JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

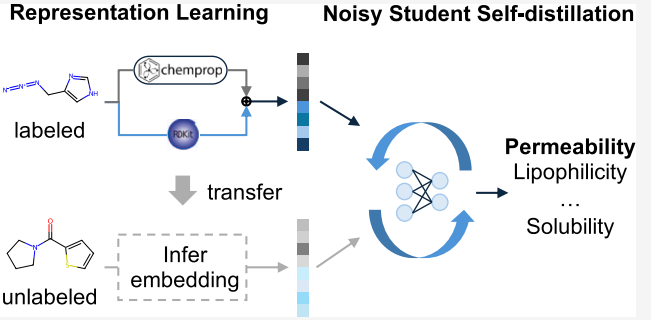
MycoPermeNet-v2: Improved Prediction of Mycomembrane Permeation Using Fusion Noisy Student Self-Distillation

Nelson Evbarunegbe,† Shiyun Wa,† Isha Karn, and Anna G. Green*

Cite This: J. Chem. Inf. Model. 2026, 66, 2985−2996 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine learning (ML) techniques offer a promising path for accelerating antibiotic discovery by computationally predicting and optimizing desirable pharmaceutical properties of molecules. One difficult case for drug discovery is tuberculosis (TB), the disease caused by *Mycobacterium tuberculosis*, a bacterium with intrinsic resistance to many antibiotics. The unique outer membrane of *M. tuberculosis*, called the mycomembrane, is thought to contribute to intrinsic antibiotic resistance by establishing a permeability barrier. While recent ML works have attempted to predict the permeability of compounds through the mycomembrane, the scarcity of labeled data has led to approaches that struggle to generalize. In this work, we propose a robust two-stage model, MycoPermeNet-v2, to predict compound permeability through the mycomembrane of *M. tuberculosis*. MycoPermeNet-v2 fuses molecular descriptors with learned graph-based embeddings and incorporates Noisy Student self-disTillation (NST) to improve performance even when labeled data is limited. Compared to prior work, we achieve significantly higher performance (RMSE from 0.755 ± 0.024 to 0.719 ± 0.021, adjusted $p < 0.0001$). We systematically evaluate the robustness of the overall model across different data split strategies, as well as the contribution of each component in the proposed feature fusion and NST framework. Furthermore, the interpretability analysis shows that the learned representations capture chemically meaningful features relevant to *M. tuberculosis* permeability. We show the generalizability of our approach over multiple models and physical chemistry property prediction tasks, demonstrating strong applicability in data-constrained molecular learning scenarios.



INTRODUCTION

Machine learning (ML) has accelerated antibiotic discovery, a field which had been struggling to find new classes of compounds with antimicrobial activity.[1−3] A uniquely challenging case for antibiotic discovery is *Mycobacterium tuberculosis*, which causes tuberculosis (TB), a disease that is responsible for over a million deaths annually.[4] *M. tuberculosis* is intrinsically resistant to antibiotics that are effective against other bacteria[5] due to its unique outer membrane, called the mycomembrane, which restricts molecular entry into the cell.[6−8]

Initial attempts to predict mycomembrane permeability inferred permeability labels indirectly from other experimental data,[9,10] with limited validation of model robustness. Lepori et al.[11] were the first to train a neural network to predict mycomembrane permeability based on experimentally generated cell permeability data. Their model, called MycoPermeNet, was trained on a data set of 1558 molecules and demonstrated strong predictive performance on within-distribution compounds. While the training data set was the first of its kind, it was composed of relatively small, azide-tagged molecules. The ability of MycoPermeNet to generalize out-of-distribution compounds, a problem particularly important in antibiotic discovery, has not yet been assessed.

The performance of ML models is often constrained by the scarcity of labeled data and limitations in the models' ability to generate useful representations of said data. When data is scarce, supervised learning can benefit from unsupervised pretraining on unlabeled data sets,[12−16] but the pretraining task requires careful design to avoid negative transfer.[17] Semisupervised learning offers a compromise, aimed at improving the model performance by combining limited labeled data with abundant unlabeled data. Noisy Student Training is a semisupervised method that involves predicting labels for unlabeled data, which are then used to train a new version of the model. This method was initially proposed in computer vision[18] and later adapted to molecular[19] and protein[20,21] tasks. A second challenge facing ML for molecular property prediction is the representational power of the graph neural networks (GNNs) typically used to generate molecular

Received: October 7, 2025
Revised: January 20, 2026
Accepted: January 21, 2026
Published: March 2, 2026

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

embeddings. GNNs can struggle to extract global molecular features[2] but show promising improvement by fusing fixed fingerprints to learned neural representations,[22,23] indicating that the two molecular representations may capture different information about the input molecule.

Our work introduces MycoPermeNet-v2, which improves permeability prediction by leveraging chemically informative feature fusion and Noisy Student self-disTillation (NST),[24] a variant of Noisy Student Training, while still correctly recapitulating domain knowledge. We show that feature fusion exposes the model to information not captured by embeddings alone and that models can achieve improved performance by injecting only a few domain-relevant features. We show that NST improves performance using two different unlabeled data sets and is robust to the hardness of the train-validate-test split. Finally, we demonstrate that the effectiveness of feature fusion and NST is statistically significant and generalizable over widely adopted molecular GNNs and multiple molecular property prediction tasks. Overall, our model addresses the limitations of previous work by enabling better generalization to structurally and chemically diverse compounds, thus reducing uncertainty and improving the permeability prediction accuracy for out-of-distribution molecules.

METHODS

Data Set Preparation and Splits

Permeability Data Set Generation. For this work, we use the mycomembrane permeability data set previously generated in Lepori et al.[11] This study measured the permeability of chemical compounds through the mycomembrane using a bio-orthogonal click-chemistry assay called PAC-MAN performed in live *M. tuberculosis* bacteria. PAC-MAN was originally introduced by Liu et al.,[25] and related versions have been used in other bacterial species.[26,27] All compounds in this data set include an azide group, a commonly used compact bio-orthogonal tag,[28,29] which enables the experimental readout.[25]

In Lepori et al.,[11] permeability is measured by fluorescence readout. Reported values are normalized to control for differences in reactivity between different azide-tagged compounds, by computing the residual between the fluorescence readout and the line of best fit predicted by the log-reactivity. Residuals are then standardized based on the mean and standard deviation per assay to control for assay heterogeneity. A subset of compounds was measured in duplicate or triplicate with the mean value across replicates reported. Final reported values are the standardized residuals of the apparent permeability, with values ranging from -3.1 to 1.58, and a mean of -1.82×10^{-17}.

Permeability Data Set and Benchmark Data Sets. Our permeability data set, generated by prior work described above, consists of 1558 labeled chemical compounds after using RDKit software[30] to filter for valid Simplified Molecular Input Line Entry System (SMILES) strings.

To evaluate the generalizability of our method over molecular property prediction tasks, we also employed three physical chemistry data sets (ESOL, FreeSolv, and Lipo (Table S1)) from MoleculeNet benchmark.[31]

Unlabeled Data Sets. We use two external unlabeled data sets for semisupervised training: the Molecular Library Small Molecule Repository (MLSMR)[32] and an in-house enamine-azide data set derived from the Enamine chemical library.[33] The MLSMR data set contains approximately 215,000 chemically diverse small molecules and is commonly used for cheminformatics analyses. The in-house enamine-azide data set consists of about 1700 compounds, where azide-functionalized tags are appended to molecular scaffolds in a manner that mimics the chemical patterns seen in the labeled train set, while spanning a broader molecular weight range.

Data Splitting. Many molecular property prediction models achieve strong performance because they split data randomly in terms of compounds,[2,34,35] rather than avoiding potential data leakage by splitting in terms of molecular scaffold. We experiment with three data split strategies, which we term scaffold-exclusive, scaffold-agnostic, and scaffold-inclusive. This study uses the scaffold-exclusive strategy to train models, which involves splitting the test set from the train set based on molecular scaffolds.[36] Specifically, we adopt the implementation of Heid et al.,[37] which puts all the scaffold groups larger than half the test size into the train set, then randomly uses the remaining scaffold groups to fill the train, val, and test sets. As a result, the test set contains structurally distinct molecules from the train set, encouraging the model to generalize better to predict permeability in an unseen chemical space. The test set is held out for both stages of model training: graph representation learning and downstream prediction. Scaffold-agnostic splitting is a commonly used strategy,[2,34,35] where data are randomly partitioned into train and test sets. Although the test set is still held out from the first-stage GNN pretraining, scaffold overlap exists across subsets, leading to possible data leakage. Finally, a scaffold-inclusive strategy trains GNN on the entire data set in the first stage, while using a scaffold-based train-test split for the second stage. This can maximize the use of limited labeled data and improve the model performance.[11] Overall, we hold out the test set for all the splitting methods and split the train and validation sets within the remaining subset in each independent run at runtime. Table 1 displays the number of molecules and scaffolds in the holdout test set and the remaining train-validation set. Figure 1 shows the detailed data preprocessing flowchart.

