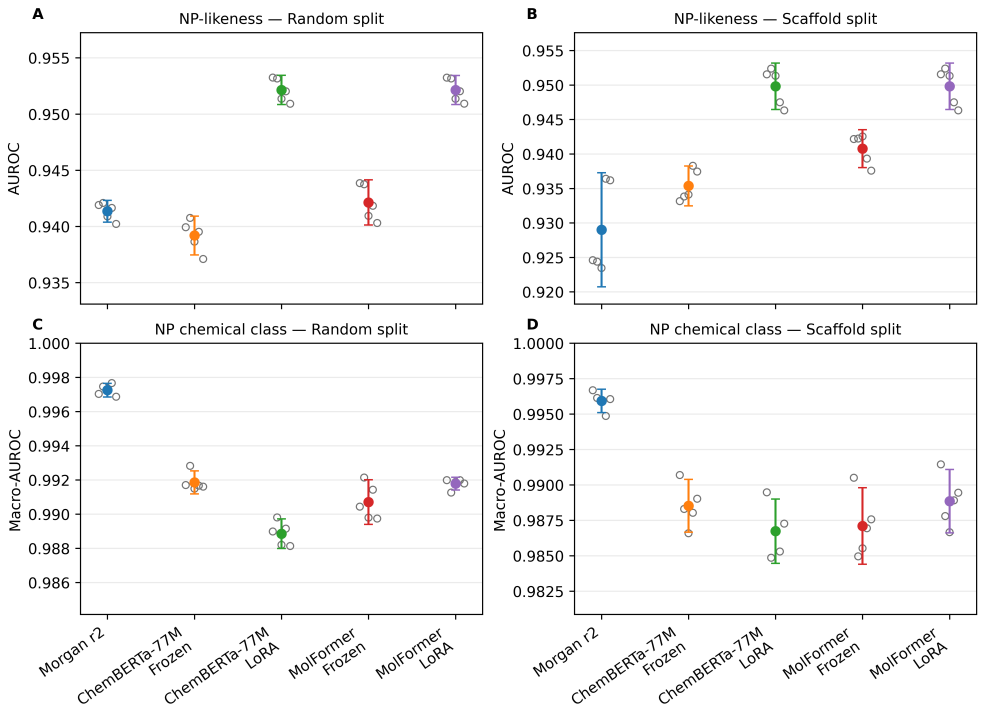


Fig. A2 Comparison of frozen foundation-model representations and task-specific adaptation across two natural-product prediction tasks. Frozen representations and LoRA-adapted models were compared with Morgan r2 for natural-product-likeness prediction under random and scaffold splitting (A–B) and natural-product chemical-class prediction under random and scaffold splitting (C–D). Open symbols represent individual split seeds, whereas filled symbols and error bars represent the mean and 95% confidence interval across five split seeds.

References

[1] Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. Journal of Natural Products. 2020 03;83(3):770–803. https://doi.org/10.1021/acs.jnatprod.9b01285. https://pubs.acs.org/jnprdf/article-pdf/83/3/770/14572523/np9b01285.pdf.

[2] Atanasov AG, Zotchev SB, Dirsch VM, Orhan IE, Banach M, Rollinger JM, et al. Natural products in drug discovery: advances and opportunities. Nature Reviews Drug Discovery. 2021 mar;20(3):200–216. https://doi.org/10.1038/s41573-020-00114-z.

[3] Paddon CJ, Keasling JD. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. Nature Reviews Microbiology. 2014 may;12(5):355–367. https://doi.org/10.1038/nrmicro3240.

[4] Seshadri K, Abad AND, Nagasawa KK, Yost KM, Johnson CW, Dror MJ, et al. Synthetic Biology in Natural Product Biosynthesis. Chemical Reviews. 2025 03;125(7):3814–3931. https://doi.org/10.1021/acs.chemrev.4c00567. https://pubs.acs.org/chreay/article-pdf/125/7/3814/41209299/cr4c00567.pdf.

