7 August 2026

ActiveFusion: Fused representations improve active learning for molecular property prediction

Nelson Evbarunegbe[1], Shiyun Wa[1], Luke Taylor[1], Anna G. Green[1]

1. University of Massachusetts Amherst

Abstract

Active learning provides an efficient strategy for molecular property prediction by iteratively prioritizing compounds for experimental evaluation. However, the effectiveness of active learning pipelines depends strongly on the choice of molecular representation, and systematic understanding of how representation families affect the active learning process in uncertainty and predictive performance remain limited.In this work, we introduce ActiveFusion, a framework for integrating heterogeneous molecular representations within active learning workflows for molecular property prediction. The framework enables systematic evaluation of physicochemical descriptors, molecular fingerprints, learned graph neural network (GNN) representations, and pretrained Transformer-based representations, as well as feature-level fusion strategies that combine complementary chemical information sources.ActiveFusion evaluates models across four molecular property prediction regression tasks. Across datasets, we demonstrate that feature fusion between learned graph representations and physicochemical descriptors consistently improves prediction performance and discovery (average final iteration R^2 of 0.71, 0.67 and 0.54 for our overall best representations Chemprop+RDKit, Chemprop, and RDKit respectively). We show that exploration-driven acquisition strategies enhance scaffold coverage and promote sampling of structurally novel regions of chemical space, and that model-agnostic acquisition of new compounds based on diversity has strong performance. Notably, both pretrained and fine-tuned Transformer-based embeddings do not consistently outperform physicochemical features or GNN-learned representations in our setting, highlighting the continued relevance of chemically interpretable features and learned features from supervised, task-specific models for active learning application in molecular property prediction.Overall, ActiveFusion provides a systematic framework for studying representation–acquisition interactions in molecular discovery with representation fusion capabilities. Our study offers practical guidance for designing active learning pipelines that balance prediction accuracy with chemical space exploration.

 https://doi.org/10.26434/chemrxiv.15003911/v2

ActiveFusion: Fused representations improve active learning for molecular property prediction

Nelson Evbarunegbe[1], Shiyun Wa[1], Luke Taylor[1], and Anna G. Green*[1]

[1]Manning College of Information and Computer Sciences, University of Massachusetts Amherst, 140 Governors Dr, Amherst, 01003, MA, United States

*Email: annagreen@umass.edu

Abstract

Active learning provides an efficient strategy for molecular property prediction by iteratively prioritizing compounds for experimental evaluation. However, the effectiveness of active learning pipelines depends strongly on the choice of molecular representation, and systematic understanding of how representation families affect the active learning process in uncertainty and predictive performance remain limited.

In this work, we introduce ActiveFusion, a framework for integrating heterogeneous molecular representations within active learning workflows for molecular property prediction. The framework enables systematic evaluation of physicochemical descriptors, molecular fingerprints, learned graph neural network (GNN) representations, and pretrained Transformer-based representations, as well as feature-level fusion strategies that combine complementary chemical information sources.

ActiveFusion evaluates models across four molecular property prediction regression tasks. Across datasets, we demonstrate that feature fusion between learned graph representations and physicochemical descriptors consistently improves prediction performance and discovery (average final iteration R^2 of 0.71, 0.67 and 0.54 for our overall best representations Chemprop+RDKit, Chemprop, and RDKit respectively). We show that exploration-driven acquisition strategies enhance scaffold coverage and promote sampling of structurally novel regions of chemical space, and that model-agnostic acquisition of new compounds based on diversity has strong performance. Notably, both pretrained and finetuned Transformer-based embeddings do not consistently outperform physicochemical features or GNN-learned representations in our setting, highlighting the continued relevance of chemically interpretable features and learned features from supervised, task-specific models for active learning application in molecular property prediction.

Overall, ActiveFusion provides a systematic framework for studying representation–acquisition interactions in molecular discovery with representation fusion capabilities. Our study offers practical guidance for designing active learning pipelines that balance prediction accuracy with chemical space exploration.

Keywords

Representation learning, active learning, uncertainty-based acquisition, diversity sampling, cheminformatics, drug discovery, machine learning, deep learning, model generalizability

Introduction

The identification of molecules with desirable biological and physicochemical properties remains a central challenge in drug discovery. In many such contexts, labeled data is scarce because experimental measurements are costly and time-consuming [1], creating a low-data regime in which predictive models must be built from only a small initial labeled set – hundreds to low thousands of molecules. Machine learning models for molecular property prediction have become increasingly important for prioritizing candidate compounds prior to experimental testing, thereby reducing the cost and time associated with laboratory screening [2]. Active learning provides a natural strategy for this setting by guiding the iterative selection of compounds for experimental evaluation [3–5]. By allocating labeling effort to high-value candidates, active learning can accelerate discovery while reducing the number of experiments required. Acquisition strategies based on uncertainty estimation, entropy, compound diversity, or expected improvement have been widely explored in chemical discovery applications [6–9]. However, the effectiveness of active learning is closely linked to the quality of the underlying molecular representation [10, 11]: if the representation fails to capture relevant chemical information, uncertainty estimates and selection strategies may become unreliable.

The choice of molecular representation has a major effect on performance of molecular property prediction models [10, 12–15]. Traditional cheminformatics representations, including molecular fingerprints and physicochemical descriptors [16, 17], which are commonly known as fixed representations, encode global molecular properties such as polarity, lipophilicity, atomic content and surface area that are often directly related to biological activity [18–20]. In contrast, learned representations derived from deep learning models capture relational and structural information through data-driven training procedures. For example, graph neural networks (GNNs) [21–23] and Transformer-based architectures [24, 25] learn molecular representations from molecular graphs or Simplified Molecular Input Line Entry System (SMILES) strings, depending on the input representation and model [26]. These representation families may therefore provide complementary perspectives on molecular structure, and indeed no single representation consistently performs best across different chemical tasks or data regimes [12–14, 27]. Integrating heterogeneous molecular representations may therefore provide a promising strategy for improving prediction performance [28].

While representation fusion has demonstrated clear benefits in standard supervised molecular property prediction [21, 28, 29], it has not been systematically studied within active learning frameworks and across molecular representation families. Prior studies in active learning for molecular property prediction have primarily focused on evaluating acquisition strategies or surrogate models using a single representation [11, 30–32], without systematically examining how representation choice influences active learning dynamics. Recent studies have explored combining multiple molecular features within active learning through stacking ensembles or meta-learning frameworks to improve performance on individual prediction tasks [33, 34]. However, these approaches perform fusion at the prediction level rather than at the representation level, and do not systematically benchmark the effect of representation-level fusion across multiple molecular property prediction tasks, molecular representations, and acquisition strategies.

In this work, we introduce ActiveFusion, a framework for integrating learned molecular embeddings and physicochemical descriptors in active learning for molecular property prediction. The approach is designed to leverage complementary information encoded by different representation families to improve prediction performance. We evaluate ActiveFusion across multiple low-data regime molecular property prediction tasks, including ESOL, FreeSolv, and Lipophilicity from MoleculeNet [35], and a mycomembrane permeability prediction dataset [36]. We show that a Chemprop+RDKit fusion active learning model has the highest predictive performance. For a case study to further validate our best molecular representation active learning

setup from our ActiveFusion framework, we used the mycomembrane permeability prediction dataset as a motivating application relevant to antibiotic discovery [36], where labeling efforts are challenging due to slow organismal growth in culture. We show that our ActiveFusion framework acquires new compounds to recommend for experimental labeling that are more diverse, and therefore likely more useful for modeling, relative to other unlabeled compounds in the unlabeled pool.

To the best of our knowledge, this work presents the first comprehensive benchmark and analysis of representation-level fusion for active learning in molecular property prediction, systematically evaluating multiple fused representations, acquisition functions, and datasets. Through systematic evaluation of representation fusion strategies and active learning dynamics, we demonstrate that integrating heterogeneous molecular features can improve model performance while facilitating the discovery of novel chemical scaffolds.

Methods

Datasets

To evaluate the ActiveFusion framework, we conducted experiments on a mycomembrane permeability dataset and multiple benchmark molecular property prediction datasets. The permeability dataset is made up of a collection of 1,558 compounds annotated with experimentally measured permeability values obtained using a bio-orthogonal click-chemistry assay performed in live *Mycobacterium tuberculosis* cells [37]. The assay measures intracellular compound accumulation while controlling for baseline chemical reactivity, enabling quantitative assessment of permeability across the mycobacterial cell envelope. Previous works have demonstrated the effectiveness of machine learning on this dataset, and the identification of model-learned chemical features that aligns with scientific intuition [28, 36]. We evaluate performance on three additional low-data regime and widely used molecular property prediction benchmark datasets from MoleculeNet [35]: ESOL [38], FreeSolv [39], and Lipophilicity [40]. Table 1 shows the detailed information of these datasets. These four datasets provide diverse regression tasks spanning different physicochemical properties and chemical space distributions. Datasets were preprocessed as in [28], where invalid SMILES strings were filtered out by using RDKit [41], single-atom data without bond information were removed, entries with duplicated SMILES strings were dropped, and graphs were featurized with atom and bond features.

Table 1: Summary of the datasets used to evaluate the generalizability of ActiveFusion.

Dataset	Task Type	Target Property	#Mol.	#Scaffold
Permeability	Regression	Mycomembrane permeability	1,558	217
ESOL	Regression	Aqueous solubility (logS)	1,128	269
FreeSolv	Regression	Hydration free energy	642	63
Lipophilicity	Regression	Octanol–water distribution coefficient (logD)	4,200	2444

Data splitting

For each dataset, molecules were partitioned using a scaffold-based splitting strategy to evaluate model generalization under realistic distribution shifts. We use the same train-test splits generated by [28], where Bemis–Murcko [42] scaffolds were first computed for all molecules, and entire scaffold families were assigned exclusively to the held-out test set. The remaining molecules formed the active learning training pool and validation set.

Active learning initialization and iterations

Active learning experiments were designed to use approximately 20% of the available training data as the initial labeled set, with the remaining 80% forming the unlabeled pool. The unlabeled pool was sampled using a constant acquisition batch size over 10 active learning iterations. The batch size for each dataset was computed according to Eq. 1.

$$B = \left\lfloor \frac{N - N_{\text{in}}}{k} \right\rfloor \tag{1}$$

where N denotes the total number of available training molecules, N_{in} is the size of the initial labeled set (20% of N), and k is the number of active learning iterations. The floor function ensures that a constant integer batch size is used throughout the active learning process. Consequently, all available training molecules were utilized for the Permeability and Lipophilicity datasets, while only a small number of molecules remained unlabeled for ESOL (8 molecules) and FreeSolv (9 molecules) due to integer rounding of the batch size. The initial labeled set was selected by random sampling without replacement from the available training pool. Table 2 summarizes the active learning settings used in this study.

Table 2: **Summary active learning experimental settings across different datasets.** The active learning starts with the initial training set in the first iteration.

Dataset	Initial training set, N_{in}	Batch size	Iterations	Total acquired #Mol.	Validation #Mol.	Test #Mol.
Permeability	228	91	10	1138	284	136
ESOL	145	57	10	715	180	225
FreeSolv	75	29	10	365	140	128
Lipophilicity	538	215	10	2688	672	840

Molecular representations

SMILES strings were converted into different molecular representations. Table 3 shows the three categories of molecular feature representations we evaluated in this study: fixed representations, learned latent representations, and pretrained embeddings. The baseline molecular representations were selected to span the major categories of molecular representations commonly used in molecular property prediction. Specifically, RDKit descriptors represent expert-engineered physicochemical features, extended connectivity fingerprints (ECFP) fingerprints represent traditional substructure-based fingerprints, Chemprop provides graph neural network-derived molecular embeddings, MoLFormer and ChemBERTa represent Transformer-based sequence embeddings learned from SMILES, and UniMol provides three-dimensional geometry-aware molecular representations. This diverse collection of representation families enables a comprehensive evaluation of representation fusion across complementary sources of chemical information

Fixed feature representations

Physicochemical descriptors were computed using the RDKit cheminformatics toolkit [41], yielding 208 features. ECFP representations were generated with radius 2 and 1024 bits to encode local atomic environments and molecular topology [16].

Table 3: Summary of molecular feature representations evaluated in this study.

Representation	Category	Model Type	Input Modality	Dimension
RDKit Descriptors	Fixed	Physicochemical descriptors	2D graph	208
ECFP	Fixed	Circular fingerprint	2D graph	1,024
AttentiveFP	Learned	Graph neural network	Molecular graph	300
Chemprop (D-MPNN)	Learned	Graph neural network	Molecular graph	300
ChemBERTa	Pretrained (frozen)	SMILES-based Transformer	SMILES string	768
MoLFormer	Pretrained (frozen)	SMILES-based Transformer	SMILES string	768
UniMol2	Pretrained (frozen)	3D molecular Transformer	3D conformer	768

Learned latent representations

We evaluated learned latent molecular representations using GNNs, which are trained directly on the target property prediction task. These models learn task-relevant molecular features from graph-structured inputs in a supervised manner. We used two graph neural network architectures: AttentiveFP [43] and directed message passing neural networks (D-MPNN) [44], implemented in Chemprop [22]. These models operate on molecular graphs, where atoms are treated as nodes and bonds as edges, and learn relational chemical patterns through iterative message passing and attention mechanisms. The latent representations are obtained by average-pooling all the atom embeddings at the GNN's last layer.