Table 1. Dataset Statistics for Train-Validation and Test Splits

	train-validation		test	
data set	#Mol.	#Scaffold	#Mol.	#Scaffold
FreeSolv	512	19	127	44
ESOL	902	163	225	106
permeability	1422	173	136	44
Lipo	3360	1930	840	514

MycoPermeNet-v2 Model

Model Architecture. MycoPermeNet-v2 is a two-stage prediction model that generates neural representations of compounds and then predicts compound permeability (Figure 2). We use a PyTorch Geometric[38] implementation of Chemprop that takes SMILES strings as the input (Table S2) and trains a directed message-passing neural network (D-MPNN) with the same architecture as Chemprop.[37] After pretraining, Chemprop's encoder is used to generate embedding $\mathbf{h}_i \in \mathbb{R}^d$ for each labeled compound and embedding $\tilde{\mathbf{h}}_j \in \mathbb{R}^d$ for each unlabeled compound. Both types of embeddings are concatenated with their corresponding normalized descriptors computed by the RDKit, forming the labeled and unlabeled feature vectors as $\mathbf{x}_i = [\mathbf{h}_i; \mathbf{r}_i] \in \mathbb{R}^{d+k}$ and $\tilde{\mathbf{x}}_j = [\tilde{\mathbf{h}}_j; \tilde{\mathbf{r}}_j] \in \mathbb{R}^{d+k}$, respectively. Among the full RDKit descriptor library, some descriptors evaluate to zero for all molecules in a given molecular property data set because the corresponding chemical substructures do not appear in the molecules. We therefore retrain only the chemically informative, nonzero RDKit descriptors. This filtering step is applied across all data sets and all descriptor-related models used in this study. There are 179, 188, 191, and 184 nonzero descriptors for FreeSolv, Lipophilicity, ESOL, and our permeability data sets, respectively

The second phase is the prediction phase, where the fused features are passed into a four-layer multilayer perceptron (MLP) as input for prediction.

Model Training. The Chemprop encoder and the MLP are trained independently by minimizing the mean square error (MSE) between the ground truth and the predicted permeability. Particularly, we train the MLP using the adapted Noisy Student self-disTillation (NST) in a semisupervised fashion following Algorithm 1. Our NST

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

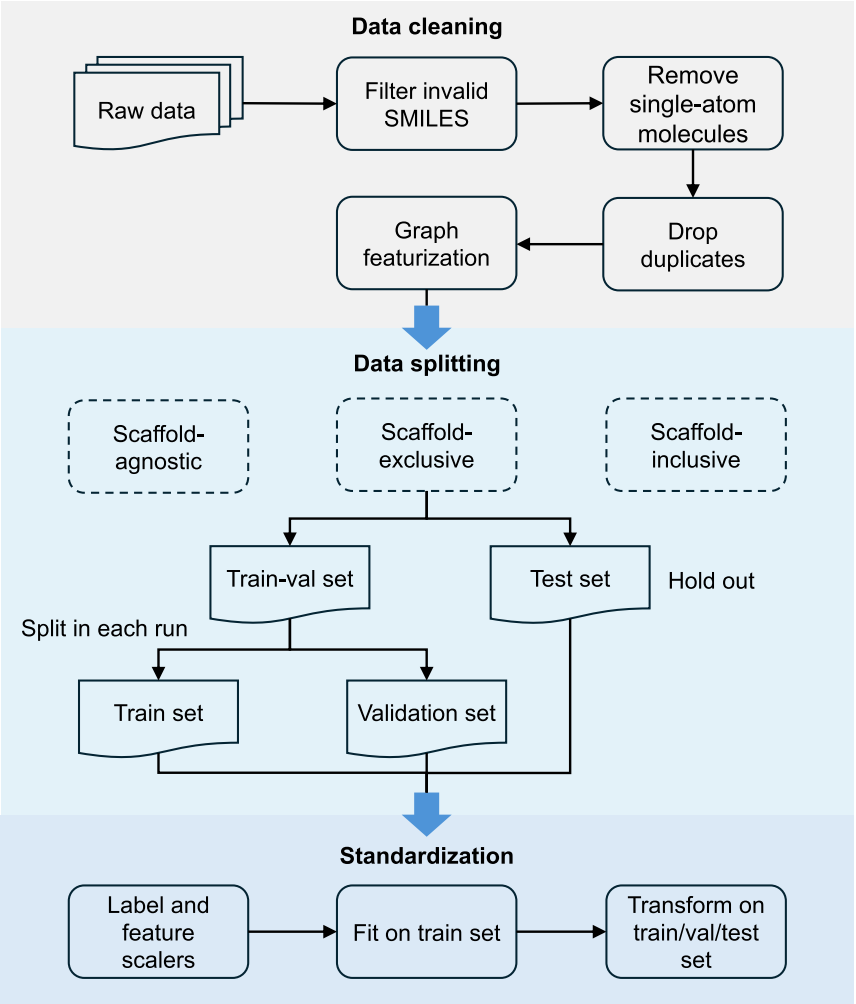


Figure 1. Overview of the data set preprocessing workflow. (1) Data cleaning: we clean all molecular property data sets by using RDKit to filter molecules with invalid SMILES strings, removing single-atom data (as some GNNs require bond information), dropping entries with duplicated SMILES strings, and featurizing graphs with atom and bond features. (2) Data splitting: the scaffold-exclusive method is used to split the hold-out test set and split the train and validation sets in each independent run. (3) Standardization: we fit feature and label standardization scalers on the train set and apply them to transform all the subsets. The label transformations are inverted before computing evaluation metrics.

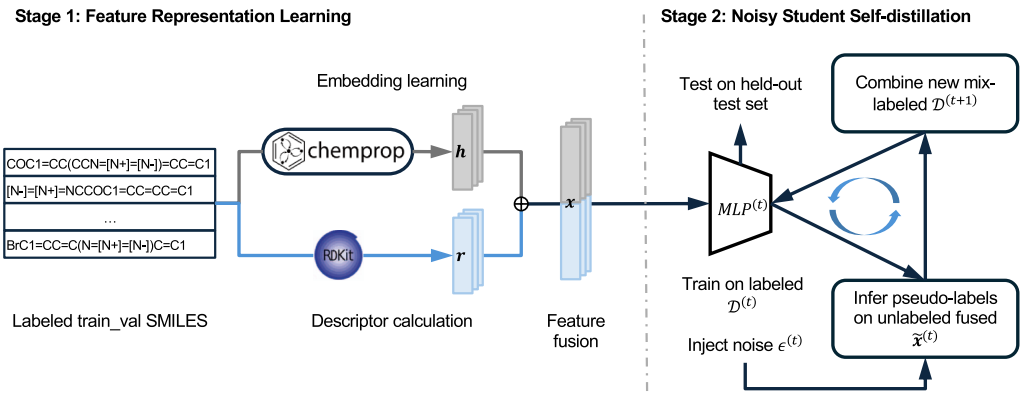


Figure 2. The workflow of the MycoPermeNet-v2 Model.

models are evaluated on the validation set at each iteration, and the final performance is the test result of the best validation checkpoint.

Following the practices reported in refs 18 and 20 we experimented with different NST procedures, including using traditional teacher-student distillation, adding Dropout to the student model, and injecting Gaussian noise into the predicted pseudolabels. Because pseudolabel noising had the highest test performance, we report this method throughout (Table S8). In addition, we treated the number of iterations and the volume of unlabeled data per iteration as hyperparameters of NST. Hyperparameters were selected based on best validation performance (detailed procedure in the Training Framework and Hyperparameters section). Our procedure selected three NST iterations with 500 unlabeled samples per iteration for the

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

primary permeability prediction task (Table S5), and the same selection strategy was applied to all other benchmark data sets.

To ensure generalization, we follow the scaffold-exclusive data split strategy described in the section Data Set Preparation and Splits, which prevents any scaffold overlap between the training and held-out test sets. Detailed model architectures, hyperparameters, and other Tables S3 and S4.