[5] Zheng S, Zeng T, Li C, Chen B, Coley CW, Yang Y, et al. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. Nature Communications. 2022 jun;13(1):3342. https://doi.org/10.1038/s41467-022-30970-9.

[6] Medema MH, Osbourn A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. Natural Product Reports. 2016 aug;33(8):951–962. https://doi.org/10.1039/c6np00035e.

[7] Qin K, Liu F, Zhang C, Deng R, Fernie AR, Zhang Y. Systems and synthetic biology for plant natural product pathway elucidation. Cell Reports. 2025;44(6):115715. https://doi.org/https://doi.org/10.1016/j.celrep.2025.115715.

[8] Lin F, Manisseri C, Fagerström A, Peck ML, Vega-Sánchez ME, Williams B, et al. Cell Wall Composition and Candidate Biosynthesis Gene Expression During Rice Development. Plant and Cell Physiology. 2016 oct;57(10):2058–2075. https://doi.org/10.1093/pcp/pcw125.

[9] Gao J, Gou Y, Huang L, Lian J. Reconstitution and optimization of complex plant natural product biosynthetic pathways in microbial expression systems. Current Opinion in Biotechnology. 2024;87:103136. https://doi.org/https://doi.org/10.1016/j.copbio.2024.103136.

[10] Nützmann HW, Huang A, Osbourn A. Plant metabolic clusters – from genetics to genomics. New Phytologist. 2016;211(3):771–789. https://doi.org/https://doi.org/10.1111/nph.13981. https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.13981.

[11] Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, et al. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. Journal of Natural Products. 2021 10;84(11):2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399. https://pubs.acs.org/jnprdf/article-pdf/84/11/2795/18215233/np1c00399.pdf.

[12] Bao H, Zhao J, Zhao X, Zhao C, Lu X, Xu G. Prediction of plant secondary metabolic pathways using deep transfer learning. BMC Bioinformatics. 2023 sep;24(1):348. https://doi.org/10.1186/s12859-023-05485-9.

[13] Shen X, Wang R, Xiong X, Yin Y, Cai Y, Ma Z, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nature Communications. 2019 apr;10(1):1516. https://doi.org/10.1038/s41467-019-09550-x.

[14] Ahmad W, Simon E, Chithrananda S, Grand G, Ramsundar B.: ChemBERTa-2: Towards Chemical Foundation Models. Available from: https://arxiv.org/abs/2209.01712.

[15] Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence. 2022 dec;4(12):1256–1264. https://doi.org/10.1038/s42256-022-00580-7.

[16] Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence. 2022 mar;4(3):279–287. https://doi.org/10.1038/s42256-022-00447-x.

[17] Liu S, Wang H, Liu W, Lasenby J, Guo H, Tang J. Pre-training Molecular Graph Representation with 3D Geometry. In: International Conference on Learning Representations; 2022. Available from: https://openreview.net/forum?id=xQUe1pOKPam.

[18] Xia J, Zhao C, Hu B, Gao Z, Tan C, Liu Y, et al. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. In: The Eleventh International Conference on Learning Representations; 2023. Available from: https://openreview.net/forum?id=jevY-DtiZTR.

[19] Ding Y, Qiang B, Li S, Zhou Y, Yu J, Li Q, et al. Pretraining a foundation model for small-molecule natural products. Nature Machine Intelligence. 2026 may;8(5):777–788. https://doi.org/10.1038/s42256-026-01226-8.

[20] Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. Journal of Medicinal Chemistry. 1996 07;39(15):2887–2893. https://doi.org/10.1021/jm9602928. https://pubs.acs.org/jmcmar/article-pdf/39/15/2887/455949/jm9602928.pdf.

[21] Hawkins C, Xue B, Yasmin F, Wyatt G, Zerbe P, Rhee S. Plant Metabolic Network 16: expansion of underrepresented plant groups and experimentally supported enzyme data. Nucleic Acids Research. 2025 01;53(D1):D1606–D1613. https://doi.org/10.1093/nar/gkae991. https://academic.oup.com/nar/article-pdf/53/D1/D1606/60949562/gkae991.pdf.