Pretrained embedding representations

We evaluated pretrained molecular representations obtained from Transformer-based models trained on large-scale molecular datasets. These models are pretrained using self-supervised objectives from unlabeled data. Transformer-based embeddings capture global structural context and long-range dependencies that may not be explicitly encoded in traditional fingerprints or graph-based models [24, 25, 45].

In this work, we used MoLFormer-XL [25], ChemBERTa-3-MLM-100M [24], and UniMol2-164M [46] to generate molecular embeddings. We obtain embeddings by applying mean pooling over the last hidden states, where token embeddings are averaged using the attention mask to exclude padding tokens. We first used pretrained models as frozen encoders without fine-tuning on the downstream task. This setting allows us to evaluate the intrinsic quality of pretrained representations in active learning scenarios. MoLFormer and ChemBERTa were also fine-tuned in each active learning iteration to evaluate the effectiveness of adapted pretrained embeddings.

Representation fusion strategy

To capture complementary chemical information encoded by different molecular representations, ActiveFusion implements a feature fusion strategy. Learned latent representations or pretrained embeddings derived from neural networks (*e.g.*, GNNs or Transformer encoders) are integrated with fixed molecular representations such as RDKit descriptors and ECFP.

For each molecule i, let $\mathbf{h}_i \in \mathbb{R}^{d_1}$ denote the neural network representation and $\mathbf{v}_i \in \mathbb{R}^{d_2}$ denote the descriptor vector. Prior to fusion, each descriptor representation is independently standardized using Min-max normalization to ensure comparable scaling across each feature dimension. The resulting descriptor vectors were then concatenated with the corresponding learned molecular representations, after which PCA was applied. Feature fusion is performed through direct vector concatenation as shown in Equation 2. The scaling parameters were estimated exclusively from the training partition, and the fitted transformation was

subsequently applied to the held-out test set. Thus, no information from the test set was used during feature preprocessing.

$$\mathbf{z}_i = \text{concat}\,(\mathbf{h}_i,\ \mathbf{v}_i) \in \mathbb{R}^{d_1+d_2} \tag{2}$$

This fusion strategy preserves the full informational content of both representation families while enabling the predictive model to jointly exploit global physicochemical properties and relational structural patterns captured by learned embeddings.

Importantly, learned representations are updated throughout the active learning process as the labeled training set expanded. Consequently, fused feature vectors evolve dynamically across acquisition iterations, allowing the model to progressively refine its representation of chemical space.

ActiveFusion framework

We consider a regression task where each molecule has a SMILES string and a continuous target. The dataset is partitioned into: (i) a training pool from which labels are acquired during active learning, (ii) a validation set used for model selection and monitoring, and (iii) a scaffold-based held-out test set used for reporting generalization and retrieval metrics (Figure 1)

Algorithm 1 describes our ActiveFusion workflow and active learning iterations. At iteration t, a surrogate model f_t is trained using only the currently labeled subset of the pool, represented by either single or fused representations. The trained model is inferred on the remaining unlabeled pool to get predictions and variance, and an acquisition function selects a new batch of unlabeled molecules to label.

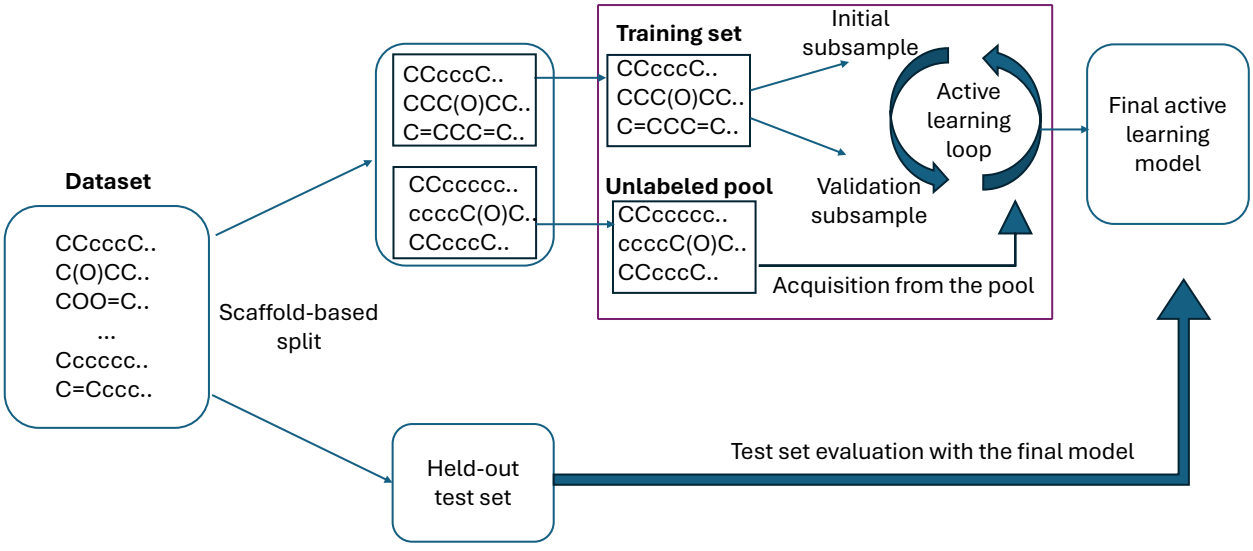

Figure 1: **Active learning workflow.**

Representation learning and retraining protocol. For experiments involving learned molecular embeddings, GNNs were retrained from scratch at every active learning iteration using only the currently labeled subset of the training pool. After training, updated molecular embeddings were recomputed for the entire pool, validation set, and test set before surrogate model training. This procedure ensured that rep-

Algorithm 1 ActiveFusion Active Learning Pipeline

Initialize labeled set $\mathcal{L}_0$ with N_{in} molecules from training pool $\mathcal{U}$
Set unlabeled pool $\mathcal{U}_0 = \mathcal{U} \setminus \mathcal{L}_0$
for $t = 0, 1, \ldots, T-1$ **do**
    Extract molecular representations for $\mathcal{L}_t$ and $\mathcal{U}_t$ using one or more featurization methods
    If multiple representations are used, apply feature-level fusion (e.g., concatenation after normalization)
    Train surrogate model f_t on fused (or single) representations of $\mathcal{L}_t$
    Use trained model f_t to infer on the unlabeled $\mathcal{U}_t$, and compute acquisition scores for all $x \in \mathcal{U}_t$
    Select batch $\mathcal{B}_t$ of size k based on the acquisition function
    Query labels y for $\mathcal{B}_t$
    Update labeled set: $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \mathcal{B}_t$
    Update unlabeled set: $\mathcal{U}_{t+1} = \mathcal{U}_t \setminus \mathcal{B}_t$
end for
Evaluate performance on held-out test set

resentation learning remained consistent with the evolving labeled dataset and avoided information leakage from unlabeled compounds.

Surrogate modeling. A Gaussian process regressor[47] with radial basis function (RBF) kernel and additive white noise term was employed as the surrogate model. In addition to the established success of building Gaussian processes on embeddings seen in past works [48–50], the choice of this surrogate model stems from its Bayesian formulation, which naturally provides both a predictive mean and an associated uncertainty estimate through the posterior variance which is convenient for active learning setups. Prior to surrogate training, dimensionality reduction was performed using principal component analysis fitted exclusively on the labeled training subset at each iteration, and after feature fusion was applied. This was implemented to reduce the feature space to $min(D, 50)$ components, where D is the original dimensionality.

Acquisition and dataset update. At each iteration, acquisition scores were computed over all unlabeled molecules using the specified acquisition function (see Acquisition functions. A fixed batch of top-ranked compounds was then selected and added to the labeled set. The unlabeled pool was correspondingly updated by removing the acquired molecules.

Baseline comparison. To contextualize active learning performance, a baseline model was trained using the full training dataset without active learning. This baseline represents the upper bound achievable when all labeled data are available and no active learning strategy is employed.

Evaluation protocol. Model performance was evaluated after every active learning iteration using both prediction metrics (R^2, RMSE) and discovery-oriented metrics (test retrieval rate, pool-based top retrieval rate, and scaffold retrieval). Prediction metrics were computed on the held-out scaffold test set to quantify generalization to structurally novel compounds. Retrieval metrics quantified the ability of acquisition strategies to prioritize high-performing molecules, while scaffold retrieval quantified chemical space exploration across iterations. Experimental settings and hyperparameters are detailed in Experimental setup and hyperparameters in the supporting information.

Acquisition functions

Each acquisition function defines a selection strategy that is evaluated across active learning iterations against the proposed performance and retrieval metrics. We considered five acquisition strategies representing different exploration–exploitation trade-offs, the balance between selecting compounds with favorable predicted values for the label of interest (exploitation) and those that are uncertain or underrepresented in the current training set (exploration) [3, 51]. These acquisition strategies are as follows: greedy acquisition, expected improvement (EI), entropy sampling, diversity sampling, and random acquisition.

For the exploitation-driven acquisition functions, we use greedy acquisition and EI. **Greedy acquisition** represents a purely exploitation-driven strategy, selecting compounds solely based on the surrogate model's predictive mean. In this case, candidates with the most favorable predicted values are prioritized, without consideration of model uncertainty as shown in Equation 3.

$$\mathcal{B}^* = \arg\operatorname{top}_k\{f(x) : x \in \mathcal{U}\} \tag{3}$$

Where $\mathcal{B}^*$ indicates the selected top-k compounds from the unlabeled pool. $f(x)$ denotes the predicted score given by the surrogate model (e.g., $f(x) = \mu(x)$ for maximization tasks and $f(x) = -\mu(x)$ for minimization tasks, where $\mu(x)$ is model's mean predicted score for unlabeled candidate x).

EI was designed to balance exploitation and exploration by selecting compounds that were expected to improve upon the current best observed target value while accounting for predictive uncertainty. Specifically, EI prioritizes candidates with both favorable predicted property values relative to the current optimum and high uncertainty, thereby promoting efficient discovery of high-value compounds while still exploring uncertain regions of chemical space (Equations 4 - 6).

$$\mathrm{EI}(x) = (\mu(x) - f^*)\Phi(z) + \sigma(x)\phi(z), \tag{4}$$

$$z = \frac{\mu(x) - f^*}{\sigma(x)}, \tag{5}$$

$$\mathcal{B}^* = \arg\operatorname{top}_k\{\mathrm{EI}(x) : x \in \mathcal{U}\} \tag{6}$$

Here, $\mu(x)$ and $\sigma(x)$ denote the predicted mean and standard deviation of the model for unlabeled candidate x, respectively. f^* represents the best observed target value in the current labeled set. $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function (CDF) and probability density function (PDF) of the standard normal distribution, z is the standardized improvement term. In practice, $\mu(x)$ and $\sigma(x)$ can be obtained from a probabilistic surrogate model that provides both predictions and uncertainty estimates, such as Gaussian processes [52] in this study, or neural networks with uncertainty estimation by ensemble variance [53] or Monte Carlo dropout [54].

For the exploration-driven acquisition functions, we used entropy-based sampling and diversity sampling. **Entropy-based sampling** is an exploration-driven strategy that prioritizes compounds with high predicted uncertainty, helping improve the model's knowledge base in finding compounds. This strategy utilizes the uncertainty from the surrogate model's prediction to rank the next batch of compounds to label. Under a Gaussian predictive distribution, the entropy of a candidate x is given by Equation 7[55]:

$$H(x) = \frac{1}{2}\log\left(2\pi e\,\sigma^2(x)\right) \tag{7}$$

We then select the top-k compounds with the highest entropy in Equation 8:

$$\mathcal{B}^* = \arg\operatorname{top}_k \{H(x) : x \in \mathcal{U}\} \tag{8}$$

Diversity sampling is also a purely exploration-driven strategy that prioritizes compounds that are most dissimilar to the labeled compounds. For each compound x in the unlabeled pool $\mathcal{U}$, we compute its maximum Tanimoto similarity [56, 57] to any compound x' in the labeled set $\mathcal{L}$. Compounds with smaller maximum similarity values are more novel with respect to the labeled set, since even their most similar labeled neighbor is relatively dissimilar. We therefore select the k compounds with the smallest maximum similarity values, favoring candidates that are dissimilar to all labeled compounds (Equation 9)

$$\mathcal{B}^* = \arg\operatorname{bottom}_k \left\{ \max_{x' \in \mathcal{L}} s(x, x') : x \in \mathcal{U} \right\} \tag{9}$$

Here, $s(x, x')$ denotes the Tanimoto similarity between compounds x and x'. Notably, diversity sampling is model-agnostic, as it does not rely on the surrogate model for ranking.