Algorithm 1 Noisy Student self-disTillation for MLP

Require: Labeled data $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, unlabeled dataset $\mathcal{U}$, unlabeled sample size m, number of iterations T, noise variance σ^2

Train the initial teacher $f_{\text{MLP}}^{(0)}$ on $\mathcal{L}$, set $\mathcal{D}^{(0)} \leftarrow \mathcal{L}$

for $t = 0$ **to** $T - 1$ **do**

  Sample unlabeled $\{\tilde{\mathbf{x}}_j^{(t)}\}_{j=1}^{m} \subset \mathcal{U}$

  Infer pseudo-labels for the unlabeled samples and inject noise:

$$\tilde{y}^{(t)} \leftarrow f_{\text{MLP}}^{(t)}(\tilde{\mathbf{x}}^{(t)}) + \epsilon^{(t)},\ \epsilon^{(t)} \sim \mathcal{N}(0, \sigma^2)$$

  Combine "mix-labeled" data for the next training round:

$$\mathcal{D}^{(t+1)} \leftarrow \mathcal{D}^{(t)} \cup \{(\tilde{\mathbf{x}}_j^{(t)}, \tilde{y}_j^{(t)})\}_{j=1}^{m}$$

  Use teacher model to initialize student model:

$$f_{\text{MLP}}^{(t+1)} \leftarrow f_{\text{MLP}}^{(t)}$$

  Train new teacher $f_{\text{MLP}}^{(t+1)}$ on $\mathcal{D}^{(t+1)}$ (self-distillation)

end for

Baselines and Variants

Two-Stage Baseline. We use the previously published MycoPermeNet[11] as the two-stage baseline model, which uses Chemprop to generate graph embeddings in the first stage, and then uses an MLP to predict compound permeability in the second stage. We attempt to improve the baseline model by adding RDKit descriptors to expand the feature space or by incorporating unlabeled data to increase the input data.

Feature Space Expansion. To enhance the representation power of the two-stage baseline model, after GNN pretraining, we concatenate each compound's RDKit descriptors with its graph embeddings to incorporate both local structural information and global molecular properties. The fused feature vector serves as the input to the second-stage MLP predictor, constituting the Fusion variant model.

Furthermore, to guide the encoder toward learning graph representations aligned with global descriptors, we model their dependencies during graph representation learning using mutual information neural estimator (MINE).[39] MINE uses a fully connected network $\mathcal{T}$ to estimate the mutual information I between the learned graph embedding $\mathbf{h}$ and the RDKit descriptor $\mathbf{r}$. It is then trained by maximizing a lower bound of $I(\mathbf{h}; \mathbf{r})$, where the optimization objective uses the Donsker-Varadhan representation

$$\arg\max_{\mathcal{T}} I_{\mathcal{T}}^{\text{DV}}(\mathbf{h}, \mathbf{r}) = \arg\max_{\mathcal{T}} \mathbb{E}_{(\mathbf{h},\mathbf{r})\sim\mathbb{P}_{\mathbb{HR}}}[\mathcal{T}(\mathbf{h}, \mathbf{r})] - \log(\mathbb{E}_{(\mathbf{h},\bar{\mathbf{r}})\sim\mathbb{P}_{\mathbb{H}}\otimes\mathbb{P}_{\mathbb{R}}}[e^{\mathcal{T}(\mathbf{h},\bar{\mathbf{r}})}]) \quad (1)$$

where $\mathbb{P}_{\mathbb{HR}}$ is the joint distribution and $\mathbb{P}_{\mathbb{H}} \otimes \mathbb{P}_{\mathbb{R}}$ is the product of marginals. In practice, we approximate samples from $\mathbb{P}_{\mathbb{H}} \otimes \mathbb{P}_{\mathbb{R}}$ by shuffling samples from the joint distribution along the batch axis.[39] The negative value of estimated mutual information works as an auxiliary loss $\mathcal{L}^{\text{MINE}}$, which is weighted by a hyperparameter α and added to the original MSE loss $\mathcal{L}^{\text{MSE}}$ to perform MINE-guided GNN learning, as shown in eq 2.

$$\mathcal{L}^{\text{total}} = \mathcal{L}^{\text{MSE}} + \alpha \cdot \mathcal{L}^{\text{MINE}} \quad (2)$$

Incorporating Unlabeled Data. We also increase the effective input data via semisupervised and unsupervised training techniques. Specifically, we adopt NST (Algorithm 1) to train the second-stage MLP, which iteratively introduces unlabeled data and trains the student model using noisy pseudolabeled samples to improve generalization. Additionally, we develop an unsupervised GNN pertaining strategy by optimizing only $\mathcal{L}^{\text{MINE}}$, making mutual information the sole training metric. After being pretrained on large-scale unlabeled data, the encoder weights are transferred to the supervised two-stage baseline framework.

GNN Encoder Baselines. In addition to MycoPermeNet, which uses Chemprop as the GNN encoder, we employ three other GNN encoder baselines, GCN,[40] GINE,[15] and Attentive FP,[41] to further evaluate the feature fusion and NST's generalizability over different graph embeddings.

Descriptor-Based ML Baselines. To evaluate the benefits of GNN embeddings, we included two strong machine-learning baselines, XGBoost and Random Forest. Both models are trained on nonzero RDKit descriptors concatenated with 1024 bit extended-connectivity fingerprints (ECFPs)[42] of a radius 3, forming a baseline commonly used in classical molecular property prediction. Hyperparameters and implementation details are provided in the Training Framework and Hyperparameters section.

Randomness-Aware Repetition Strategy

Random seeds in model initialization or data shuffling can negatively affect the stability of models trained on limited labeled data, resulting in large performance variances across training runs.[43] Therefore, we investigate the effects of two types of randomness influenced by random seed selection: data randomness, which controls splitting and dataloader shuffling, and PyTorch randomness, which controls model initialization and training steps taken. Our results (Tables S6 and Figure S2) suggest that data randomness is the critical factor that harms the stability of the final performance.

To mitigate the impacts of randomness on final performance, we randomly sampled 7 seeds for Chemprop's data randomness and another 7 independent seeds for PyTorch randomness, resulting in 49 unique seed pairs for Chemprop representation learning. For each repetition, a unique random seed is also assigned to control all of the random states in the prediction stage. Due to the computational cost when evaluating our methods on large benchmark data sets, we reduce our repetitions to 9 total runs (3 data random seeds by 3 PyTorch random seeds) for the Lipophilicity data set. Wilcoxon signed-rank hypothesis testing[44] was conducted between critical model pairs to reliably determine the proposed method's effectiveness. Unless otherwise specified, we used one-sided Wilcoxon signed-rank tests (Wilcoxon) with a significance level $\alpha = 0.05$, where the alternative hypothesis is that the variant model achieves lower RMSE than the baseline. For comparisons aimed at characterizing distributional differences rather than directional improvement (Figure 4 and Table S15), a two-sided test with the same significance level was applied. To control the family wise error rate, all pairwise model comparisons were treated as one hypothesis family, then we adjusted the p-values within each family using the Holm−Bonferroni procedure.[45,46] We used a measure of rank−biserial correlation[47] to report the absolute effect size (see hypotheses, absolute effect sizes, raw and adjusted p-values in Hypothesis testing). We reported adjusted p values (denoted as adjusted p) as the main statistical testing metric in this paper.

RESULTS

MycoPermeNet-v2 Significantly Improves Prediction of Mycomembrane Permeation

We experimented with the proposed MycoPermeNet-v2 model and its variants for the prediction of mycomembrane permeation. Our MycoPermeNet-v2 model—two-stage baseline with Fusion + NST—achieved the best performance, with an RMSE score of 0.719 ± 0.021. Using feature fusion and NST augments the feature space and incorporates unlabeled data, significantly improving the performance and stability over the baseline MycoPermeNet model (baseline vs baseline Fusion + NST, adjusted $p < 0.0001$, Wilcoxon). The combined Fusion + NST strategy also statistically outperforms using either Fusion or NST component alone (baseline + Fusion/baseline + NST vs baseline Fusion + NST, adjusted $p < 0.0001$, Wilcoxon).

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

Table 2. Comparison of Final Performance on the Test Set[a]

Model	R^2 (↑)	RMSE (↓)	Spearman (↑)
baseline (MycoPermeNet)	0.526 ± 0.030	0.755 ± 0.024	0.733 ± 0.021
XGBoost	0.503 ± 0.031	0.757 ± 0.024	0.713 ± 0.022
random forest	0.486 ± 0.025	0.771 ± 0.019	0.697 ± 0.022
MINE-based baseline	0.515 ± 0.029	0.763 ± 0.023	0.730 ± 0.016
unsupervised + baseline	0.467 ± 0.029	0.800 ± 0.022	0.690 ± 0.015
baseline + fusion	0.554 ± 0.032	0.731 ± 0.028	0.755 ± 0.017
baseline + NST	0.533 ± 0.029	0.749 ± 0.023	0.734 ± 0.021
baseline fusion + NST (MycoPermeNet-v2)	**0.570 ± 0.025**	**0.719 ± 0.021**	**0.759 ± 0.016**

[a]Performance is measured by the coefficient of determination R^2, root mean square error (RMSE), and Spearman's rank correlation between model predictions and experimental measurements. Baseline Fusion + NST is the MycoPermeNet-v2 model, baseline is the original MycoPermeNet model.

