# MODEL-DRIVEN COLLABORATIVE INFORMATION ENHANCES DRUG ACTIVITY PREDICTION IN LOW-RESOURCE SCENARIOS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Drug development is hindered by the high cost and time consumption of identifying bioactive molecules, particularly in zero-shot and few-shot scenarios where experimental data is scarce. Existing methods often fail to generalize to novel proteins or assays, limiting their applicability in real-world drug discovery. Furthermore, these approaches struggle with the simultaneous processing of heterogeneous data types, including assay descriptions, molecular structures, and protein sequences. To address these limitations, we propose *Assay Similarity-based Activity Prediction* (ASAP), a novel framework that leverages a model-driven assay similarity and ranking loss to predict bioactivity. By learning the collaborative information between assays and mitigating batch effects, ASAP enables robust predictions for low-resource scenarios, such as novel assays with only text or protein sequence information. Extensive experiments demonstrate the superiority of ASAP in both zero-shot and few-shot settings. Our proposed approach provides a scalable, generalizable solution for accelerating drug discovery. Source code is available here.

## 1 INTRODUCTION

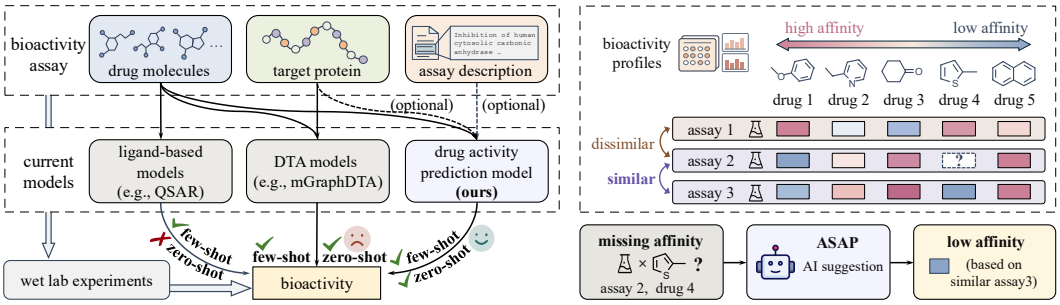
Drug development is an expensive and time-consuming process, spanning over a decade and costing approximately $2.8 billion per successful drug (DiMasi et al., 2016; Sun et al., 2022; Sertkaya et al., 2024). A critical step is identifying bioactive drug molecules, or ligands, that can interact with specific biological targets, such as proteins or nucleic acids, to modulate their function. *Bioactivity assays*, which quantify drug-target interactions through various protocols and readout mechanisms, generate valuable but heterogeneous data that requires sophisticated computational approaches. The ability to accurately predict ligand bioactivity before experimental validation can greatly accelerate drug discovery by allowing researchers to focus on the most promising candidates (Cumming et al., 2013). Despite its importance, this task remains challenging due to complex and often non-linear relationships between chemical structures and their biological activities (Nantasenamat et al., 2010).

Traditional ligand-based approaches, such as quantitative structure-activity relationship (QSAR) modeling, rely on structure and physicochemical properties to predict drug bioactivity (Perkins et al., 2003; Sahoo et al., 2022). Recent deep learning methods including convolutional neural networks (CNNs) (Son & Kim, 2021), graph neural networks (GNNs) (Torng & Altman, 2019; Liu et al., 2021), and Transformers (Tang et al., 2024) have shown promise in drug-target affinity (DTA) prediction, due to their capacity to model high-dimensional data and uncover latent features (Jiménez et al., 2018; Nguyen et al., 2021; Li et al., 2021). In real-world applications, these models usually suffer from *poor transferability*, particularly for understudied targets with sparse training data (Dalkıran et al., 2023). To overcome this limitation, recent advances have introduced more sophisticated learning techniques. ActFound learns within-assay activity differences through pairwise meta-learning (Feng et al., 2024), AdaMBind uses task-adaptive meta-learning for new targets (Wan et al., 2026), and FS-CAP predicts activities by conditioning on a few measured compounds (Eckmann et al., 2024).

Assay descriptions provide complementary information about experimental conditions. CLAMP and TwinBooster combine descriptions with molecular representations for bioactivity prediction (Seidl et al., 2023; Schuh et al., 2024). Assay-aware modeling further shows that experimental context can explain variation between measurements of the same compound–protein pair (Schoenmaker et al.,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




(a) The comparison with current related models.