Random acquisition selects compounds uniformly at random from the unlabeled pool and serves as a pure exploration baseline (Equation 10). Although it does not leverage model predictions or uncertainty, it provides a reference for quantifying the benefit of informed acquisition strategies.

$$\mathcal{B}^* \sim \text{Uniform}(\mathcal{U}),\ |\mathcal{B}^*| = k \tag{10}$$

Evaluation metrics

Importantly, no single metric fully captures active learning performance. We analyze prediction accuracy, retrieval efficiency, and scaffold diversity, to obtain a comprehensive view of active learning behavior across exploration and exploitation dimensions.

Prediction performance metrics

We evaluate model performance with coefficient of determination (R^2) and root mean squared error (RMSE). We evaluate the active learning cycle by counting the number of iterations to meet the baseline, as well as the maximum R^2 attained, and also the maximum R^2 attained in all the active learning cycles, for all the datasets.

Exploitation metrics

To evaluate exploitation of our active learning setup, we implement the **test retrieval rate** and the **pool-based top retrieval rate** metrics. (Table 4)

The **test retrieval rate** measures how many high-scoring compounds are identified in the top-ranked subset of a held-out test set across iterations, *i.e.*, top-k percent of the y labels, based on the surrogate model's predictions. It evaluates the model's ability to prioritize truly valuable compounds in unseen data, based on ranking. This metric is standard in the field across various discovery setups.

Across acquisition iterations, the **pool-based top retrieval rate** measures the fraction of globally top-ranked compounds from the original candidate pool that have been acquired by the current iteration. In our implementation, this global top set is fixed before active learning begins and contains the top n_{batch}

compounds (see batch size in Table 2) ranked by the ground-truth pool labels. High pool-based top retrieval indicates that the acquisition function favors regions of highest-scoring compounds.

Exploration metrics

We implemented **scaffold retrieval rate** to evaluate the amount of exploration performed by our active learning models (Table 4). Across acquisition iterations, scaffold retrieval rate measures the fraction of unique Bemis–Murcko scaffolds [42] in the entire training pool that have been covered by the cumulative labeled set up to the current iteration. It quantifies chemical diversity in the acquired set. High scaffold retrieval indicates that the acquisition strategy samples broadly across chemical space rather than repeatedly selecting structurally similar compounds.

Table 4: Summary of evaluation metrics used to assess prediction accuracy, exploration, and exploitation in active learning.

Metric	Category	Definition	Purpose
R^2	Predictive	$1 - \frac{\sum_{x\in\mathcal{T}}(y-\hat{y})^2}{\sum_{x\in\mathcal{T}}(y-\bar{y})^2}$	Overall prediction accuracy
RMSE	Predictive	$\sqrt{\frac{1}{\lvert\mathcal{T}\rvert}\sum_{x\in\mathcal{T}}(y-\hat{y})^2}$	Prediction error magnitude
Scaffold Retrieval Rate	Exploration	$\frac{\lvert\{S(x) : x \in \mathcal{L}_t\}\rvert}{\lvert\{S(x) : x \in \mathcal{U}\}\rvert}$	Chemical space coverage
Test Set Retrieval Rate	Exploitation	$\frac{\lvert TopK^{\text{test,pred}} \cap TopK^{\text{test,true}}\rvert}{\lvert TopK^{\text{test,true}}\rvert}$	Ranking quality on unseen compounds
Pool Top Retrieval Rate	Exploitation	$\frac{\lvert\mathcal{L}_t \cap TopK^{\text{pool,true}}\rvert}{\lvert TopK^{\text{pool,true}}\rvert}$	Prioritization of high-value pool compounds

Notation.
x denotes a molecular compound.
y and $\hat{y}$ denote the true and predicted property values.
$\bar{y}$ denotes the mean true property value in the test set.
$\mathcal{T}$ denotes the held-out test set.
$S(x)$ denotes the Bemis–Murcko scaffold.
$\mathcal{L}_t$ denotes the cumulative labeled set up to acquisition iteration t.
$\mathcal{U}$ denotes the entire training pool.
$TopK^{\text{test,true}}$ denotes the ground truth of top k% compounds in the test set.
$TopK^{\text{test,pred}}$ denotes the predictions of top k% compounds in the test set at an iteration.
$TopK^{\text{pool,true}}$ denotes the ground truth of top k% compounds in the entire unlabeled pool.

Chemical Interpretability Analysis of ActiveFusion Representation

To evaluate whether the PCA-compressed molecular representation retained chemically meaningful information, we interpreted the Gaussian process model trained on the 50-dimensional PCA representation. SHAP (SHapley Additive exPlanations) values were computed for the principal components to quantify their contribution to permeability predictions. For each principal component, the mean absolute SHAP value was calculated across the evaluation molecules, and the components were ranked according to their importance to the prediction of permeability. The highest-ranking principal components were subsequently correlated with a set of 23 interpretable physicochemical properties known to be predictive of mycomembrane permeability [28, 58] using Spearman correlation analysis.

Chemical space coverage analysis

To quantify the structural diversity of each dataset and characterize the extent of chemical space explored during active learning, we performed a chemical space coverage analysis based on molecular similarity and scaffold diversity metrics. Molecular structures were represented using extended connectivity fingerprints (ECFP, radius = 2, 1024-bits), which provide a topology-aware encoding of local chemical environments. Pairwise Tanimoto similarity between compounds was used to assess structural redundancy and diversity within each dataset.

To visualize global chemical space organization, we employed Uniform Manifold Approximation and Projection (UMAP)[59] on fingerprint representations. UMAP projections enabled qualitative assessment of how acquisition strategies navigated chemical space and whether selected compounds populated dense regions or sparsely sampled boundaries.

Results and Discussion

Representation performance and suitability for low-data active learning

We compare performance of all acquisition functions across all datasets, models, and performance metrics. To simplify downstream interpretation, we select the strongest representative from each representation family based on the the full-training baseline performance (Figure 2), which is generally close to peak R^2 attained during active learning (Figure 4, Figures S20 - S27). The selected representatives are Chemprop for GNN-based representations, MoLFormer for Transformer-based representations, and RDKit for the fixed representations.

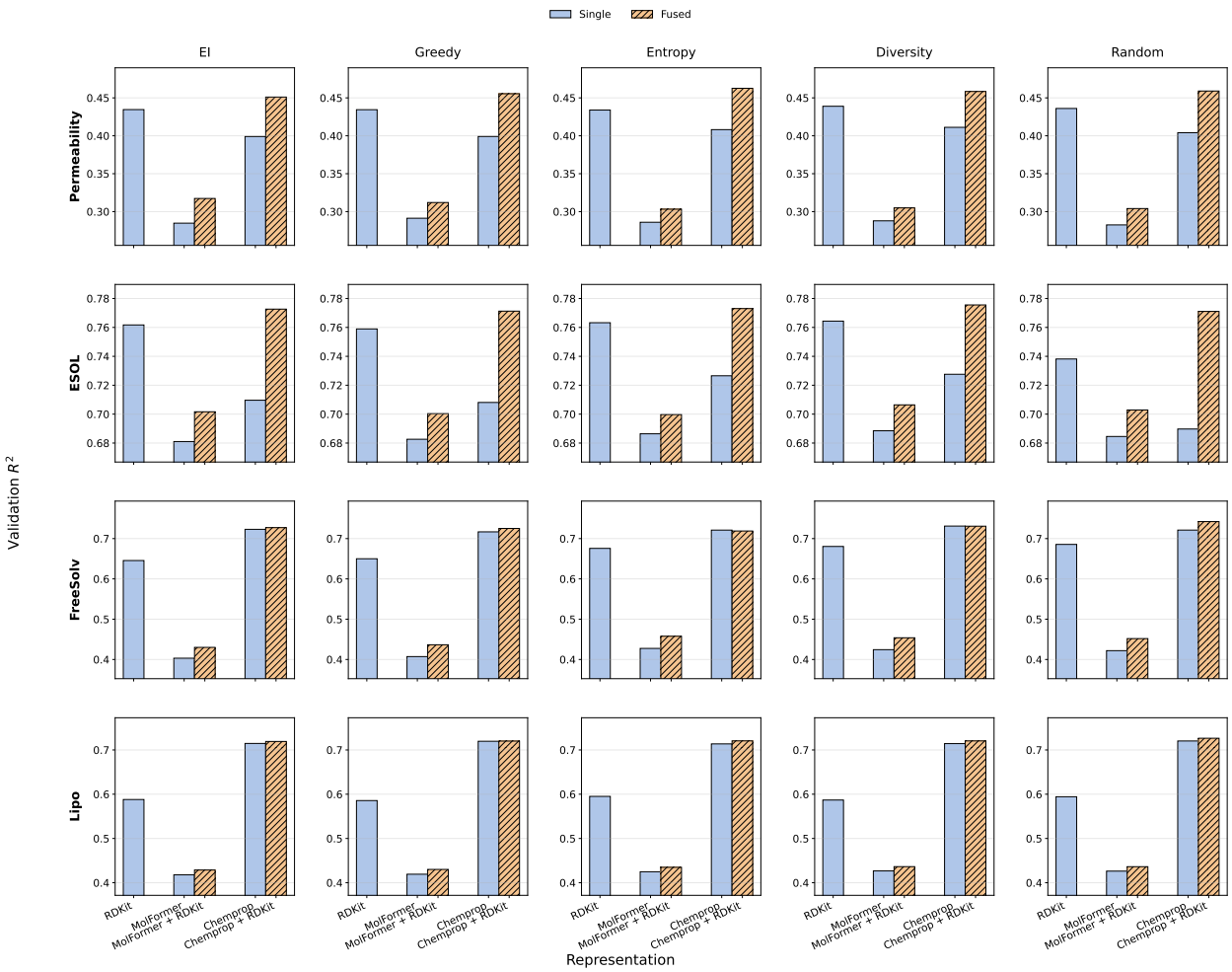

Figure 2: **Validation R^2 performance across datasets and acquisition functions.** R^2 on validation set shows the result of training on the entire training dataset without active learning.

GNNs outperform Transformers across tasks

We report an average baseline R^2 of 0.67 for Chemprop, our highest-performing GNN model, against 0.38 for MoLFormer, our highest-performing Transformer model ($p < 1 \times 10^{-19}$, Mann-Whitney U) as shown in Table S1. We also report maximum R^2 attained across all active learning iterations, averaged across all the datasets and acquisition functions in Table 5, where we see significant performance of Chemprop over MoLFormer (R^2 of 0.68 vs. 0.39, $p < 1 \times 10^{-43}$, Mann-Whitney U). We report the results for all the independent acquisition functions and for all the independent datasets in Tables S3 - S6. Other GNN architectures also outperform Transformer architectures in terms of maximum R^2 (Table S2). The comparatively lower performance of Transformer-based representations may be related to the mismatch between their pretrained knowledge and the task-specific objective. While these models are trained on large molecular corpora using self-supervised objectives, the resulting representations may not be well aligned with the downstream property prediction tasks considered in this study. These findings suggest that Transformer-based representations generated by frozen pretrained models may not consistently provide advantages in low-data active learning scenarios.

To evaluate whether finetuning can help the pretrained models adapt to the downstream tasks, we select

the best perfoming transformer models based on baseline model prediction performance on the validation set (Figure 2) – MoLFormer and ChemBERTa – to conduct supervised finetuning in each iteration of active learning (see Experimental setup and hyperparameters in the supporting information). Figure 3 shows that finetuning improves the performance of the pretrained models over their frozen states, but they are still inferior to Chemprop. This is likely because finetuning large pretrained models requires a sufficient amount of task-specific data. The used dataset size is small, and the data volume in each active learning iteration is also limited, providing insufficient signal to fully adapt their general-purpose representations to the downstream property prediction objective.