We performed an ablation study to determine the effect of model choices on final performance, investigating the effects of feature space expansion and incorporating unlabeled data at various training stages. Table 2 and Figure 3 demonstrate the final prediction performance of all of the models on the held-out test set (see statistical test results in Table S11).

Fusing descriptors with learned embeddings expands the feature space and significantly improves the baseline performance (baseline vs baseline + Fusion, adjusted $p < 0.0001$, Wilcoxon). However, training non-GNN regressors, XGBoost and Random Forest, on concatenated RDKit descriptors and ECFPs yields worse performance than the GNN-based baseline MycoPermeNet model. MINE attempts to induce the learned embeddings to better capture the information in RDKit features. But, the MINE-based variant generally underperforms the baseline, and the performance differences are not statistically significant (baseline vs MINE-based baseline, adjusted $p > 0.05$, Wilcoxon).

We find that the effect of incorporating unlabeled data is strategy-dependent. Baseline with NST achieves significant improvement and more stable predictions compared to the baseline (baseline vs baseline + NST, adjusted $p < 0.0001$, Wilcoxon). In contrast, the unsupervised pretraining strategy—where Chemprop is first trained on extensive data and then transfers its weights to initialize the GNN encoder—results in degraded performance relative to the baseline.

MycoPermeNet-v2 is Robust to Split Strategy

We investigate the robustness of our Fusion + NST setup to different split strategies (Table 3), finding that regardless of the data split strategy, our best model consistently achieves significant performance improvements (see statistical test results in Table S12). We note that this paper adopts the scaffold-exclusive split strategy, which is the most rigorous, resulting in the lowest predictive performance. Both scaffold-agnostic and scaffold-inclusive strategies reduce the prediction difficulty by exposing some test set scaffolds or the test set to the first-stage model, leading to higher performance.

Fusion + NST Generalizes across Data Sets and Architectures

We investigate whether our proposed Fusion + NST generalizes to other GNN encoder baselines and molecular property prediction tasks (Table 4). Across the three molecular property regression tasks in MoleculeNet (ESOL, Lipo, and FreeSolv), Fusion + NST achieves statistically significant performance improvements (Figure 4 and Table S15). While Attentive FP with Fusion + NST performs the best on ESOL and Lipo, Chemprop Fusion + NST (MycoPermeNet-v2) is the best choice for FreeSolv and our permeability task. The success of the Fusion + NST strategy is not limited to a specific type of graph representation or a particular molecular property.

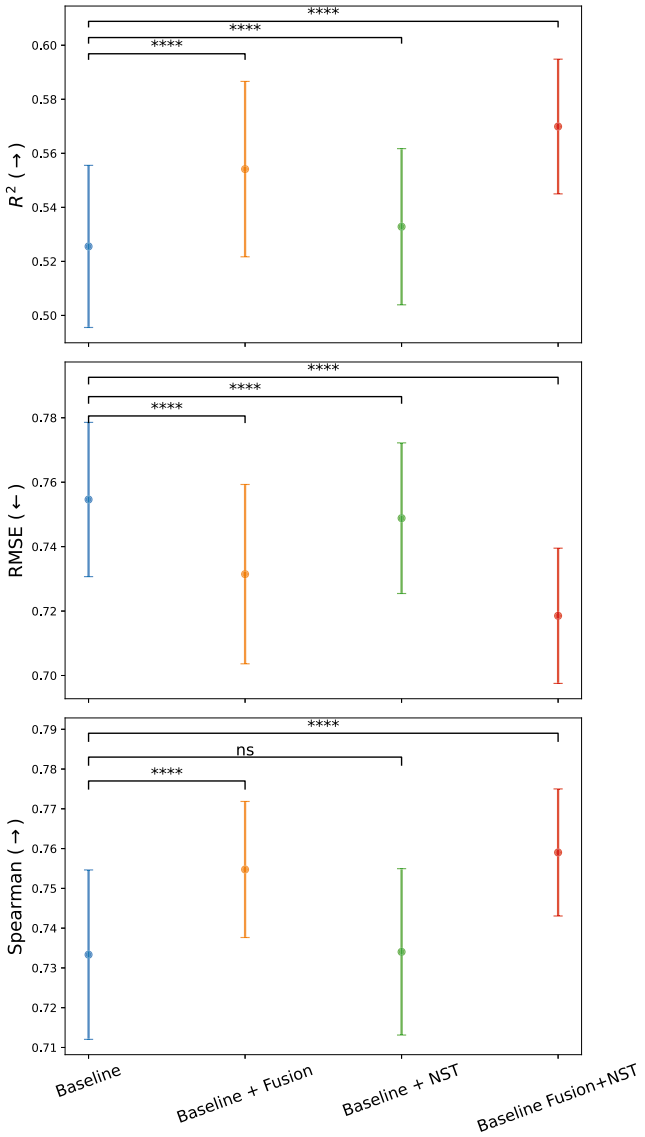


Figure 3. Critical model prediction performance on the test set. The statistical test results of one-sided Wilcoxon signed-rank test between baseline and its primary variants are annotated: ns,* **, ***, and ****, indicating adjusted $p > 0.05$, $p \leq 0.05$, 0.01, 0.001, and 0.0001, respectively.

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

Table 3. Comparison of Final Performance on the Test Set across Different Splits

split strategy	model	R^2 (↑)	RMSE (↓)	Spearman (↑)
Scaffold-exclusive	baseline	0.526 ± 0.030	0.755 ± 0.024	0.733 ± 0.021
	baseline fusion + NST	0.570 ± 0.025	0.719 ± 0.021	0.759 ± 0.016
Scaffold-agnostic	baseline	0.572 ± 0.022	0.666 ± 0.022	0.758 ± 0.014
	baseline fusion + NST	0.597 ± 0.014	0.646 ± 0.016	0.775 ± 0.008
Scaffold-inclusive	baseline	0.728 ± 0.028	0.541 ± 0.029	0.857 ± 0.015
	baseline fusion + NST	0.740 ± 0.021	0.529 ± 0.024	0.864 ± 0.011

Table 4. Test RMSE of Different Models with and without Fusion + NST across Multiple MoleculeNet Datasets[a]

		RMSE (↓)			
GNN encoder	fusion + NST	ESOL	FreeSolv	Lipo	permeability
GCN	×	1.534 ± 0.117	3.158 ± 0.599	0.805 ± 0.030	0.910 ± 0.053
	√	0.967 ± 0.059	2.593 ± 0.409	0.752 ± 0.033	0.741 ± 0.040
GINE	×	1.325 ± 0.166	3.314 ± 0.819	0.834 ± 0.054	0.943 ± 0.087
	√	1.058 ± 0.059	2.612 ± 0.456	0.783 ± 0.036	0.796 ± 0.073
Chemprop	×	1.062 ± 0.056	2.362 ± 0.189	0.694 ± 0.022	0.755 ± 0.024
	√	0.903 ± 0.035	**2.296 ± 0.173**	0.654 ± 0.013	**0.719 ± 0.021**
attentive FP	×	0.910 ± 0.041	2.654 ± 0.616	0.661 ± 0.026	0.804 ± 0.051
	√	**0.899 ± 0.042**	2.408 ± 0.484	**0.648 ± 0.018**	0.769 ± 0.034

[a] √ or × indicates the two-stage GNN encoder + MLP framework with or without fusion + NST strategy.