(b) The idea illustration of our work.

Figure 1: (a) Our ASAP compared with current ligand-based and drug-target affinity (DTA) models. The ligand-based models cannot be adapted to new assays. The current DTA models have poor generalization ability to new assays and require target protein information as input. (b) An illustration of the idea of our work. Assays with similar ligand rankings can help the model make predictions.

2025). These advances motivate learning from the assay as well as the molecule and target. However, two challenges remain when using historical assay data in real-world applications.

**C1: Identifying transferable information across assays**. As illustrated in Figure 1a, current DTA methods perform adequately in transductive settings but generalize poorly to unseen proteins or novel drug molecules, despite recent advances (Feng et al., 2024; Wan et al., 2026). Historical assays differ in their targets, chemical space, and experimental conditions. Differences in measurement endpoints, units, and scales further complicate direct comparison. This severely restricts their applicability in practical drug discovery.

**C2: Effectively integrating heterogeneous experimental data.** Conventional DTA models rely heavily on the availability of protein structural or interaction data, which is often incomplete or absent for novel target proteins. This dependency limits their applicability to novel drug candidates or assays without well-characterized target proteins. For example, phenotypic drug discovery focuses on identifying compounds that induce desired phenotypic changes without prior knowledge of specific target (Vincent et al., 2022). Without accurate bioactivity prediction for uncharacterized target proteins or assays, the efficient in-silico identification of phenotypic hits becomes a significant bottleneck. CLAMP and TwinBooster use assay descriptions but do not explicitly exploit available protein-sequence information (Seidl et al., 2023; Schuh et al., 2024). Technical differences in experimental protocols can also introduce batch effects, hindering effective multi-source data integration.

**Presented Work.** Motivated by these challenges, we propose *Assay Similarity-based Activity Prediction* (ASAP), a novel framework for bioactivity prediction. In this context, an **assay** refers to a specific experimental procedure featured by a target protein sequence and/or a textual assay description. ASAP is designed to leverage the inherent collaborative relationships among assays (Figure 1b). To achieve this, we introduce a model-driven approach that quantifies assay similarity based on their bioactivity data, using these meaningful relationships to guide the pre-training of assay embeddings with a contrastive margin loss. This step enables the *inductive generalization* of well-characterized assays to novel, less-informed assays. Additionally, to mitigate the batch effects, ASAP learns ranking across drug molecules in an assay instead of regressing absolute activity values.

By combining these innovative strategies, our approach provides a robust framework for bioactivity prediction in low-resource scenarios, i.e., zero-shot and few-shot, where a new assay has only text or protein sequence information. ASAP not only overcomes the limitations of the pair-input models, but also bridges the gap between heterogeneous experimental data and real-world drug discovery applications. Our main contributions are summarized as follows:

- We introduce ASAP, a bioactivity prediction framework that utilizes assay similarity as collaborative signals to enable inductive generalization, allowing accurate bioactivity predictions for novel assays.
- We propose a model-driven similarity between assays and use it to pre-train assay embeddings, which enables bioactivity prediction given only the text or sequence modality.
- Our experiments on various datasets address existing approaches' limitations by including both zero-shot prediction for novel assays and few-shot prediction with limited activity data, demonstrating the effectiveness of ASAP.