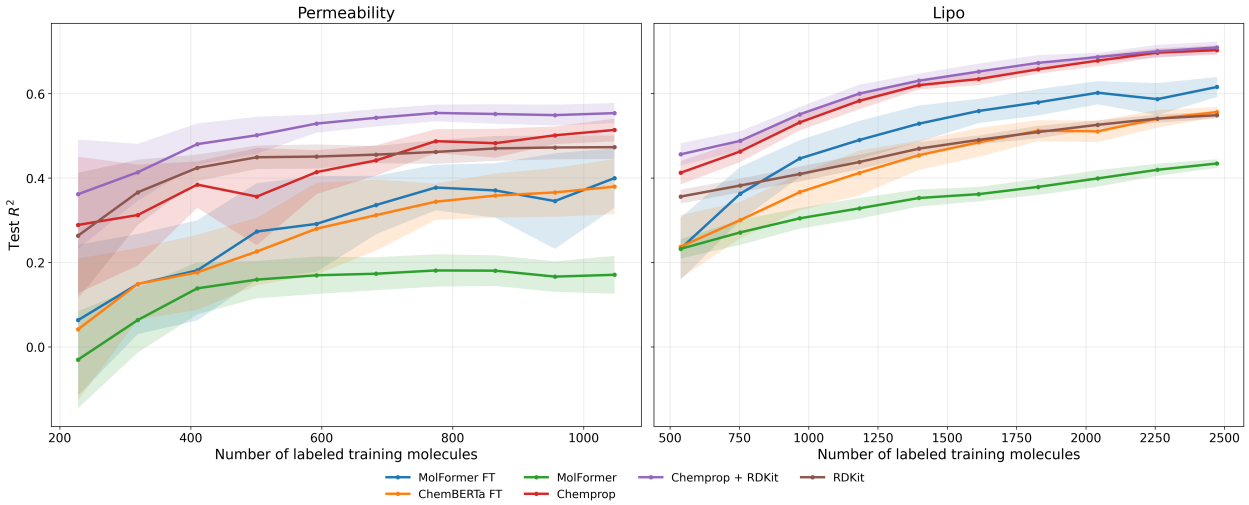

Figure 3: **Evaluation of best molecular representations with finetuned transformers.** We show results for the Permeability and the Lipophilicity datasets. FT stands for 'finetuned' and the acquisition function shown here is the expected improvement.

Table 5: **Comparison of average maximum attained R^2_{test} values across critical molecular representation pairs. P-values are computed by using the Mann–Whitney U test.**

Representation Pairs	$\overline{\text{Max}_{\text{repr1}}}^{a}$	$\overline{\text{Max}_{\text{repr2}}}$	p-value
MoLFormer vs. Chemprop	0.39	**0.68**	4.23×10^{-43}
MoLFormer+RDKit vs. Chemprop+RDKit	0.41	**0.71**	1.35×10^{-40}
RDKit vs. Chemprop	0.55	**0.68**	5.46×10^{-12}
RDKit vs. Chemprop+RDKit	0.55	**0.71**	3.97×10^{-19}
MoLFormer vs. MolFormer+RDKit	0.39	**0.41**	2.23×10^{-4}
Chemprop vs. Chemprop+RDKit	0.68	**0.71**	1.57×10^{-5}

[a] $\overline{\text{Max}_{\text{repr}}}$ denotes the average maximum attained R^2_{test} for a representation. For each run, the maximum R^2_{test} was first taken across active learning iterations, and the resulting values were then averaged across acquisition strategies, datasets, and random seeds.

Descriptor representations provide strong standalone performance

We observe strong standalone performance fixed descriptor-based representations, which in general meet or exceed the performance of Transformer-based models in terms of R^2 but lag behind our best GNN model

(Figures 2 and 4). We also note from Figure 4 that fixed descriptors yield more stable performance trajectories across iterations relative to the learned representations, particularly in early iterations. This indicates that fixed descriptors may be more suitable for uncertainty estimation and compound ranking early in active learning workflows.

Representation fusion improves performance across datasets

We find that combining GNN-learned and Transformer-based embeddings with physicochemical descriptors generally improves prediction performance across active learning iterations. In particular, the fused Chemprop+RDKit representation emerges as the most effective feature combination for three out of the four datasets, with a mean final iteration R^2 of 0.71 for the fused representation, followed by Chemprop (0.67), and RDKit (0.54) (Table S1) across all the datasets, with the finegrained results reported in Tables S3 - S6. The exception is the ESOL dataset, where maximum R^2 performance attained of the fused representation is slightly worse than RDKit standalone representation (0.82 vs. 0.83). For this dataset, since both RDKit and Chemprop exhibit strong performance in maximum R^2 attained (0.83 and 0.76 respectively, Table S4), we suspect that their fusion do not provide new modeling information to significantly improve the active learning process. However, we do note that fusion appears to smooth the more jagged curves produced by the learned representations. For FreeSolv, with a weak RDKit representation (0.34) compared to a strong Chemprop representation (0.74), the fusion benefit is slight (0.76) but still present.

We compare fusion performance when using ECFP representations instead of RDKit, and observe similar performance trends (Figure S2). Notably, this phenomenon still holds even though ECFP has weaker standalone baseline R^2 performance than RDKit, indicating that the performance gains are not specific to a single descriptor family. These improvements generalize across all benchmark datasets, suggesting that ActiveFusion effectively leverages complementary structural and physicochemical information during iterative model refinement.

To evaluate the effect of PCA on surrogate modeling, we repeated the representation fusion experiments on the permeability dataset without dimensionality reduction, allowing the Gaussian process to operate directly on the original high-dimensional fused representations. The same overall conclusions were observed, with fused representations consistently outperforming their corresponding single-representation baselines (Figure S7). These results indicate that the performance gains from representation fusion are robust to the presence or absence of PCA. We therefore retained PCA in the proposed framework because it substantially reduces the computational cost and memory requirements associated with Gaussian process training on high-dimensional fused representations, while preserving the overall performance trends observed across datasets.

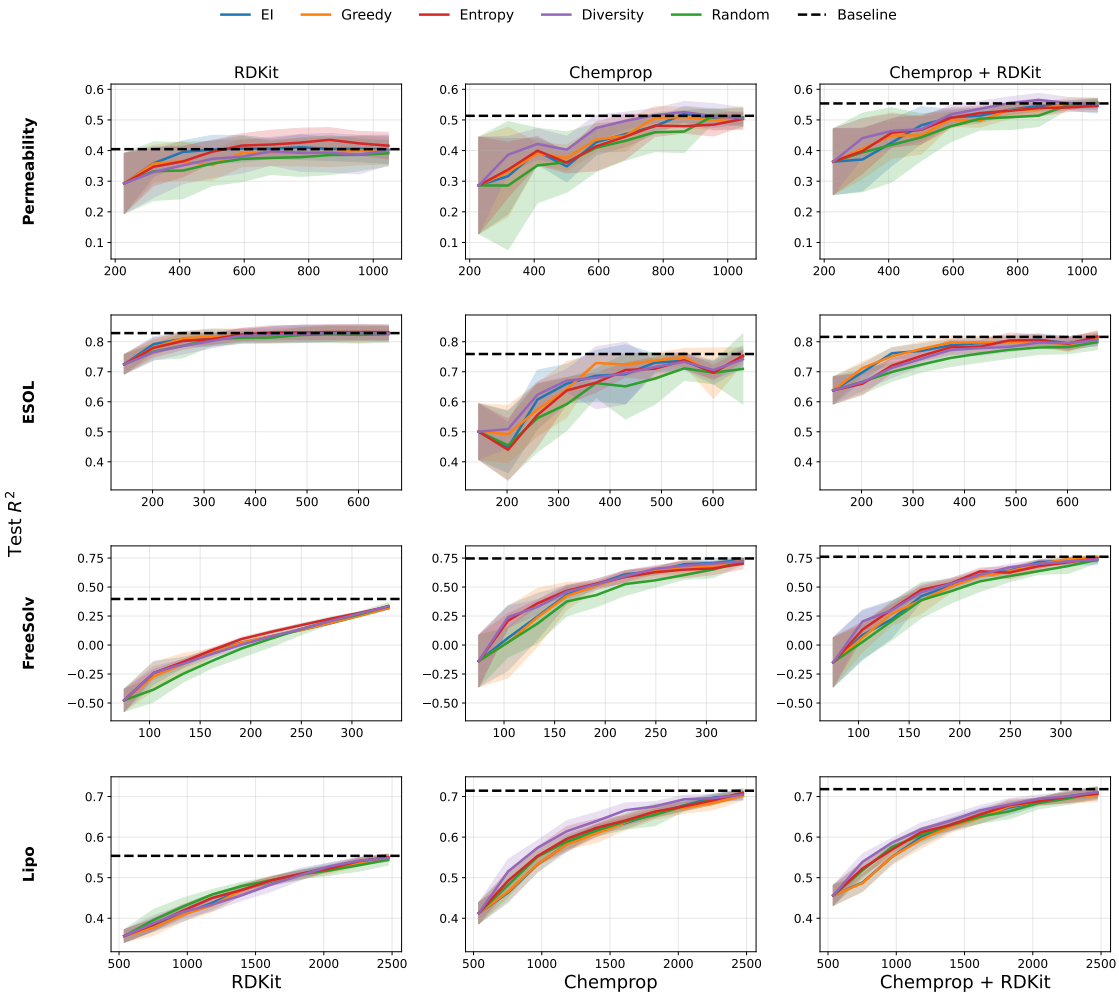

Figure 4: R^2 **test set performance across active learning cycles for all the datasets for the top 3 best performing representations.** The dashed line represents the model baseline which is training the entire dataset without active learning. Consistent fusion improvement across datasets shows generalizability.

Exploration performance of ActiveFusion

Exploration-driven acquisition strategies outperform random acquisition at scaffold retrieval across all datasets, as depicted in Figure 5. At 50% of the labeled pool acquired, diversity-based acquisition beats random acquisition at scaffold retrieval in 12/12 cases, and entropy-based acquisition beats random acquisition in 11/12 cases. Diversity-based acquisition beats entropy-based acquisition in 11/12 cases at 50% of the pool acquired. Both diversity-based sampling and entropy-based uncertainty sampling lead to improved coverage of unique chemical scaffolds during active learning cycles. These findings indicate that exploration strategies enable broader traversal of chemical space compared to uninformed sampling.

We observe that the model-agnostic diversity acquisition function retrieves structurally unique compounds earlier in the active learning process than the model-aware entropy acquisition strategy. We observe this for all scenarios except for the RDKit and Chemprop+RDKit representations on the ESOL dataset where entropy reaches complete retrieval of all the scaffolds earlier compared to the diversity acquisition function, however this is not statistically significant due to overlapping confidence intervals.

We investigate the apparent underperformance of the entropy acquisition function that uses the estimate of the Gaussian process surrogate model uncertainty. To achieve this, we measured the Spearman rank correlations between the mean absolute error and the Gaussian process uncertainty scores across the active learning iterations (Figure S4). Overall, we found that absolute error only weakly correlated with uncertainty. Most notably, we found that the correlation was near zero early in the active learning pipeline. This indicates that surrogate model uncertainty estimates are poorly calibrated, not accurate estimates of actual uncertainty, potentially explaining their underperformance compared to the diversity acquisition function. Given the strong performance of the diversity acquisition function, we recommend implementing diversity acquisition early in the pipeline if poor calibration is observed.

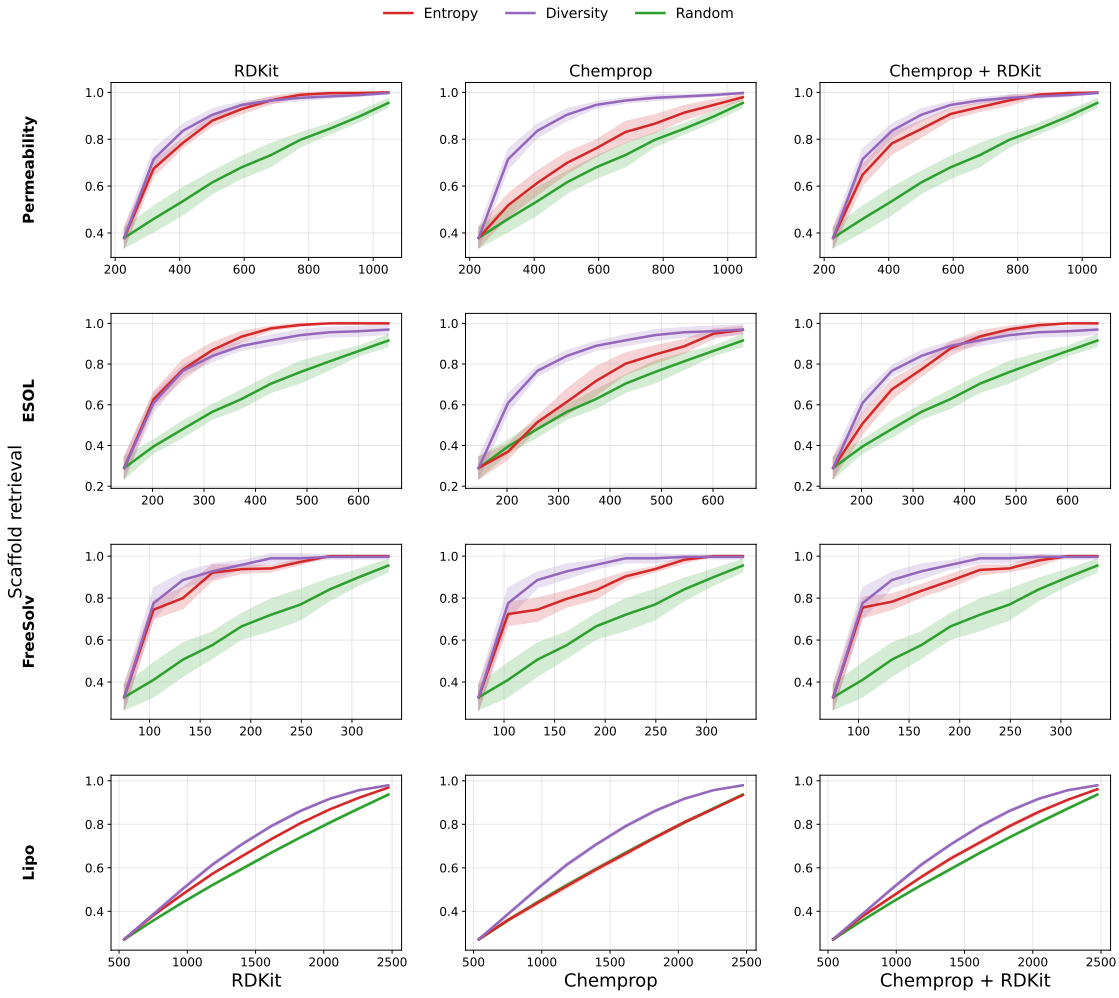

Figure 5: **Scaffold retrieval performance across active learning cycles for the all the datasets for the top 3 best performing representations.** Fusion facilitates scaffold retrieval. Shown here are all the exploration-based acquisition functions. Extended results for other representations can be found in Figures S8 - S11.