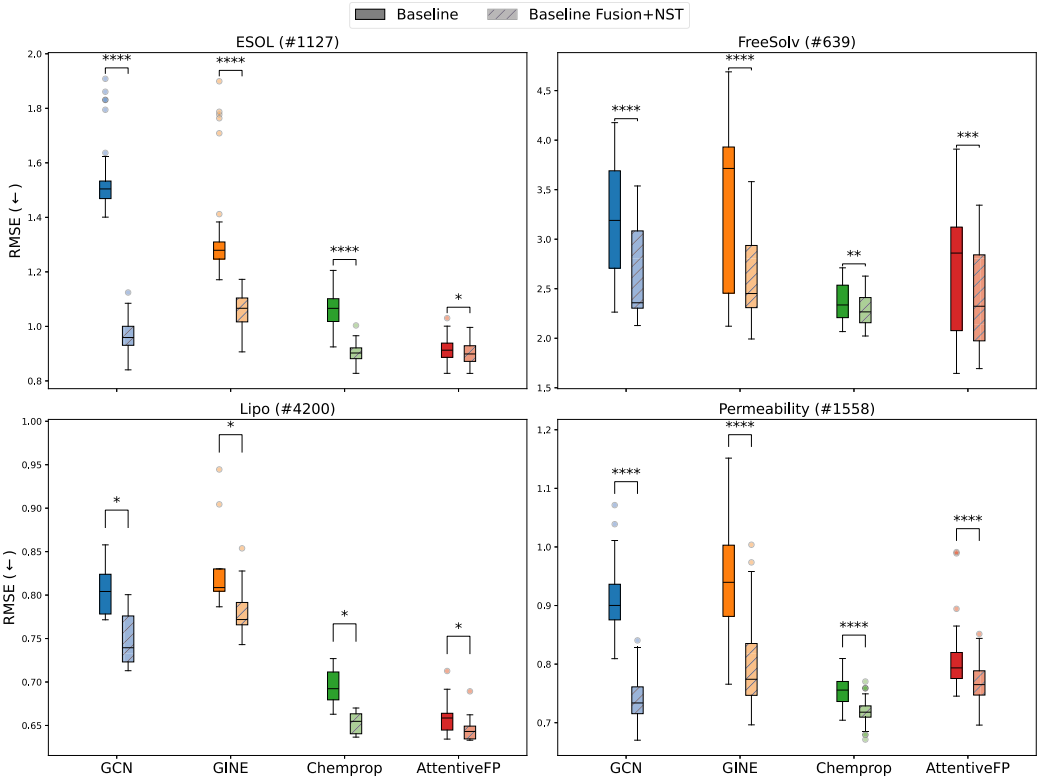


Figure 4. Two-stage baseline with or without Fusion + NST test performance box plots across different GNN encoders and MoleculeNet data sets. The statistical test results of two-sided Wilcoxon signed-rank test between each model pair are annotated: ns,* **, ***, and ****, indicating adjusted $p > 0.05$, $p \leq 0.05$, 0.01, 0.001, and 0.0001, respectively.

For all subsequent experiments on these data sets, we use the best-performing GNN encoder identified for each data set in the two-stage model.

Feature Fusion Leverages Chemically Important Information

We hypothesize that the effect of feature fusion depends on the relevance of the chemical information to the prediction task. To test this, we curate 23 descriptors that are tailored to *M. tuberculosis*-related properties, including 3D features such as globularity and plane of best fit, based on previous observations about permeability-associated features.[11] As a control for the selected *M. tuberculosis*-related set, we also construct an RDKit random set by randomly sampling 23 descriptors from the full set of 184 nonzero descriptors, obtained after removing RDKit descriptors not present in any molecule in the data set (see MycoPermeNet-v2 Model). The

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

Table 5. Model Performance under Different Descriptor Fusion Settings

descriptors	model	R^2 (↑)	RMSE (↓)	Spearman (↑)
-	baseline	0.526 ± 0.030	0.755 ± 0.024	0.733 ± 0.021
RDKit full (184)	baseline + fusion	0.554 ± 0.032	0.731 ± 0.028	0.755 ± 0.017
	baseline fusion + NST	0.570 ± 0.025	0.719 ± 0.021	0.759 ± 0.016
M. tuberculosis-related (23)	baseline + fusion	0.542 ± 0.030	0.742 ± 0.025	0.740 ± 0.019
	baseline fusion + NST	0.549 ± 0.028	0.736 ± 0.023	0.742 ± 0.018
RDKit random (23)	baseline + fusion	0.532 ± 0.032	0.749 ± 0.026	0.738 ± 0.024
	baseline fusion + NST	0.541 ± 0.032	0.742 ± 0.026	0.741 ± 0.021

Table 6. NST Generalizability across Different Injected Unlabeled Datasets

injected data set	model	R^2 (↑)	RMSE (↓)	Spearman (↑)
-	baseline	0.526 ± 0.030	0.755 ± 0.024	0.733 ± 0.021
MLSMR	baseline + NST	0.533 ± 0.029	0.749 ± 0.023	0.734 ± 0.021
	baseline fusion + NST	0.570 ± 0.025	0.719 ± 0.021	0.759 ± 0.016
in-house	baseline + NST	0.531 ± 0.032	0.750 ± 0.025	0.734 ± 0.023
	baseline fusion + NST	0.564 ± 0.030	0.723 ± 0.024	0.759 ± 0.016

Table 7. Baseline + NST Test RMSE across Different Iteration Counts and Labeled Datasets[a]

		RMSE (↓)			
data sets	volume	Iter 0	Iter 1	Iter 2	Iter 3
FreeSolv	639	2.362 ± 0.189	**2.350 ± 0.192**	2.381 ± 0.182	2.394 ± 0.188
ESOL	1127	0.910 ± 0.041	0.908 ± 0.039	**0.907 ± 0.037**	0.910 ± 0.042
permeability	1558	0.755 ± 0.024	0.753 ± 0.025	**0.749 ± 0.023**	0.750 ± 0.024
Lipo	4200	0.661 ± 0.026	0.661 ± 0.023	0.661 ± 0.025	**0.657 ± 0.025**

[a]We use the GNN encoder found optimal for each dataset in prior experiments: Chemprop for permeability and FreeSolv, and Attentive FP for ESOL and Lipo.

M. tuberculosis-related properties and 184 nonzero RDKit descriptors are listed in Table S17 and S16.

Table 5 demonstrates that fusing only 23 *M. tuberculosis*-related 23 descriptors has superior performance to the two-stage baseline model and to the random 23 features models. However, its performance is less than that of the full set of 184 nonzero RDKit descriptors. Importantly, our fusion strategy significantly improves the baseline model (see statistical test results in Table S13) and remains robust across descriptors of different sizes and contents, indicating its general applicability.

Performance Analysis of NST under Different Conditions

Generalizability on Different Unlabeled Data Sets. To assess the generalizability of NST across different unlabeled data sets, we experiment with two unlabeled data sets of different sizes (Section Data Set Preparation and Splits). 500 unlabeled data points are randomly sampled from the in-house data set. As shown in Table 6, all the models trained by NST have significant improvements over the baseline model's performance (see statistical test results in Table S14). This suggests that NST's ability to enhance prediction performance does not rely on a specific unlabeled data set, nor does it require a large quantity of unlabeled samples.

Performance across Labeled Data Sets with Different Volumes. Besides the generalizability over unlabeled data sets, we incorporate three labeled data sets from MoleculeNet benchmark to show NST's capability. We use the same approach shown in Table S5 to select the unlabeled data volume of each NST iteration for these data sets—500 for FreeSolv, 400 for ESOL, and 2000 for Lipo—which also shows a proportional relationship with the data set volume. Table 7 demonstrates that NST consistently improves the two-stage baseline performance of different molecular property prediction tasks. We also note that as the data set volume expands, NST requires more iterations to reach the best performance.

Gains of Using NST under Different Teacher Models. We evaluate the gains of using NST on top of different teacher models, that is, the two-stage baseline and baseline + Fusion models. Table S9 indicates that NST tends to be more effective when applied to baseline + Fusion than to the baseline, leading to larger reductions in RMSE. Furthermore, we directly compute the performance improvement (RMSE reduction) of NST over either baseline + Fusion across multiple runs. The results in Table 8 show that NST generally yields greater gains when applied on top of baseline + Fusion (2 out of 4 data sets), while in one case, the improvements are comparable, and in another, the baseline alone benefits more.

Table 8. Performance Improvements (RMSE Reduction) of NST over Two-Stage Baselines and Baseline + Fusion[c]

data set	Δ_{Baseline}[a]	$\Delta_{\text{Baseline+Fusion}}$[b]
FreeSolv	0.028 ± 0.036	0.050 ± 0.062
ESOL	0.013 ± 0.013	0.012 ± 0.015
permeability	0.010 ± 0.009	0.019 ± 0.019
Lipo	0.011 ± 0.007	0.011 ± 0.006

[a]The averaged RMSE reduction of baseline + NST and baseline over multiple runs. [b]The averaged RMSE reduction of baseline Fusion + NST and baseline + Fusion over multiple runs. [c]Select the best NST result under each random state, compute its difference from the corresponding baseline, and then aggregate the statistics. We use the GNN encoder found optimal for each dataset in prior experiments: Chemprop for permeability and FreeSolv, and Attentive FP for ESOL and Lipo.