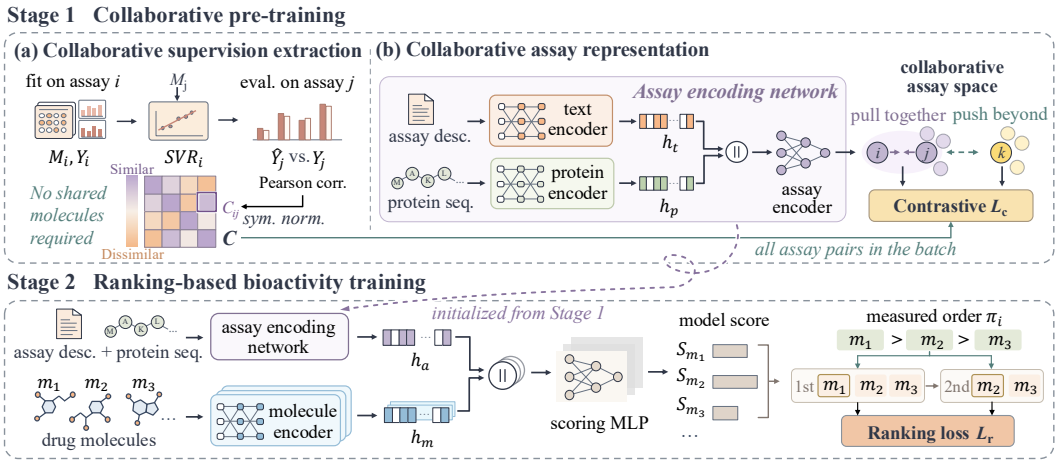



Figure 2: Our proposed pipeline ASAP. (1) Upper: The collaborative pre-training stage learns the assay similarity for the assay encoding. (2) Down: The pre-trained assay embedding is then used for ranking-based bioactivity training with the molecule embedding.

## 2 Problem Setting

**Bioactivity Prediction.** Bioactivity prediction is typically formulated as a classical regression task, where the aim is to assess the interaction strength between a *drug molecule* and a *target protein*. For a given *assay* regarding a specific target protein $p \in \mathcal{P}$, which comes with a textual description $t \in \mathcal{T}$ about its experimental procedure, the machine learning model $f$ is trained on paired samples $\{(m_1, y_1), \cdots, (m_N, y_N)\}$, where $m_n \in \mathcal{M}$ is a drug molecule from chemical space, and $y_n$ is a continuous value quantifying the activity value (e.g., $\text{IC}_{50}$, $\text{K}_i$, and $\text{K}_d$).

Drug molecules can be represented in various formats, including SMILES strings (Weininger, 1988), molecular graphs, or 3D conformations containing atomic coordinates. Similarly, target proteins can be encoded as either amino acid sequences or 3D structural information. In this work, we adopt SMILES strings for drug molecule representations and amino acid sequences for target proteins.

**Few-shot Bioactivity Prediction.** In few-shot settings, we predict bioactivities in a new assay with limited training samples. Given a test assay $A_{\text{test}} = \{t_{\text{new}}, p_{\text{new}}, \mathcal{M}_{\text{test}}\}$ where $\mathcal{M}_{\text{test}} = \{m_1, ..., m_N\}$, we have only $K$ labeled measurements $\{(m_1, y_1), \cdots, (m_K, y_K)\}$ where $K \ll N$. Our goal is to leverage these few measurements alongside a large set of historical assays $\mathcal{A}_{\text{train}} = \{A_1, \cdots, A_D\}$ to predict the bioactivities for the remaining drug molecules $\{m_{K+1}, ..., m_N\}$ in $A_{\text{test}}$.

**Zero-shot Bioactivity Prediction.** In zero-shot scenarios, we aim to predict the bioactivities for a new assay without any labeled measurements. Formally, given a test assay $A_{\text{test}} = \{t_{\text{new}}, p_{\text{new}}, \mathcal{M}_{\text{test}}\}$, we need to predict bioactivities for all drug molecules in $\mathcal{M}_{\text{test}} = \{m_1, ..., m_N\}$ solely based on the description $t_{\text{new}}$ and the target protein $p_{\text{new}}$, together with historical assays $\mathcal{A}_{\text{train}} = \{A_1, \cdots, A_D\}$.

While in many real experiments, assay targets can be biological molecules beyond proteins or a phenotype like cell proliferation. Considering that assay descriptions and target proteins may not always be available, we investigate two sub-settings of zero-shot prediction in our experiments: (1) Target-drug setting: when assay description is unavailable, i.e., $A_{\text{test}} = \{p_{\text{new}}, \mathcal{M}_{\text{test}}\}$; (2) Text-drug setting: when target protein information is missing, i.e., $A_{\text{test}} = \{t_{\text{new}}, \mathcal{M}_{\text{test}}\}$.