Using entropy-based acquisition, we find that per-iteration gains in scaffold retrieval are positively correlated with per-iteration gains in prediction performance (Figure 6). We see this generalize across all datasets, with significant positive Spearman correlations across 10 random states for each iteration - For the Chemprop+RDKit representation we find spearman correlations of 0.33 for Permeability, 0.65 for ESOL,

0.75 for FreeSolv, and 0.90 for Lipophilicity. We also find similar conclusions using both the RDKit and Chemprop representations individually (Figures S5 and S6). This relationship shows that iterations where more novel scaffolds are acquired tend to be the iterations with the greatest performance gains, highlighting the importance of exposing the active learning model to novel chemical knowledge through structurally diverse compounds.

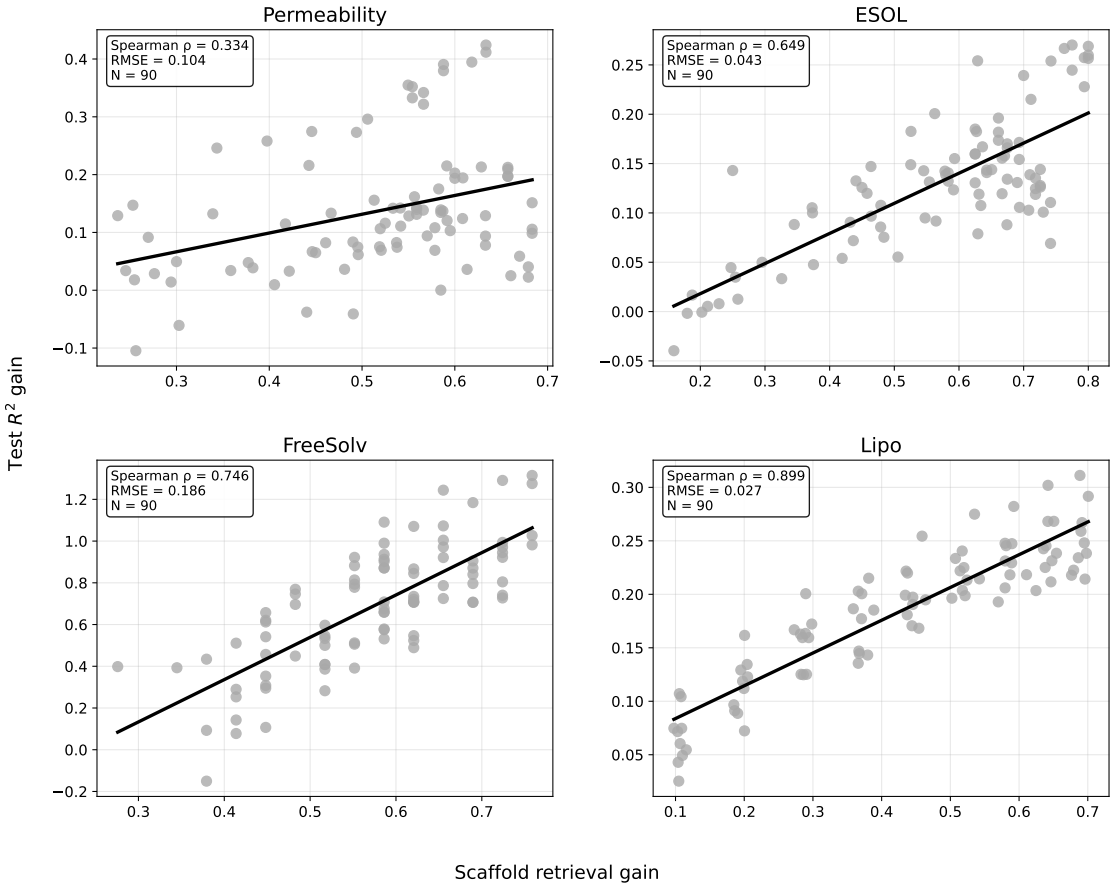

Figure 6: **Unique scaffold gain across the active learning cycles is positively correlated with R^2 improvement across all datasets.** Here we used the entropy acquisition function to measure if the retrieval of unique scaffolds improve active learning model performance. Each dot in the plot represent an active learning cycle for a particular model random state. The axes represent changes in each from the prior iteration. The feature representation used here is Chemprop+RDKit. Figures S5 and S6 show the same conclusion for the single feature representations - RDKit and Chemprop.

Exploitation performance of ActiveFusion

Exploitation describes the ability of an active learning strategy to efficiently identify high-value compounds during iterative data acquisition. We assess exploitation performance using retrieval-based metrics that measure how effectively the framework prioritizes compounds with desirable target properties.

For fused representations, exploitation-driven acquisition functions (EI and greedy acquisition) consistently outperform random acquisition at both retrieval metrics across all datasets (Figure 7 and 8). Exploitation-driven acquisition functions applied to individual representations also outperform random ac-

quisition for all datasets. In general, increased performance over random acquisition is expected, as these strategies leverage surrogate model predictions to focus sampling on regions associated with favorable molecular properties. The improved retrieval efficiency demonstrates the suitability of ActiveFusion for compound prioritization tasks in low-data discovery settings.

As shown in Figure 7, at first iteration, the performance of Chemprop is worse for permeability, FreeSolv and ESOL datasets when compared to RDKit. Fusion is more stable than the Chemprop standalone representation, and more exploitative in early retrieval than the RDKit standalone representation. In all these cases, fusion of these representations improves retrieval performance to compensate for one of the representations being poor. For pool retrieval analysis (Figure 8), all 12/12 cases of both exploitative acquisition functions show stong retrieval performance over random acquisition, possessing more area-under-the-curve. This further proves suitability for our ActiveFusion framework in high-value prioritization active learning setups.

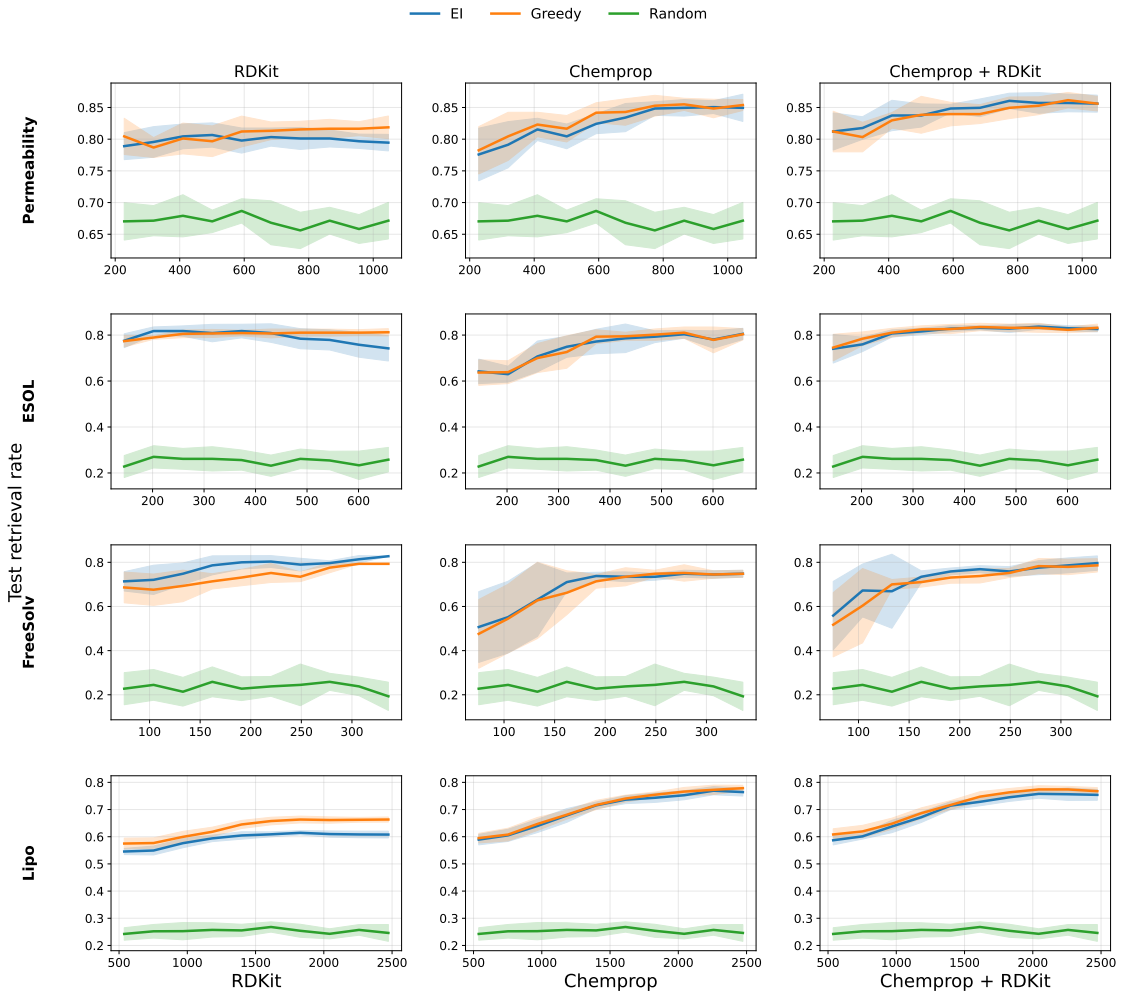

Figure 7: **Test set retrieval performance across active learning cycles for the all the datasets for the top 3 best performing representations.** Shown here are all the exploitation-based acquisition functions. Extended results for other representations can be found in Figures S12 - S15.

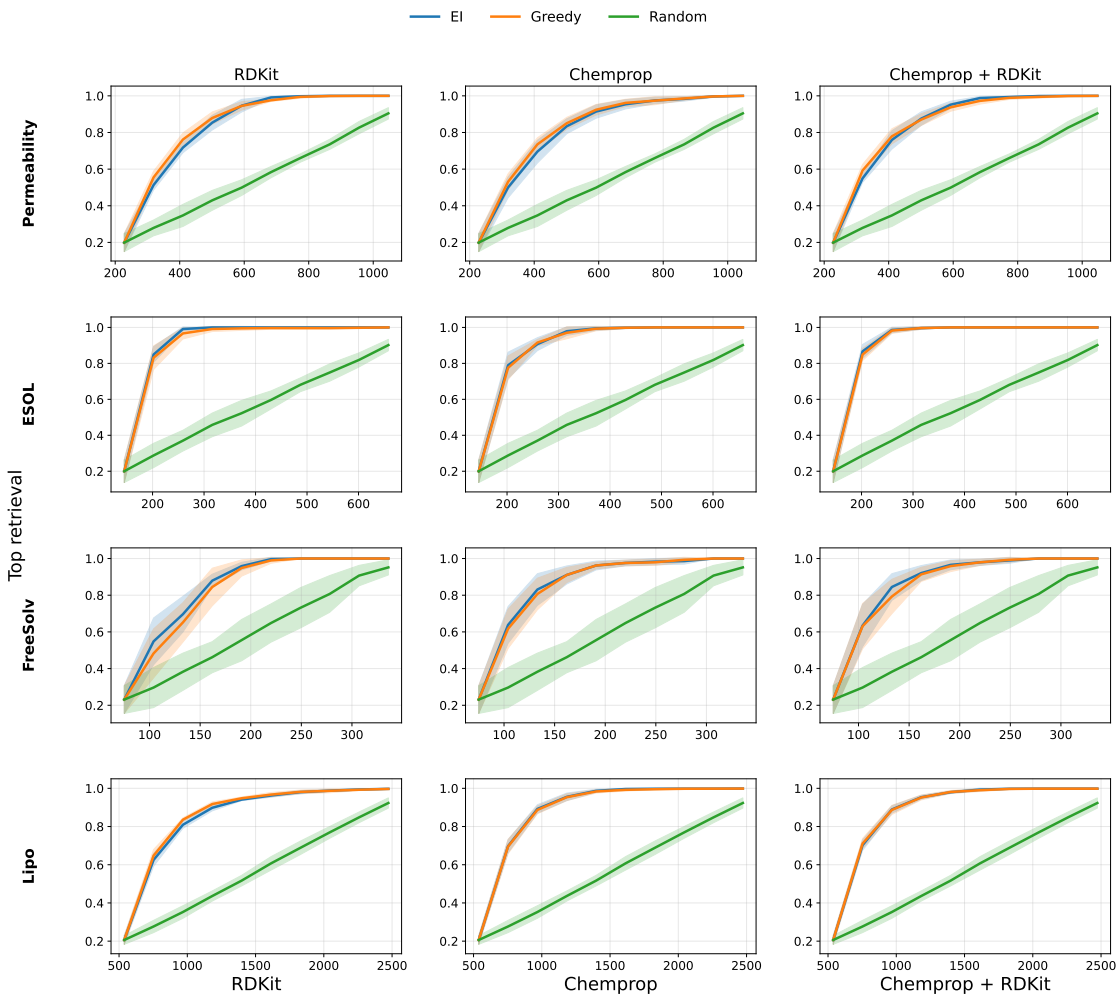

Figure 8: **Pool retrieval of the top high-performing compounds across active learning cycles for the all the datasets for the top 3 best performing representations.** Shown here are all the exploitation-based acquisition functions. Extended results for other representations can be found in Figures S16 - **??**