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

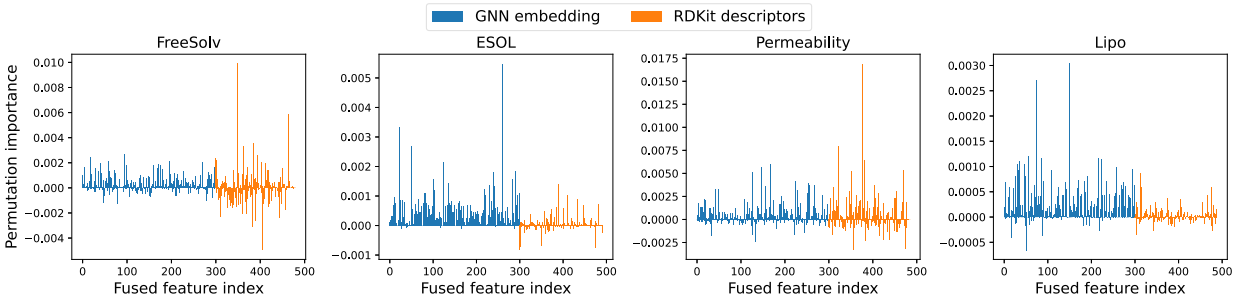


Figure 5. Permutation importance analysis of the fused feature space across four data sets. The fused representation contains learned GNN embeddings (blue) and RDKit descriptors (orange). For each feature, the importance is computed as the decrease in the model's R^2 after random permutation of that feature while keeping all others fixed. We use the GNN encoder found optimal for each data set in prior experiments: Chemprop for permeability and FreeSolv, and Attentive FP for ESOL and Lipo. The best baseline Fusion + NST model for each data set is used among all the runs.

Model Interrogation and Interpretability

Feature Importance in the Fused Feature Space. To quantify the relative contribution of the GNN embeddings and RDKit descriptors, we computed the permutation importance for each feature within the fused feature space. As shown in Figure 5, both types of representations contribute to the final prediction, and their relative importance exhibits different patterns across data sets. Therefore, we counted how many features from each group appear among the top-50 most important fused features. Table 9 shows that features from GNN embeddings dominate the important dimensions in three out of four data sets and are similar to the contribution of the RDKit features in the fourth. We find that the top three most important RDKit descriptors in the primary permeability data set were SLogP_VSA8[48] (lipophilic local surface area), BCUT2D_MRLOW[49] (electronic polarizability and global topology via molecular refractivity), and EState_VSA3[50] (spatial distribution of local electronic states). These three factors generally describe lipophilicity and local electronic states, which likely influence the ability of small molecules to partition into the mycolic acids of the mycomembrane.[7,51] While factors influencing mycomembrane permeability are still under active investigation, both lipophilicity (log P) and molecular refractivity (MR) were previously identified in Lepori et al.[11] as the most influential features in mycomembrane permeation.

Table 9. Number of GNN-Derived Features and RDKit Descriptors among the Top-50 Most Important Fused Features

feature amount	FreeSolv	ESOL	permeability	Lipo
#GNN features	32	44	26	47
#RDKit descriptors	18	6	24	3

Effectiveness of Fusion + NST in the Low-Data Regime. Because Fusion + NST has demonstrated promising performance on small data sets (Table 4), we further show its effectiveness in low-data scenarios by randomly sampling a

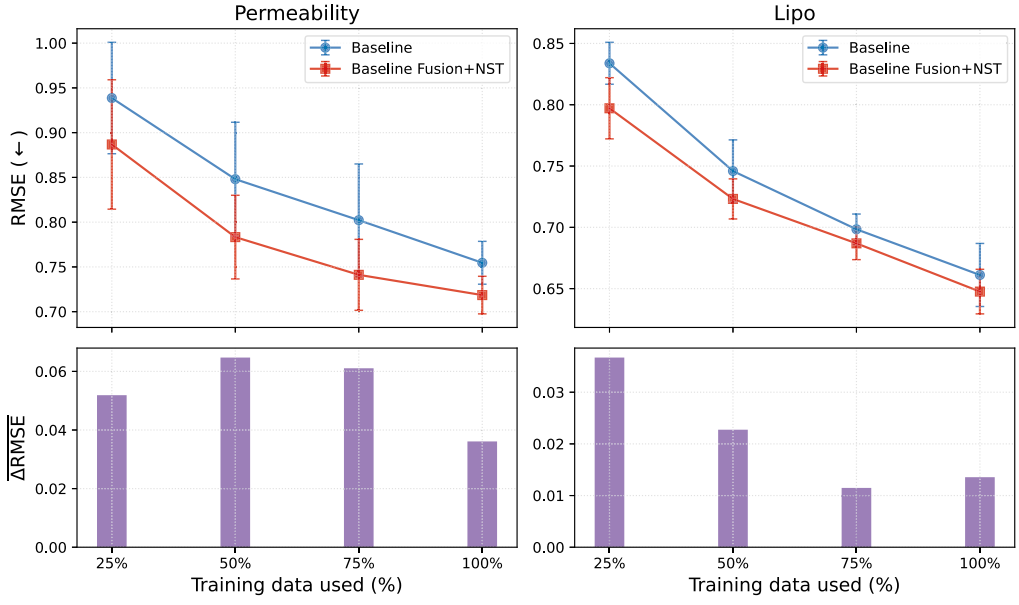


Figure 6. Performance of Fusion + NST under different labeled-data fractions on the permeability and Lipophilicity data sets. Top panels: RMSE of two-stage baseline and baseline Fusion + NST models when trained with 25%, 50%, 75%, and 100% of the labeled training data. The best GNN encoders, Chemprop and Attentive FP, are applied to permeability and Lipophilicity data sets, respectively. Bottom panels: corresponding average reduction in RMSE ($\overline{\Delta\text{RMSE}}$) achieved by Fusion + NST relative to each baseline.

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

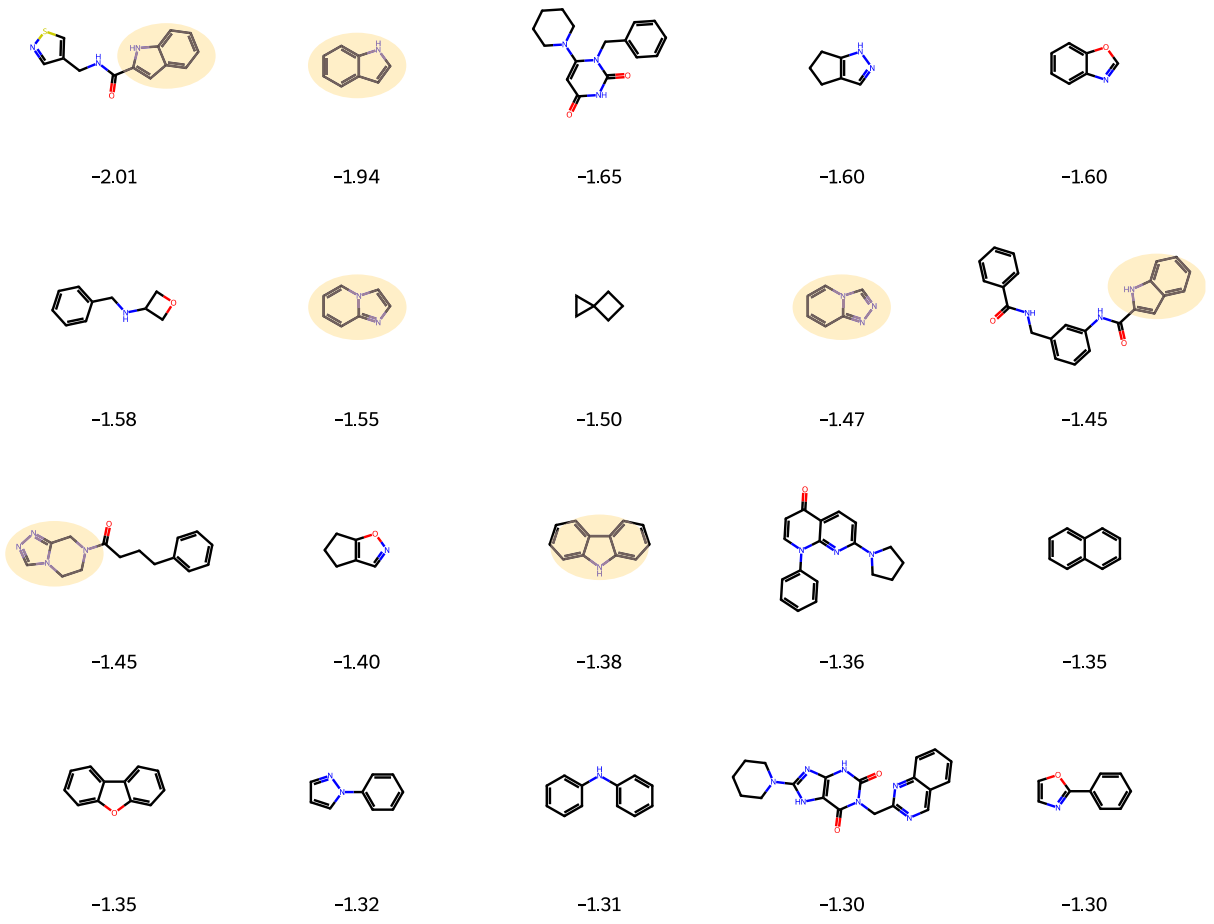


Figure 7. Scaffold model prediction analysis. The highlighted regions are the nitrogen heterocycles observed in the top 20 most permeable scaffolds predicted by the best MycoPermeNet-v2 model.