## 3 Methodology

**Preliminary Analysis.** We observe a natural correlation existing in activity prediction tasks. When two assays demonstrate similar rankings in their drug activity patterns across their respective target proteins, it indicates an underlying correlation in their drug response profiles. Such *collaborative relationships* among assays suggest that assays with comparable behavior may share similar predictive

signals. By capturing these relationships, a model can effectively leverage historical data from related assays when predicting activities for a *previously unseen new assay*. Specifically, given the assay description and target protein, if a model could identify analogous assays in the collaborative latent space and utilize their information to inform predictions, the few-shot and zero-shot prediction capabilities can be enhanced.

ASAP is designed to capture and leverage the above *bioactivity collaborative information* for prediction. Figure 2 illustrates the pipeline of ASAP, which can be divided into two parts: First, we extract the collaborative relationships from historical assay data with a model-driven assay similarity. With collaborative supervision by pre-training on large-scale datasets, assay embeddings then learn to capture the collaborative latent space features. Second, we employ ranking-based learning to jointly optimize the assay and drug molecule embeddings with collaborative information, translating the learned collaborative information into predictive capability for assay activity.

### 3.1 Stage 1: Collaborative Pre-training

Incorporating collaborative information into the pre-training stage presents two major challenges. First, obtaining reliable collaborative supervision signals from large-scale pre-training data. Second, designing an appropriate optimization objective that effectively enables the model to capture these collaborative patterns. To address these, we introduce the designs as follows.

#### 3.1.1 Collaborative Supervision Extraction

A conventional approach to extracting collaborative information from correlated assays involves explicitly capturing assay similarity through metrics like cosine similarity. However, applying this approach to assay similarity computation presents several challenges: **(1) Sparse Overlap**. Drug molecules tested across different assays can vary significantly, potentially leading to zero overlaps between assay pairs and making traditional similarity calculations infeasible. **(2) Incomparable Bioactivities**. Activity values measured in different assays often lack direct comparability, rendering similarity metrics based on absolute activity values meaningless. **(3) Transductive Nature**. Similarity calculations based on historical interactions are inherently transductive; any experimental database update requires recalculating all pairwise similarities, limiting scalability.

To address the challenges mentioned above, we propose a model-based collaborative supervision extraction approach. The core idea is to use a model to evaluate the similarity between two assays, based on the intuition that if a model trained on one assay's activity can accurately predict another assay's activity, there should be high collaborative similarity between them. Specifically, we adopt Support Vector Regression (SVR) (Hearst et al., 1998) as the predictive model and use extended-connectivity fingerprints (ECFPs) (Rogers & Hahn, 2010) to represent drug molecules, given their effectiveness in capturing molecular structural characteristics. The similarity $s(i, j)$ from assay $i$ to assay $j$ can be formalized as:

$$\begin{aligned} (\theta_i^*, b_i^*) &= \text{Training}(\text{SVR}_{\theta,b}(\boldsymbol{M}_i), Y_i), \\ \hat{Y}_j = \text{SVR}_{\theta_i^*, b_i^*}(\boldsymbol{M}_j), \quad & s(i,j) = \text{correlation}(\hat{Y}_j, Y_j), \end{aligned} \tag{1}$$

where $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$ is the ECFP feature matrix of all drug molecules in assay $i$ and assay $j$, respectively, and $\theta_i^*, b_i^*$ are the SVR parameters learned from assay $i$. $Y_i$ and $Y_j$ are their corresponding activity values. $\hat{Y}_j$ is the assay $j$'s predicted values on drug molecules, and correlation$(\cdot, \cdot)$ is the correlation function, which we implement using the Pearson coefficient to measure the linear correlation between predicted and actual values.

To ensure the similarity metric reflects the intuitive properties of assay relationships, we define the final similarity matrix $\boldsymbol{C}$ as: $\boldsymbol{C}_{ij} = 1/4\big(s(i,j) + s(j,i)\big) + 1/2$, where this transformation ensures symmetry in pairwise assay relationships and constrains similarity scores to the interval $[0, 1]$, thereby providing an interpretable probability-like measure of the collaborative similarity between assays.