Chemical Interpretability of PCA-Compressed Representations

To understand the chemical interpretability of the fused representation, we performed feature importance analysis using SHAP on the 50 principal components (PCs) obtained after PCA using the permeability dataset. SHAP analysis identified PC10, PC49, and PC3 as the most influential components contributing to model predictions.(Figure 9a). We subsequently used spearman correlation to relate the values of the learned PCs to interpretable physicochemical descriptors. (Figure 9b and Table 6)

PC10 exhibited the strongest associations with `NOCount` and TPSA, indicating that nitrogen-containing functionalities and molecular polarity are major contributors to permeability prediction. These observations are consistent with previous studies reporting the importance of nitrogen heterocycles and polar surface area in governing mycomembrane permeability [28, 36]. LogP of small molecules has also been previously reported to correlate with mycomembrane permeability [60], and we observe logP being strongly correlated with two of the most influential principal components, PC10 and PC6. (Figure 9b and Table 6)

Although PC49 was identified as an influential latent component by SHAP, it did not exhibit strong correlations with any individual physicochemical descriptor. This suggests that PC49 primarily captures information originating from the learned molecular embedding, or interactions between the embedding and descriptor spaces, that cannot be explained by any single handcrafted descriptor alone. Our previous work has also suggested this possibility. [28]

PC3 was strongly associated with molecular weight and heavy atom count, indicating that molecular size is another important determinant of mycomembrane permeability. This observation is also consistent with previous studies showing that larger molecules generally experience greater difficulty permeating the mycomembrane than smaller compounds. [58, 60, 61]

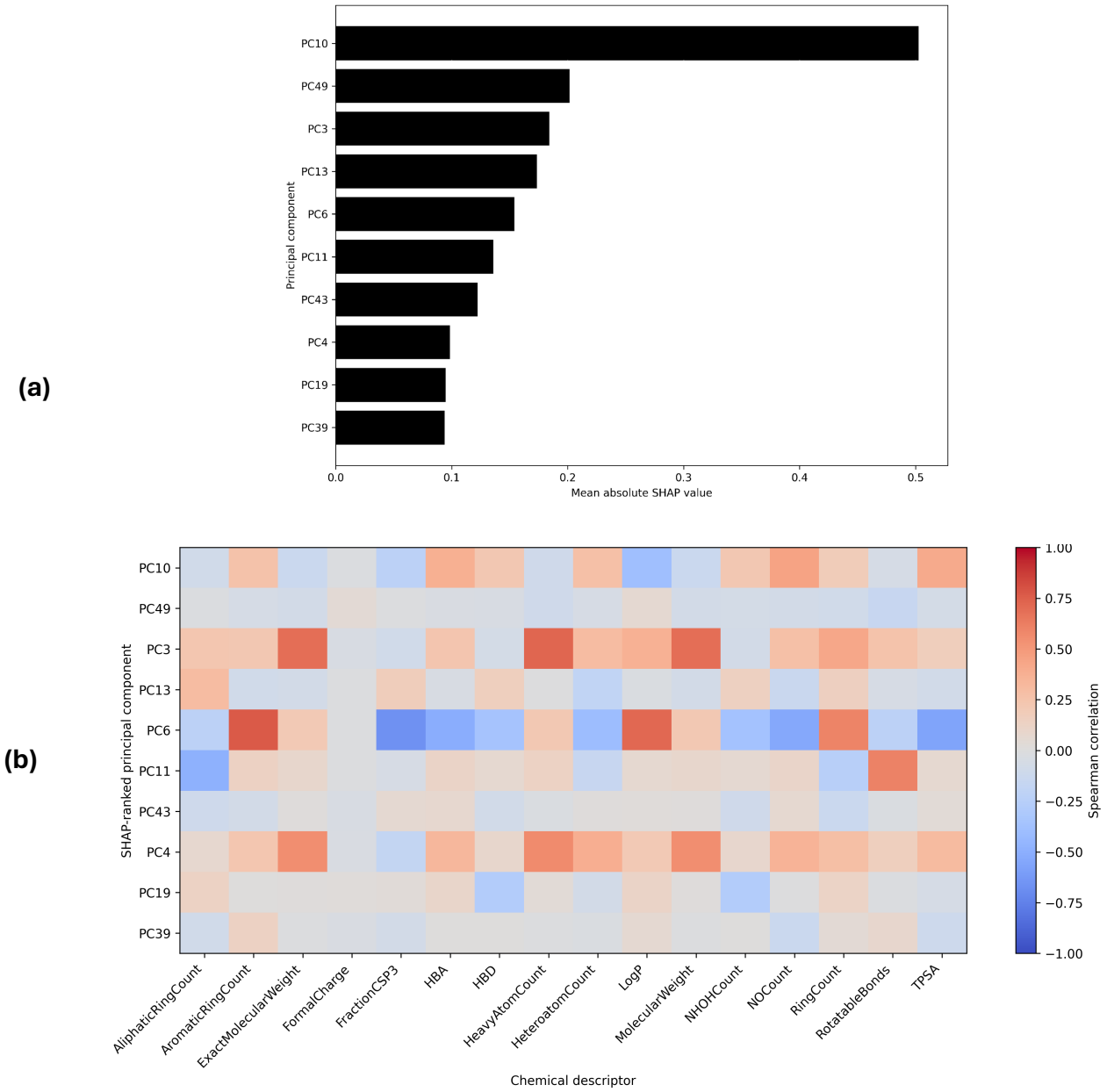

Figure 9: **PCA chemical interpretability analyses. (a) Feature importance ranking of the principal components on mycomembrane permeability (b) Heatmap correlation between top most influential principal components with physicochemical descriptors**

Table 6: Most influential principal components identified by SHAP analysis and their dominant associated molecular descriptors. Descriptor associations are determined using the magnitude of correlation between each principal component and standard molecular descriptors.

Principal Component	**Mean |SHAP|**	**Top 3 Descriptor Associations**
PC10	0.503	NOCount, TPSA, logP
PC49	0.202	No strong association with measured descriptors
PC3	0.184	HeavyAtomCount, MolecularWeight, ExactMolecularWeight
PC13	0.173	No strong association with measured descriptors
PC6	0.154	AromaticRingCount, logP, FractionCSP3

Case study: Prioritizing out-of-distribution compounds for mycomembrane permeability

To evaluate the real-world applicability of ActiveFusion, we conducted a case study using the mycomembrane permeability dataset as a central application. Mycomembrane permeability is a key prerequisite for candidate anti-tuberculosis antibiotics [62], but experimental measurement of compound permeability through the mycobacterial cell envelope is costly and time-consuming [61, 62]. Consequently, it is critical to prioritize compounds that are expected to provide the greatest benefit for model improvement when labeled.

Using the final trained surrogate model on our best performing representation, Chemprop+RDKit, predictions were generated for a pool of 1,684 unlabeled candidate compounds selected for potential experimental testing. These 1,684 compounds are selected from the Enamine catalog [63] by domain experts as candidates for experimental testing, and were previously used to validate a permeability model [28]. The entropy acquisition function was then applied to identify compounds with the highest prediction uncertainty, reflecting regions of chemical space where the model is expected to benefit most from additional information. The entropy-based acquisition function was chosen because it is the most performant non-model agnostic function for this task.

To validate that the recommended compounds indeed correspond to uncertain and under-represented regions of chemical space, we analyzed their structural similarity to the permeability training dataset and to the other members of the pool (Figure 10). The top 20 compounds recommended by ActiveFusion Chemprop+RDKit exhibit significantly lower similarity to previously labeled compounds (Mann–Whitney U test: p-value $= 1.47 \times 10^{-14}$, Kolmogorov–Smirnov test: p-value $= 1.67 \times 10^{-30}$, Cliff's delta $= -0.99$) and to the remaining compounds in the pool (Mann-Whitney U test: p-value $= 4.68 \times 10^{-2}$, Kolmogorov–Smirnov test: p-value $= 1.8 \times 10^{-1}$, Cliff's delta $= -0.26$). The negative Cliff's delta value, which is the effect size, suggests that compounds prioritized by ActiveFusion tend to occupy more structurally distinct regions of chemical space relative to the rest of the pool, indicating that the model preferentially identifies uncertain and under-represented regions for exploration.

Furthermore, projection of the selected compounds into a three-dimensional UMAP chemical space visualization shows that they are located near the boundaries of the training distribution (Figure 11) and not clustered together or within the chemical space of known labeled compounds. This observation provides additional evidence that our ActiveFusion entropy-based acquisition selects out-of-distribution compounds that expand chemical space coverage. Experimental characterization would be required to further test these selected compounds and confirm that the expansion of chemical space coverage translatees into model performance improvement.

Relationship between uncertainty-driven and diversity-driven exploration

Motivated by the strong performance of diversity-based acquisition, we investigated the overlap between compounds selected by entropy-based uncertainty sampling and those selected by diversity-based acquisition. The 20 most chemically diverse compounds exhibited substantially lower similarity to the training chemical space relative to the remaining pool (Mann–Whitney U test: p-value $= 1.39 \times 10^{-14}$, Kolmogorov–Smirnov test: p-value $= 3.40 \times 10^{-45}$, Cliff's delta $= -1.00$), confirming that they occupy highly isolated regions of chemical space. In addition, the effect size is larger compared to entropy-based acquisition. This suggests that entropy does not simply prioritize the most extreme diverse outliers, but instead suggests identification of compounds that balance structural novelty with regions of elevated model uncertainty.