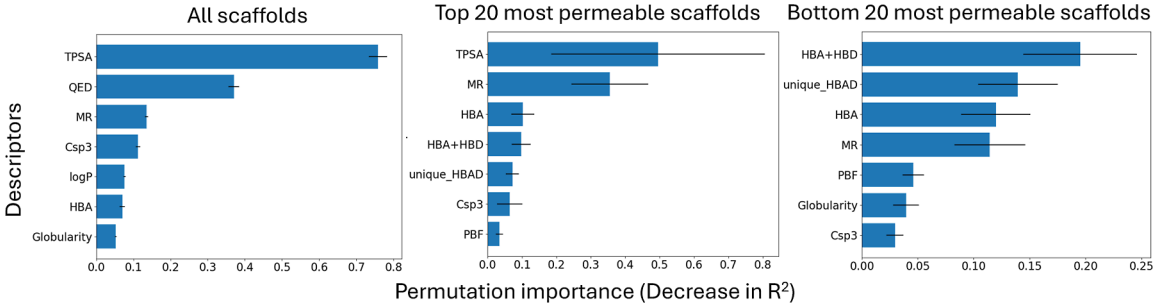


Figure 8. Top ranking physicochemical features. Permutation importance using a surrogate model trained on MycoPermeNet-v2 model predictions against the 23 descriptors. This shows the importance of the top (middle) and bottom (leftmost) most permeable scaffolds, further examining which of the physicochemical features are ranked high as the most important.

percentage of training samples to train the corresponding two-stage baseline with or without the Fusion + NST approach. The validation and test set volumes remain fixed across all training-fraction conditions. We conducted experiments on our permeability data set and the largest of the MoleculeNet data sets, Lipophilicity data set. For both tasks, as Figure 6 illustrates, Fusion + NST offers performance gain over the baseline across all training set fractions, but gains over the baseline are generally larger in the low-data settings.

Model Reliability Analysis for Permeability Prediction. We then interrogated the robustness of MycoPermeNet-v2 across the permeability and feature values. The parity plot comparing ground-truth values with the mean and standard deviation of MycopermeNet-v2 predictions on the held-out test set demonstrates consistent performance across ground-truth values (Figure S4). We further assessed robustness by examining the model's sensitivity to variation in representative physicochemical descriptors. No systematic increase in prediction error was observed across descriptor values (Figure S5).

Scaffold-Level Evaluation. To validate that MycoPermeNet-v2 learns biologically relevant features, we evaluated its performance across different molecular scaffolds via the Bemis−Murcko scaffolding process.[36] Following the methodology from Lepori et al.,[11] we used the trained best model to predict permeability for all scaffold classes in the data set and

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

ranked these classes by decreasing permeability. Our analysis showed that 7 of the top 20 most permeable scaffolds were nitrogen-containing heterocycles (Figure 7), consistent with the findings from the baseline MycoPermeNet model.[11]

We hypothesized that molecular features contributing to enhanced mycomembrane permeability would be consistently ranked as important both in the full data set and among the highest-permeability scaffolds, while being less important among the lowest-permeability scaffolds, and vice versa for features associated with reduced permeability. To test this hypothesis, permutation importance analyses using a surrogate Random Forest model were performed on the full data set, fitting the 23 descriptors and their corresponding predictions from the MycoPermeNet-v2 model (Table S17), as well as separately on the top 20 and bottom 20 scaffolds ranked by predicted permeability. Under this framework, features that are highly ranked in both the full data set and the top 20 scaffolds but absent or reduced in importance in the bottom 20 are interpreted as contributing to increased mycomembrane permeability. This analysis revealed that changes in topological surface area (TPSA) showed a great impact in raising mycomembrane permeability, being present in both the full permutation importance and top 20 scaffolds permeability test but not in the bottom 20 (Figure 8). This is in agreement with Lepori et al.[11] who have also reported that TPSA possesses strong correlation with the mycomembrane.

DISCUSSION

This paper introduces MycoPermeNet-v2, a two-stage model to predict the compound permeability through *M. tuberculosis*'s mycomembrane, achieving statistically significant improvement compared to the baseline. We show that the improvement is due to two model design choices: (1) fusing chemically informative features from RDKit with learned molecular embeddings and (2) training the model with an adapted Noisy Student Training, to allow learning from unlabeled data. We show that our approach generalizes to other GNN encoder architectures and to physical chemistry property prediction tasks from MoleculeNet. Comprehensive analyses were conducted to validate our method's robustness, generalizability, and sensitivity to random seeds.

We hypothesize that our feature fusion approach achieved success by enriching the representations learned by graph neural networks, which reflect the local chemical structure, with molecular descriptors that reflect the global chemical structure and high-level physicochemical and topological properties. Even though Attentive FP, a GNN with an attention mechanism, has shown success in extracting global features, our fusion strategy still provides modest improvement, particularly in cases where Attentive FP was not the best model (see results in Table S10). The effectiveness of feature fusion does not stem from simply increasing the number of input features but from incorporating meaningful and task-relevant information (Table 5). We also experimented with MINE, which directly estimates the mutual information between embeddings and descriptors and incorporates it as an auxiliary loss to supervise the representation learning phase. However, this degrades model performance, likely because forcing the local and global information to align via mutual information maximization may distort the learned representation.

The semisupervised NST framework utilizes extensive unlabeled data to help the model extract useful information beyond the limited labeled set, improving model performance and reducing variability. We show that the model benefits from NST regardless of the unlabeled data set used. A natural hypothesis is that under limited supervision, increasing unlabeled data and NST iterations would enhance model performance. However, as shown in Table S5, baseline + NST typically reaches peak validation performance in the first iteration across different data volumes, suggesting limited gains from further scaling. Table 7 shows that with more labeled data, the baseline + NST model needs more iterations to reach optimal performance (3 iterations, Lipo). In contrast, on smaller data sets, the model needs fewer iterations to obtain the best NST performance (1 iteration, FreeSolv; 2 iterations, permeability and ESOL). We attribute this phenomenon to two main factors. First, a larger data set may contain richer features, requiring more iterations to distill. Second, more data may result in a more accurate teacher model, which in turn demands more iterations for effective knowledge transfer to the student models.

Besides the scaling characterization of NST, we also observe that the degree of NST's improvement varies across different teacher models (Tables S9 and 8). Specifically, the improvement brought by NST is generally larger when applied to baseline + Fusion compared to baseline. We assume that the effectiveness of NST is highly related to the teacher model's performance, as a more capable teacher model provides better guidance and improves the student model more, whereas a weaker teacher yields smaller gains on the same task. This phenomenon should be understood as a characteristic of NST rather than a limitation, since its effectiveness is closely tied to the quality of the teacher model.

Currently our technique is limited to regression tasks and has not been evaluated on classification tasks. Additional improvements in prediction may be yielded by fusing additional chemical features or more advanced semisupervised learning techniques. While our model had statistically significant improvements in generalization to unseen scaffolds, due to data availability, we have yet to assess the model's performance on compounds from an external test set. Measurement on such a set would be the most rigorous evaluation of generalizability, and differences in the data distribution could lead to differences in performance. Furthermore, any biases present in the training data, such as potential effects of the azide tag on permeability of compounds in the permeability data set, would be inherited by the trained models. Additional applications could then involve the use of our model to predict permeability-enhancing modifications to antibiotic compounds.

ASSOCIATED CONTENT

Data Availability Statement

The source code, data, and best pretrained model weights are available on GitHub at https://github.com/SAGE-Lab-UMass/MycoPermeNet-v2-pub. Our PyTorch Geometric version of Chemprop is adapted from https://github.com/itakigawa/pyg_chemprop.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c02435.