**Case Study.** To investigate whether our extracted collaborative similarity ($\boldsymbol{C}_{ij}$) is meaningful and what it specifically signifies from biological mechanism perspectives, we review specific assay pairs with varying similarity scores. ① Similar case: two assays targeting `5-HT1a receptor` (CHEMBL616344) and the `Dopamine D2 receptor` (CHEMBL3599873) have high similarity. Indeed, these two receptors share structural, functional, and pharmacological overlaps despite their

distinct neurotransmitter systems. Drugs acting as agonists or antagonists at either receptor may produce analogous functional readouts (e.g., cAMP inhibition) in assays, leading to correlated rankings. ② Dissimilar case: conversely, assays targeting the `5-HT1a receptor` (CHEMBL616344) and `PI3K`$\gamma$ (CHEMBL4416687) exhibit low collaborative similarity. Although these two targets have different functions, research into their structures provides some support for the negatively correlated assay rankings. The compounds that are optimized to fit into the 5-HT1a receptor pocket will be terrible at forming the specific hydrogen bonds required at the PI3K$\gamma$ hinge. Conversely, the flat, neutral molecules designed to be ATP-mimetics for PI3K$\gamma$ will lack the positive charge needed to stick to the 5-HT1a receptor.

These case analyses demonstrate that our collaborative information aligns with biological reality (e.g., similar target, similar mechanisms), confirming that our model-based extraction captures meaningful collaborative relationships for subsequent training stages. More robustness and interpretability analyses of collaborative similarity are provided in Appendix E.

#### 3.1.2 Contrastive Margin Loss

To design an appropriate optimization objective that enables the model to capture collaborative patterns, we use a contrastive margin loss that adaptively adjusts the learning signals based on the similarity between assays. The key idea is to pull similar (i.e., collaborating) assays closer in the embedding space while pushing dissimilar (i.e., non-collaborative) ones apart beyond a margin, which helps better differentiate the degree of collaborative relationships between assays. Given assay embeddings and their similarity matrix $\boldsymbol{C}$, the loss function can be formulated as:

$$\mathcal{L}_c = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} [w_{ij} d_{ij}^2 + (1 - w_{ij}) \max(0, \delta - d_{ij})^2], \tag{2}$$

where $B$ is the batch size, $d_{ij} = \|\mathbf{h}_{a_i} - \mathbf{h}_{a_j}\|$ denotes the Euclidean distance between assay embeddings, $\delta$ is a margin parameter, and $w_{ij}$ represents the weight that determines the push-pull behavior between assay pairs. We adopt a sigmoid-based weighting strategy to leverage collaborative similarity with a smooth transition between collaborative and non-collaborative pairs:

$$w_{ij} = \sigma(\beta(\boldsymbol{C}_{ij} - 0.5)), \tag{3}$$

where $\sigma$ is the sigmoid function. Parameter $\beta$ controls the steepness of the transition around the midpoint threshold of 0.5, allowing nuanced adjustments.

#### 3.1.3 Assay Encoding

While training a large model from scratch is possible, we opt to initialize our encoder with pre-trained models due to their established advantages. Pre-trained models have already captured rich domain knowledge from massive corpora and demonstrated superior generalization capabilities, particularly in low-resource scenarios. This aligns well with our intended scenario where labeled data is limited.

Various architectures exist for text and protein encoding, we adopt two specialized pre-trained models in this work, leaving the exploration of alternative encoders for future investigation. For assay descriptions, we leverage PubMedBERT (Gu et al., 2021), which processes textual information from biomedical literature. For target proteins, we utilize the pre-trained ESM2 model (Lin et al., 2023), which encodes evolutionary information from protein sequences:

$$\mathbf{h}_t = \text{PubMedBERT}([\text{CLS}], tt_1, tt_2, \ldots, tt_b), \tag{4}$$

$$\mathbf{h}_p = \text{ESM2}([\text{CLS}], pt_1, pt_2, \ldots, pt_c), \tag{5}$$

where $(tt_1, \ldots, tt_b)$ and $(pt_1, \ldots, pt_c)$ are tokenized sequences of the assay description and target protein. Final representation is obtained by average pooling over the last hidden states of all tokens.

To integrate the semantic information from the assay description with the sequence information of the target protein, we employ a two-layer perceptron network with ReLU activation to obtain the embedding of the assay that captures combined insights from both data sources: $\mathbf{h}_a = \text{MLP}(\mathbf{h}_t || \mathbf{h}_p)$, where $||$ represents concatenation operation.