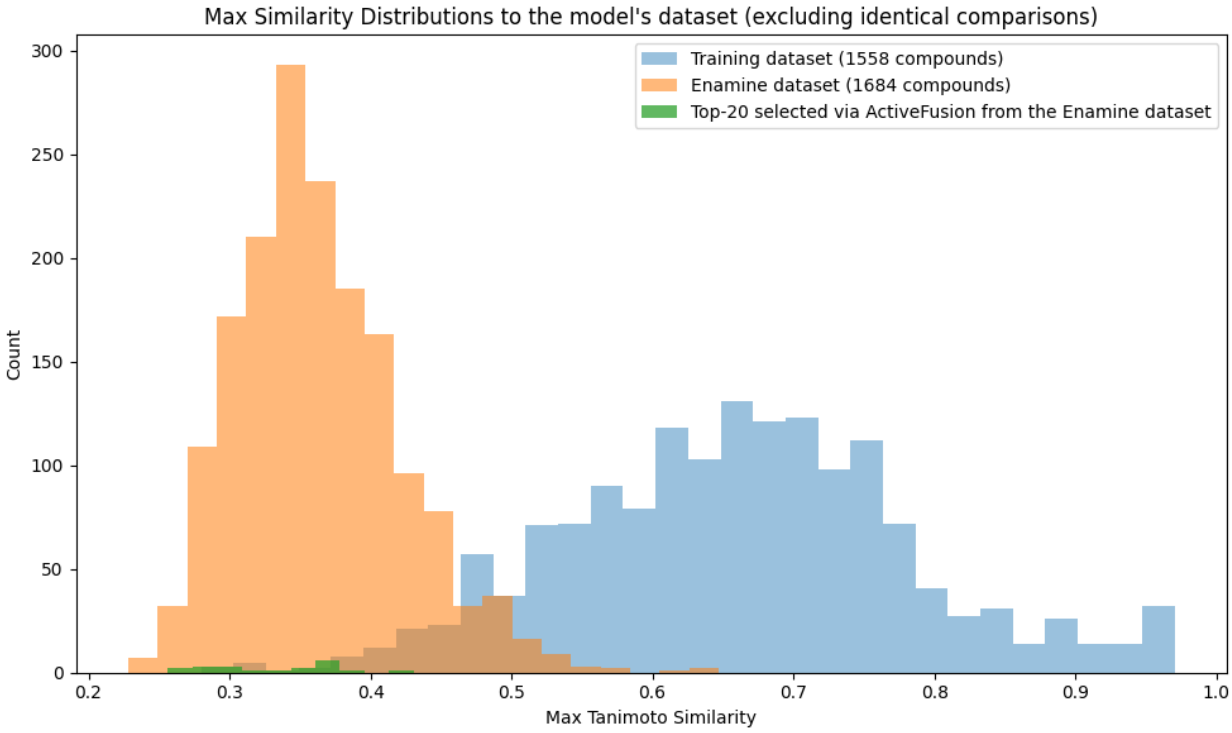

Figure 10: **Maximum similarity of the active learning-selected 20 compounds from the 1,684 Enamine library compounds with the permeability training dataset.**

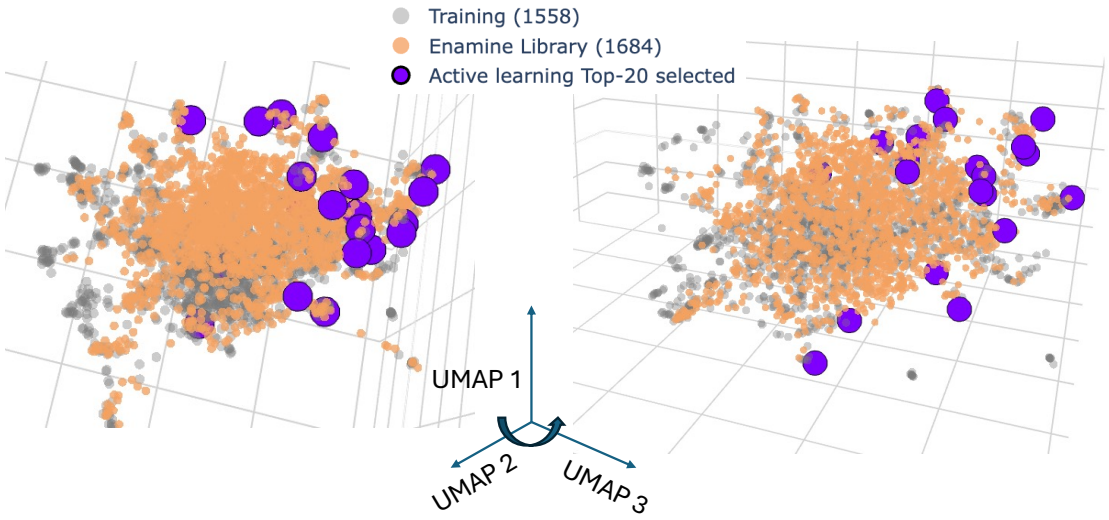

Figure 11: **UMAP analysis reveals the 20 compounds selected by ActiveFusion entropy acquisition function exist on the boundary of the chemical space relative to the permeability model training data.**

Among the top 20 compounds recommended by each strategy, one compound was found to overlap. To assess the statistical significance of this overlap, we conducted a random sampling analysis by repeatedly selecting two sets of 20 compounds from the candidate pool and computing their intersection. Across 10,000

random trials, the average overlap was approximately zero compounds with a standard deviation of 0.48, indicating that the observed overlap of the compound is unlikely to arise by chance. These results suggest that uncertainty-driven and diversity-driven acquisition strategies partially converge on chemically informative regions of the search space.

Conclusions

This work provides a comprehensive evaluation of how molecular representation choice influences active learning behavior in low-data molecular property prediction tasks. Rather than focusing solely on prediction accuracy, the ActiveFusion framework enables analysis of how representations affect exploration of chemical space, compound prioritization efficiency, and model generalization under distribution shift.

The observed benefits of feature fusion further suggest that complementary molecular information sources can be integrated to improve both predictive performance and acquisition dynamics. By combining learned embeddings with global physicochemical descriptors, ActiveFusion improved prediction accuracy during iterative data acquisition. This has practical implications for experimental discovery workflows. We find that fusing more than two representations does not yield additional benefits to predictive performance (Figure S3), especially if one of the representations is weak on its own, and may in fact lead to lower R^2 during active learning iterations.

Importantly, the positive relationship observed between scaffold exploration and prediction improvement suggests that the acquisition of chemically diverse compounds may directly contribute to improved model generalization. This insight supports the design of future active learning strategies that explicitly account for structural novelty when prioritizing compounds for experimental measurement.

Finally, the mycomembrane permeability case study illustrates how active learning frameworks can be deployed prospectively to guide compound selection in resource-constrained experimental settings. More broadly, ActiveFusion establishes a foundation for studying representation–acquisition interactions in molecular discovery pipelines and may inform the development of next-generation data-efficient drug discovery strategies.

As this study demonstrates the effectiveness of exploration-driven acquisition strategies alongside the performance gains enabled by representation fusion within active learning, future works could focus on systematically benchmarking hybrid acquisition frameworks that combine chemical diversity- and uncertainty-based selection and measuring generalization in low-data molecular property prediction settings. Prior studies suggest that such algorithmic integration may further improve data efficiency and model generalization, as this technique has been shown for computer vision and deep learning in large data regimes [64, 65]. Ultimately, future works can utilize our ActiveFusion workflow, with our best fused representation, to drive autonomous molecular property prediction in their laboratory settings.

Data and Software availability

All data and code used in this study can be found at `https://github.com/SAGE-Lab-UMass/ALBench`.

Acknowledgment

The authors thank members of the UMass SAGE lab for their valuable input and discussion, and the Siegrist lab for providing the mycomembrane permeability dataset utilized here as a case study. This work also utilized resources from Unity, a collaborative, multi-institutional high-performance computing cluster managed by UMass Amherst Research Computing and Data.

Supporting Information

Experimental setup and model hyperparameters; statistical analysis procedures; extended comparisons of molecular representations using Mann–Whitney U tests; RMSE analyses; representation fusion analyses with RDKit and ECFP; uncertainty calibration analyses; scaffold retrieval versus R^2 analyses; complete active learning evaluation results across datasets, acquisition functions, and molecular representations; additional representation performance plots; Enamine library dataset information.

References

(1) Van Tilborg, D.; Brinkmann, H.; Criscuolo, E.; Rossen, L.; Özçelik, R.; Grisoni, F. Deep learning for low-data drug discovery: Hurdles and opportunities. *Current Opinion in Structural Biology* **2024**, *86*, 102818, DOI: 10.1016/j.sbi.2024.102818, https://www.sciencedirect.com/science/article/pii/S0959440X24000459.

(2) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* **2019**, *18*, 463–477, DOI: 10.1038/s41573-019-0024-5, https://www.nature.com/articles/s41573-019-0024-5.

(3) Settles, B., *Active Learning*; Synthesis Lectures on Artificial Intelligence and Machine Learning 1, Vol. 6; Morgan & Claypool: 2012, pp 1–114, DOI: 10.2200/S00429ED1V01Y201207AIM018.

(4) Fralish, Z.; Reker, D. Taking a deep dive with active learning for drug discovery. *Nature Computational Science* **2024**, *4*, 727–728, DOI: 10.1038/s43588-024-00704-6, https://www.nature.com/articles/s43588-024-00704-6.

(5) Reker, D. Practical considerations for active machine learning in drug discovery. *Drug Discovery Today: Technologies* **2019**, *32–33*, 73–79, DOI: 10.1016/j.ddtec.2020.06.001, https://www.sciencedirect.com/science/article/pii/S1740674920300019.

(6) Bailey, M.; Moayedpour, S.; Li, R.; Corrochano-Navarro, A.; Kötter, A.; Kogler-Anele, L.; Riahi, S.; Grebner, C.; Hessler, G.; Matter, H.; Bianciotto, M.; Mas, P.; Bar-Joseph, Z.; Jager, S. Deep Batch Active Learning for Drug Discovery. *eLife* **2024**, Reviewed Preprint, DOI: 10.7554/eLife.89679, https://elifesciences.org/reviewed-preprints/89679.

(7) Yin, T.; Panapitiya, G.; Coda, E. D.; Saldanha, E. G. Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. *Journal of Cheminformatics* **2023**, *15*, DOI: 10.1186/s13321-023-00753-5, https://link.springer.com/article/10.1186/s13321-023-00753-5.

(8) Fralish, Z.; Reker, D. Finding the most potent compounds using active learning on molecular pairs. *Beilstein J. Org. Chem* **2024**, PMC11368049, DOI: https://doi.org/10.3762/bjoc.20.185, https://www.beilstein-journals.org/bjoc/articles/20/185.

(9) Wen, Y.; Li, Z.; Xiang, Y.; Reker, D. Improving molecular machine learning through adaptive subsampling with active learning. *Digital Discovery* **2023**, 1134–1142, DOI: 10.1039/D3DD00037K, https://doi.org/10.1039/D3DD00037K.

(10) Masood, M. A.; Kaski, S.; Cui, T. Molecular property prediction using pretrained-BERT and Bayesian active learning: a data-efficient approach to drug design. *Journal of Cheminformatics* **2025**, *17*, DOI: 10.1186/s13321-025-00986-6, https://jcheminf.biomedcentral.com/articles/10.1186/s13321-025-00986-6.

(11) Hao, Z.; Lu, C.; Huang, Z.; Wang, H.; Hu, Z.; Liu, Q.; Chen, E.; Lee, C. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2020)*, ACM: 2020, pp 731–752, DOI: https://doi.org/10.1145/3394486.3403117, https://dl.acm.org/doi/10.1145/3394486.3403117.

(12) Van Tilborg, D.; Grisoni, F. Traversing chemical space with active learning for low-data drug discovery. *Nature Computational Science* **2024**, *4*, 786–796, DOI: 10.1038/s43588-024-00697-2, https://www.nature.com/articles/s43588-024-00697-2.

(13) Sadeghi, S.; Bui, A.; Forooghi, A.; Lu, J.; Ngom, A. Can large language models understand molecules? *BMC Bioinformatics* **2024**, *25*, 225, DOI: 10.1186/s12859-024-05847-x, https://doi.org/10.1186/s12859-024-05847-x.

(14) Praski, M.; Adamczyk, J.; Czech, W. Benchmarking Pretrained Molecular Embedding Models For Molecular Representation Learning, 2026, https://arxiv.org/abs/2508.06199.

(15) Shen, J.; Nicolaou, C. A. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies* **2020**, *32–33*, 29–36, DOI: 10.1016/j.ddtec.2020.05.001.

(16) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754, DOI: 10.1021/ci100050t.

(17) Zhou, S.; Li, J.; Liu, Z.; Wang, D. Molecular Representation Matters: Comparative Evaluation of Fingerprints, RDKit Descriptors, and Hashing Effects. *The Journal of Physical Chemistry Letters* **2026**, *17*, PMID: 41653145, 1960–1971, DOI: 10.1021/acs.jpclett.5c03797, https://doi.org/10.1021/acs.jpclett.5c03797.

(18) Myint, K.-Z.; Wang, L.; Tong, Q.; Xie, X.-Q. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Molecular Pharmaceutics* **2012**, *9*, 2912–2923, DOI: 10.1021/mp300237z.

(19) Li, M.; Zeng, M.; Zhang, H.; Chen, H.; Guan, L. Biological Activity Predictions of Ligands Based on Hybrid Fingerprints. *ACS Omega* **2023**, *8*, 5561–5570, DOI: 10.1021/acsomega.2c06944.

(20) Leeson, P. D.; Springthorpe, B. The influence of lipophilicity in drug discovery and design. *Nature Reviews Drug Discovery* **2007**, *6*, 881–890, DOI: 10.1038/nrd2445.

(21) Yang, K.; Swanson, K.; Jin, W.; Coley, C. W.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K. F.; Barzilay, R. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388, DOI: 10.1021/acs.jcim.9b00237.

(22) 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. *Journal of Chemical Information and Modeling* **2023**, *64*, 9–17, DOI: 10.1021/acs.jcim.3c01250.

(23) Graff, D. E.; Morgan, N. K.; Burns, J. W.; Doner, A. C.; Li, B.; Li, S.-C.; Manu, J.; Menon, A.; Pang, H.-W.; Wu, H.; Zalte, A. S.; Zheng, J. W.; Coley, C. W.; Green, W. H.; Greenman, K. P. Chemprop v2: An Efficient, Modular Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling* **2025**, *66*, PMID: 41453060, 28–33, DOI: 10.1021/acs.jcim.5c02332, https://doi.org/10.1021/acs.jcim.5c02332.