Implementation details about the model, supplementary results, and hypothesis testing results (PDF)

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996

AUTHOR INFORMATION

Corresponding Author

Anna G. Green − *Manning College of Information & Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States;* orcid.org/0000-0001-7548-3682; Email: annagreen@umass.edu

Authors

Nelson Evbarunegbe − *Manning College of Information & Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States;* orcid.org/0000-0003-0408-5016

Shiyun Wa − *Manning College of Information & Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States;* orcid.org/0000-0003-2949-5492

Isha Karn − *Manning College of Information & Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c02435

Author Contributions

†N.E. and S.W. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the members of the UMass SAGE lab and Siegrist lab for valuable input and discussion. This work utilized resources from Unity, a collaborative, multi-institutional high-performance computing cluster managed by UMass Amherst Research Computing and Data. This work was partially funded by an award from the Models to Medicine Center in the Institute of Applied Life Sciences at the University of Massachusetts Amherst.

REFERENCES

(1) Stokes, J. M.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688−702.

(2) Liu, G.; et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nat. Chem. Biol.* **2023**, *19*, 1342−1350.

(3) Swanson, K.; Liu, G.; Catacutan, D. B.; Arnold, A.; Zou, J.; Stokes, J. M. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nat. Mach. Intell.* **2024**, *6*, 338−353.

(4) WHO Global Tuberculosis Report. 2023. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023/tb-disease-burden/1-2-tb-mortality (accessed April 16, 2025).

(5) Jarlier, V.; Nikaido, H. Mycobacterial cell wall: Structure and role in natural resistance to antibiotics. *FEMS Microbiol. Lett.* **1994**, *123*, 11−18.

(6) Lepori, I.; Jackson, K.; Liu, Z.; Chordia, M. D.; Wong, M.; Rivera, S. L.; Roncetti, M.; Poliseno, L.; Freundlich, J. S.; Pires, M. M.; Siegrist, M. S. The Mycomembrane Differentially and Heterogeneously Restricts Antibiotic Permeation. *ACS Infect. Dis.* **2025**, *11*, 1893−1906.

(7) Dulberger, C. L.; Rubin, E. J.; Boutte, C. C. The mycobacterial cell envelope-a moving target. *Nat. Rev. Microbiol.* **2020**, *18*, 47−59.

(8) Brennan, P. J.; Nikaido, H. THE ENVELOPE OF MYCOBACTERIA. *Annu. Rev. Biochem.* **1995**, *64*, 29−63.

(9) Nagamani, S.; Sastry, G. N. Mycobacterium tuberculosis Cell Wall Permeability Model Generation Using Chemoinformatics and Machine Learning Approaches. *ACS Omega* **2021**, *6*, 17472−17482.

(10) Banerjee, A.; Sharma, A.; Kamble, P.; Garg, P. Prediction of Mycobacterium tuberculosis cell wall permeability using machine learning methods. *Mol. Diversity* **2024**, *28*, 2317−2329.

(11) Lepori, I.; Liu, Z.; Evbarunegbe, N.; Feng, S.; Brown, T. P.; Mane, K.; Shivangi; Wong, M.; George, A.; Guo, T.; et al. Identification of chemical features that influence mycomembrane permeation and antitubercular activity. *bioRxiv* **2025**, 2025.02.27.640664.

(12) Wen, N.; Liu, G.; Zhang, J.; Zhang, R.; Fu, Y.; Han, X. A fingerprints based molecular property prediction method using the BERT model. *J. Cheminf.* **2022**, *14*, 71.

(13) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, USA*, 2019; pp 429−436.

(14) Zhang, X.-C.; Wu, C.-K.; Yang, Z.-J.; Wu, Z.-X.; Yi, J.-C.; Hsieh, C.-Y.; Hou, T.-J.; Cao, D.-S. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings Bioinf.* **2021**, *22*, bbab152.

(15) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph. *arXiv* **2020**, arXiv:1905.12265.

(16) Xia, J.; Zhao, C.; Hu, B.; Gao, Z.; Tan, C.; Liu, Y.; Li, S.; Li, S. Z. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. In *Eleventh International Conference on Learning Representations*, 2023.

(17) Zhang, W.; Deng, L.; Zhang, L.; Wu, D. A Survey on Negative Transfer. *IEEE CAA J. Autom. Sinica* **2023**, *10*, 305−329.

(18) Xie, Q.; Luong, M.-T.; Hovy, E.; Le, Q. V. Self-training with Noisy Student improves ImageNet classification. *arXiv* **2020**, arXiv.1911.04252.

(19) Liu, Y.; Lim, H.; Xie, L. Exploration of chemical space with partial labeled noisy student self-training and self-supervised graph embedding. *BMC Bioinf.* **2022**, *23*, 158.

(20) Jumper, J.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(21) Abramson, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493−500.

(22) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(23) Wang, J.; Zhang, L.; Sun, J.; Yang, X.; Wu, W.; Chen, W.; Zhao, Q. Predicting drug-induced liver injury using graph attention mechanism and molecular fingerprints. *Methods* **2024**, *221*, 18−26.

(24) Pham, M.; Cho, M.; Joshi, A.; Hegde, C. Revisiting Self-Distillation. *arXiv* **2022**, arXiv.2206.0849.

(25) Liu, Z.; Lepori, I.; Chordia, M. D.; Dalesandro, B. E.; Guo, T.; Dong, J.; Siegrist, M. S.; Pires, M. M. A Metabolic-Tag-Based Method for Assessing the Permeation of Small Molecules Across the Mycomembrane in Live Mycobacteria. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202217777.

(26) Ferraro, N. J.; Kim, S.; Im, W.; Pires, M. M. Systematic Assessment of Accessibility to the Surface of Staphylococcus aureus. *ACS Chem. Biol.* **2021**, *16*, 2527−2536.

(27) Ongwae, G. M.; Liu, Z.; Feng, S.; Chordia, M. D.; Dash, R.; He, Y.; Gh, M. S.; Dalesandro, B. E.; Guo, T.; Sharpless, K. B.; Dong, J.; Siegrist, M. S.; Im, W.; Pires, M. M. Click-Based Determination of Accumulation of Molecules in Escherichia coli. *bioRxiv* **2025**, 2023.06.20.545103.

(28) Jewett, J. C.; Bertozzi, C. R. Cu-free click cycloaddition reactions in chemical biology. *Chem. Soc. Rev.* **2010**, *39*, 1272−1279.

(29) Agard, N. J.; Prescher, J. A.; Bertozzi, C. R. A Strain-Promoted [3 + 2] Azide-Alkyne Cycloaddition for Covalent Modification of Biomolecules in Living Systems. *J. Am. Chem. Soc.* **2004**, *126*, 15046−15047.
(30) RDKit: Open-source cheminformatics. 2025. https://www.rdkit.org (accessed April 16, 2025).
(31) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2017**, *9*, 513.
(32) Molecular Libraries Small Molecule Repository (MLSMR). 2025. https://reporter.nih.gov/project-details/8325475 (accessed April 16, 2025).
(33) Enamine REAL Space. 2025. https://enamine.net/compound-collections/real-compounds/real-space-navigator (accessed: April 16, 2025).
(34) Xie, L.; Xu, L.; Kong, R.; Chang, S.; Xu, X. Improvement of Prediction Performance With Conjoint Molecular Fingerprint in Deep Learning. *Front. Pharmacol.* **2020**, *11*, 606668.
(35) Han, J.; Kwon, Y.; Choi, Y.-S.; Kang, S. Improving chemical reaction yield prediction using pre-trained graph neural networks. *J. Cheminf.* **2024**, *16*, 25.
(36) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(37) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 9−17.
(38) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. *arXiv* **2019**, arxiv:1903.02428.
(39) Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, R. D. MINE: Mutual Information Neural Estimation. *arXiv* **2021**, arxiv:1801.04062.
(40) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2017**, arXiv:1609.02907.
(41) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2020**, *63*, 8749−8760.
(42) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(43) Pecher, B.; Srba, I.; Bielikova, M. A Survey on Stability of Learning with Limited Labelled Data and its Sensitivity to the Effects of Randomness. *ACM Comput. Surv.* **2025**, *57*, 1.
(44) Woolson, R. F. *Wiley Encyclopedia of Clinical Trials*; John Wiley & Sons, Ltd, 2008; pp 1−3.
(45) Lydersen, S. Adjustment of p values for multiple hypotheses: why, when and how. *Ann. Rheum. Dis.* **2024**, *83*, 1254−1255.
(46) Bender, R.; Lange, S. Adjusting for multiple testing-when and how? *J. Clin. Epidemiol.* **2001**, *54*, 343−349.
(47) Cureton, E. E. Rank-biserial correlation. *Psychometrika* **1956**, *21*, 287−290.
(48) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.
(49) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(50) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(51) Marrakchi, H.; Lanéelle, M.-A.; Daffé, M. Mycolic Acids: Structures, Biosynthesis, and Beyond. *Chem. Biol.* **2014**, *21*, 67−85.

https://doi.org/10.1021/acs.jcim.5c02435
J. Chem. Inf. Model. 2026, 66, 2985−2996