[22] Rogers D, Hahn M. Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling. 2010 04;50(5):742–754. https://doi.org/10.1021/ci100050t. https://pubs.acs.org/jcisd8/article-pdf/50/5/742/11822562/ci100050t.pdf.

[23] Menke J, Massa J, Koch O. Natural product scores and fingerprints extracted from artificial neural networks. Computational and Structural Biotechnology Journal. 2021;19:4593–4602. https://doi.org/https://doi.org/10.1016/j.csbj.2021.07.032.

[24] Ertl P, Roggo S, Schuffenhauer A. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. Journal of Chemical Information and Modeling. 2008 11;48(1):68–74. https://doi.org/10.1021/ci700286x. https://pubs.acs.org/jcisd8/article-pdf/48/1/68/11545023/ci700286x.pdf.

[25] Ertl P, Schuhmann T. A Systematic Cheminformatics Analysis of Functional Groups Occurring in Natural Products. Journal of Natural Products. 2019 04;82(5):1258–1263. https://doi.org/10.1021/acs.jnatprod.8b01022. https://pubs.acs.org/jnprdf/article-pdf/82/5/1258/8364934/np8b01022.pdf.

[26] Fender I, Gut JA, Lemmin T. Beyond performance: how design choices shape chemical language models. Journal of Cheminformatics. 2025 nov;17(1):173. https://doi.org/10.1186/s13321-025-01099-w.

[27] Walsh CT. Tailoring enzyme strategies and functional groups in biosynthetic pathways. Natural Product Reports. 2023 02;40(2):326–386. https://doi.org/10.1039/d2np00048b. https://pubs.rsc.org/np/article-pdf/40/2/326/8608460/d2np00048b.pdf.

[28] Stratton CF, Newman DJ, Tan DS. Cheminformatic comparison of approved drugs from natural product versus synthetic origins. Bioorganic & Medicinal Chemistry Letters. 2015 nov;25(21):4802–4807. https://doi.org/10.1016/j.bmcl.2015.07.014.

[29] Bitchagno GTM, Nchiozem-Ngnitedem VA, Melchert D, Fobofou SA. Demystifying racemic natural products in the homochiral world. Nature Reviews Chemistry. 2022 nov;6(11):806–822. https://doi.org/10.1038/s41570-022-00431-4.

[30] Mu F, Unkefer CJ, Unkefer PJ, Hlavacek WS. Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. Bioinformatics. 2011 jun;27(11):1537–1545. https://doi.org/10.1093/bioinformatics/btr177.

[31] Li A, Casiraghi E, Rousu J. Chemical reaction enhanced graph learning for molecule representation. Bioinformatics. 2024 10;40(10):btae558. https://doi.org/10.1093/bioinformatics/btae558. https://academic.oup.com/bioinformatics/article-pdf/40/10/btae558/59716457/btae558.pdf.

[32] Wu Z, Ramsundar B, Feinberg E, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chemical Science. 2018 01;9(2):513–530. https://doi.org/10.1039/c7sc02664a. https://pubs.rsc.org/sc/article-pdf/9/2/513/5768099/c7sc02664a.pdf.

[33] Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nature Communications. 2023 oct;14(1):6395. https://doi.org/10.1038/s41467-023-41948-6.

[34] Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for Pre-training Graph Neural Networks. In: International Conference on Learning Representations; 2020. .

[35] Cai H, Zhang H, Zhao D, Wu J, Wang L. FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction. Briefings in Bioinformatics. 2022 11;23(6):bbac408. https://doi.org/10.1093/bib/bbac408. https://academic.oup.com/bib/article-pdf/23/6/bbac408/47144410/bbac408.pdf.

[36] Kroll A, Ranjan S, Engqvist MKM, Lercher MJ. A general model to predict small molecule substrates of enzymes based on machine and deep learning. Nature Communications. 2023 may;14(1):2787. https://doi.org/10.1038/s41467-023-38347-2.