(24) Singh, R.; Barsainyan, A. A.; Irfan, R.; Amorin, C. J.; He, S.; Davis, T.; Thiagarajan, A.; Sankaran, S.; Chithrananda, S.; Ahmad, W.; Jones, D.; McLoughlin, K.; Kim, H.; Bhutani, A.; Sathyanarayana, S. V.; Viswanathan, V.; Allen, J. E.; Ramsundar, B. ChemBERTa-3: an open source training framework for chemical foundation models. *Digital Discovery* **2026**, *5*, 662–685, DOI: 10.1039/D5DD00348B, http://dx.doi.org/10.1039/D5DD00348B.

(25) 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**, *4*, 1256–1264, DOI: 10.1038/s42256-022-00580-7.

(26) Weininger, D. SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36, DOI: 10.1021/ci00057a005.

(27) Masood, M. A.; Kaski, S.; Cui, T. Molecular property prediction using pretrained-BERT and Bayesian active learning: A data-efficient approach to drug design. *Journal of Cheminformatics* **2025**, *17*, 58, DOI: 10.1186/s13321-025-00986-6.

(28) Evbarunegbe, N.; Wa, S.; Karn, I.; Green, A. G. MycoPermeNet v2: Improved prediction of mycomembrane permeation using fusion noisy student self-distillation. *Journal of Chemical Information and Modeling* **2026**, DOI: 10.1021/acs.jcim.5c02435.

(29) Fang, C.; Wang, Y.; Grater, R.; Kapadnis, S.; Black, C.; Trapa, P.; Sciabola, S. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. *Journal of Chemical Information and Modeling* **2023**, *63*, PMID: 37216672, 3263–3274, DOI: 10.1021/acs.jcim.3c00160, https://doi.org/10.1021/acs.jcim.3c00160.

(30) Ding, X.; Cui, R.; Yu, J.; Liu, T.; Zhu, T.; Wang, D.; Chang, J.; Fan, Z.; Liu, X.; Chen, K.; Jiang, H.; Li, X.; Luo, X.; Zheng, M. Active Learning for Drug Design: A Case Study on the Plasma Exposure of Orally Administered Drugs. *Journal of Medicinal Chemistry* **2021**, *64*, 16838–16853, DOI: 10.1021/acs.jmedchem.1c01683, https://doi.org/10.1021/acs.jmedchem.1c01683.

(31) Mohr, B.; Shmilovich, K.; Kleinwächter, I. S.; Schneider, D.; Ferguson, A. L.; Bereau, T. Data-driven discovery of cardiolipin-selective small molecules by computational active learning. *Chemical Science* **2022**, 4498–4511, DOI: 10.1039/D2SC00116K, https://pubs.rsc.org/en/content/articlelanding/2022/sc/d2sc00116k.

(32) Williams, H. J.; Pickett, S. D.; Baxter, A.; Palmer, D. S. Query Matters: How Selection Strategies Influence Active Learning in Drug Discovery. *Journal of Chemical Information and Modeling* **2026**, *66*, 3288–3299, DOI: 10.1021/acs.jcim.5c02504, https://doi.org/10.1021/acs.jcim.5c02504.

(33) Zetta, D. N.; Srisongkram, T. Active stacking-deep learning with strategic sampling for data-efficient toxicity prediction under class imbalance. *ACS Omega* **2025**, *10*, 53907–53926, DOI: 10.1021/acsomega.5c04016, https://doi.org/10.1021/acsomega.5c04016.

(34) Zetta, D. N.; Srisongkram, T. Data-efficient learning for accurate identification of MAPK1 inhibitors using an active meta-deep learning framework. *Journal of Cheminformatics* **2026**, *18*, 34, DOI: 10.1186/s13321-026-01159-9, https://doi.org/10.1186/s13321-026-01159-9.

(35) 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. *Chemical Science* **2018**, *9*, 513–530, DOI: 10.1039/C7SC02664A.

(36) Lepori, I.; Liu, Z.; Evbarunegbe, N.; Feng, S.; Brown, T. P.; Mane, K.; Shivangi; Wong, M.; George, A.; Guo, T.; Dong, J.; Freundlich, J. S.; Im, W.; Green, A. G.; Pires, M. M.; Siegrist, M. S. Identification of chemical features that influence mycomembrane permeation and antitubercular activity. *bioRxiv* **2025**, DOI: 10.1101/2025.02.27.640664.

(37) 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. *Angewandte Chemie International Edition* **2023**, *62*, e202217777, DOI: https://doi.org/10.1002/anie.202217777, https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202217777.

(38) Delaney, J. S. ESOL: A simple model for aqueous solubility based on molecular topology. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1000–1005, DOI: 10.1021/ci034243x.

(39) Mobley, D. L.; Guthrie, J. P. FreeSolv: A database of experimental and calculated hydration free energies. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 711–720, DOI: 10.1007/s10822-014-9747-x.

(40) Hersey, A. ChEMBL Deposited Data Set - AZ dataset, ChEMBL database deposition, Accessed via ChEMBL, https://www.ebi.ac.uk/chembl/.

(41) Landrum, G. RDKit: Open-source cheminformatics, https://www.rdkit.org, 2023.

(42) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893, DOI: 10.1021/jm9602928.

(43) 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. *Journal of Medicinal Chemistry* **2019**, *63*, 7389–7400, DOI: 10.1021/acs.jmedchem.9b00959.

(44) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *Proceedings of ICML*, 2017, https://proceedings.mlr.press/v70/gilmer17a.html.

(45) Mswahili, M. E.; Jeong, Y.-S. Transformer-based models for chemical SMILES representation: A comprehensive literature review. *Heliyon* **2024**, *10*, e39038, DOI: 10.1016/j.heliyon.2024.e39038, https://www.sciencedirect.com/science/article/pii/S2405844024150694.

(46) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023, https://openreview.net/pdf?id=6K2RM6wVqKu.

(47) Rasmussen, C. E.; Williams, C. K. I., *Gaussian Processes for Machine Learning*; MIT Press: 2006, https://gaussianprocess.org/gpml/chapters/RW.pdf.

(48) Wilson, A. G.; Hu, Z.; Salakhutdinov, R.; Xing, E. P. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016, pp 370–378, https://arxiv.org/abs/1511.02222.

(49) Singh, S.; Hernández-Lobato, J. M. Deep kernel learning for reaction outcome prediction and optimization. *Communications Chemistry* **2024**, *7*, DOI: 10.1038/s42004-024-01219-x.

(50) Ghosh, A.; Ziatdinov, M.; Kalinin, S. V. Active deep kernel learning of molecular properties from structural embeddings. *APL Machine Learning* **2025**, *3*, 046103, DOI: 10.1063/5.0282700.

(51) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization* **1998**, *13*, 455–492, DOI: 10.1023/A:1008306431147.

(52) Gorantla, R.; Kubincová, A.; Suutari, B.; Cossins, B. P.; Mey, A. S. J. S. Benchmarking Active Learning Protocols for Ligand-Binding Affinity Prediction. *Journal of Chemical Information and Modeling* **2024**, *64*, PMID: 38446131, 1955–1965, DOI: 10.1021/acs.jcim.4c00220, https://pubs.acs.org/doi/10.1021/acs.jcim.4c00220.

(53) Cao, Z.; Sciabola, S.; Wang, Y. Large-Scale Pretraining Improves Sample Efficiency of Active Learning-Based Virtual Screening. *Journal of Chemical Information and Modeling* **2024**, *64*, PMID: 38442000, 1882–1891, DOI: 10.1021/acs.jcim.3c01938, https://doi.org/10.1021/acs.jcim.3c01938.

(54) Thaler, S.; Mayr, F.; Thomas, S.; Gagliardi, A.; Zavadlav, J. Active learning graph neural networks for partial charge prediction of metal-organic frameworks via dropout Monte Carlo. *npj Computational Materials* **2024**, *10*, DOI: 10.1038/s41524-024-01277-8, https://doi.org/10.1038/s41524-024-01277-8.

(55) In *Elements of Information Theory*; John Wiley & Sons, Ltd: 2005; Chapter 8, pp 243–259, DOI: https://doi.org/10.1002/047174882X.ch8, https://onlinelibrary.wiley.com/doi/abs/10.1002/047174882X.ch8.

(56) Willett, P. The calculation of molecular structural similarity: principles and practice. *Molecular Informatics* **2014**, *33*, 403–413, DOI: https://doi.org/10.1002/minf.201400024, https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201400024.

(57) Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction*; tech. rep.; IBM, 1958, http://dalkescientific.com/tanimoto.pdf.

(58) Lepori, I.; Liu, Z.; Evbarunegbe, N.; Feng, S.; Brown, T. P.; Mane, K.; Dash, R.; Shivangi; Wong, M.; Naick, A.; George, A.; Guo, T.; Shelke, A. M.; Sharpless, K. B.; Dong, J.; Freundlich, J. S.; Im, W.; Green, A. G.; Pires, M. M.; Siegrist, M. S. Identification of chemical features for improved outer membrane permeation in mycobacteria using machine learning. *Nature Microbiology* **2026**, Published online 30 June 2026, DOI: 10.1038/s41564-026-02412-5, https://doi.org/10.1038/s41564-026-02412-5.

(59) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426* **2018**, https://arxiv.org/abs/1802.03426.

(60) Draper, P. The Outer Parts of the Mycobacterial Envelope as Permeability Barriers. *Frontiers in Bioscience* **1998**, *3*, D1253–D1261, DOI: 10.2741/A360, https://pubmed.ncbi.nlm.nih.gov/9851911/.

(61) Liu, Z.; Lepori, I.; Chordia, M. I.; Dalesandro, B. E.; Guo, T.; Dong, J.; Siegrist, S. M.; Pires, M. M. A Metabolic-Tag-Based Method for Assessing the Permeation of Small Molecules Across the Mycomembrane in Live Mycobacteria. *Angewandte Chemie International Edition* **2023**, *62*, e202217777, DOI: https://doi.org/10.1002/anie.202217777.

(62) 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 Small-Molecule Permeation in Mycobacteria. *ACS Infectious Diseases* **2025**, *11*, 1893–1906, DOI: 10.1021/acsinfecdis.4c01062, https://pubs.acs.org/doi/full/10.1021/acsinfecdis.4c01062.

(63) Enamine Ltd. Compound Libraries, https://enamine.net/compound-libraries, Accessed: 2026-05-20, 2026.

(64) Wu, J.; Chen, J.; Huang, D. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp 9387–9396, DOI: 10.1109/CVPR52688.2022.00918.

(65) He, Y.; Cai, L.; Liao, J.; Foo, C.-S. Hybrid Active Learning with Uncertainty-Weighted Embeddings. *Transactions on Machine Learning Research* **2024**, https://openreview.net/forum?id=jD761b50aE.

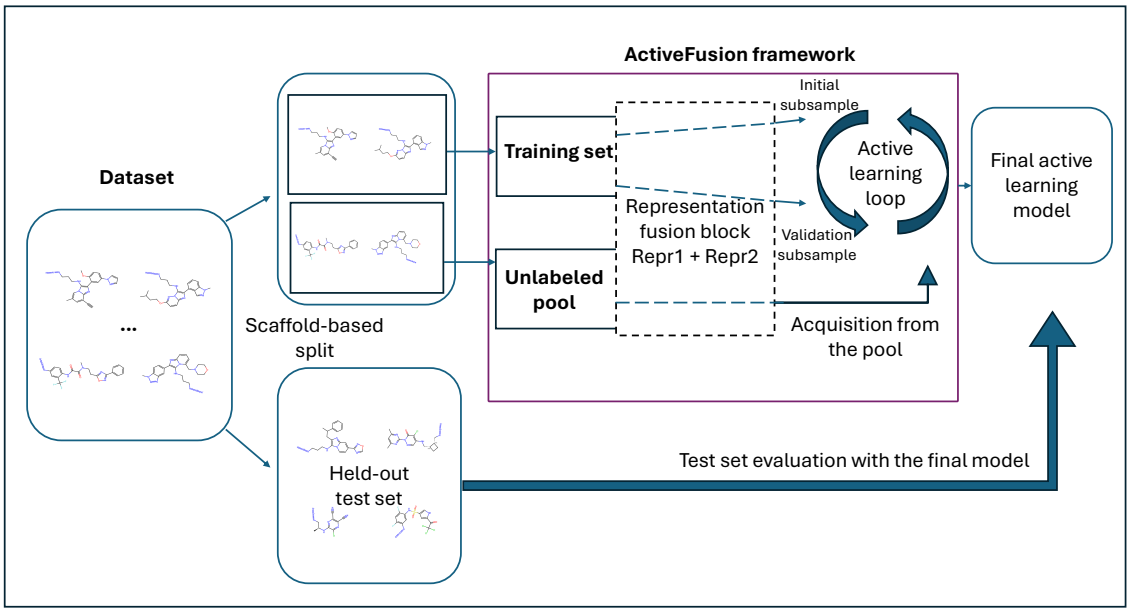

Figure 12: **Graphical Abstract**