### 3.2 Stage 2: Ranking-based Bioactivity Training

#### 3.2.1 Drug Molecule Encoding

After the pre-training in the 1st stage, the assay embedding is equipped with collaborative information. It can then be employed to enhance bioactivity prediction with the drug molecule encoding. Similar to protein encoding, we utilize a specialized pre-trained molecular model, ChemBERTa (Chithrananda et al., 2020), in this work, as it is well-suited for capturing the intricate chemical and structural properties of drug molecules as follows:

$$\mathbf{h}_m = \text{ChemBERTa}([\text{CLS}], mt_1, mt_2, \ldots, mt_d), \tag{6}$$

where $(mt_1, mt_2, \ldots, mt_d)$ denotes the tokenized sequence of the drug molecules. To extract a meaningful representation, we apply average pooling on the last hidden states of all input tokens.

#### 3.2.2 Ranking Loss

To translate the learned collaborative information into predictive capability for assay activity, we adopt the ListMLE loss from Xia et al. (2008). This ranking loss considers the entire ranked list simultaneously, making it particularly suitable for problems where relative ordering matters more than absolute values. The ranking loss for a batch of predictions is formulated as:

$$\mathcal{L}_r = -\frac{1}{B_u}\sum_{i=1}^{B_u}\sum_{k=1}^{N_{u_i}} \log \frac{\exp(s_{\pi_i(k)})}{\sum_{j=k}^{N_{u_i}} \exp(s_{\pi_i(j)})}, \quad s_{\pi_i(k)} = \text{MLP}(\mathbf{h}_{a_i} || \mathbf{h}_{m_{\pi_i(k)}}), \tag{7}$$

where $B_u$ is the number of assays in a batch, $N_{u_i}$ is the number of drug molecules for assay $i$, and $\pi_i(\cdot)$ is the permutation function that sorts molecules by their ground truth activities. The predicted activity $s_{\pi_i(k)}$ for molecule at position $k$ is computed by concatenating the embeddings of assay $i$ and molecule $\pi_i(k)$, then passing through a perceptron network. This loss effectively maximizes ground truth ranking likelihood by comparing each prediction with all remaining molecules in the assay.

### 3.3 Downstream Evaluation

After stage-2 training on a large historical dataset, we obtain a bioactivity prediction model, where the assay embeddings capture features of the collaborative latent space. For a new assay, we encode its available text description and/or protein sequence, representing any missing modality with the corresponding encoder's mask token. This assay-independent placeholder preserves the input dimension, while the observed modality supplies assay-specific context. The trained model supports zero-shot prediction without assay-specific measurements. When a small set of labeled compounds is available, it can be fine-tuned using the same within-assay ranking objective for few-shot adaptation to the new assay. The following section details our experiments from these two aspects.

## 4 Experiments

To evaluate the effectiveness of our method, we introduce experiments with different settings in this section, with the aim of answering the following research questions:

- **RQ1**: How effective is ASAP in zero-shot transfer for both target-drug and text-drug settings?
- **RQ2**: Can ASAP effectively leverage a few labeled examples to improve transfer performance?
- **RQ3**: How do different pre-training components and strategies contribute to overall performance?

### 4.1 Experimental Protocols

Table 1: The statistics of datasets.

| Dataset | # Assay | # Drug | # Protein | # Pair |
|---|---|---|---|---|
| $\text{ChEMBL}_{\text{pre-train}}$ | 17,517 | 503,317 | 2,248 | 942,534 |
| $\text{ChEMBL}_{\text{test}}$ | 4,380 | 159,800 | 1,235 | 200,689 |
| Davis | 379 | 68 | 379 | 9,026 |
| KIBA | 229 | 2,068 | 229 | 118,036 |

**Datasets.** We use ChEMBL (Zdrazil et al., 2024) for pre-training and Davis (Davis et al., 2011) and KIBA (Tang et al., 2014) for downstream evaluation, following prior work (Yang et al., 2022; Nguyen et al., 2021). From the 19 million experimentally measured activities available in ChEMBL, we select assays targeting single proteins with 20–10,000

Table 2: Zero-shot target-drug results (in %) on Davis and KIBA. Error bars represent the standard deviation across three runs. Red / yellow cells indicate the best/second-best performance.

| Dataset | Davis | | | | KIBA | | | |
|---|---|---|---|---|---|---|---|---|
| Split | random | | target sim | | random | | target sim | |
| Metric | *w*-Pearson | *w*-CI | *w*-Pearson | *w*-CI | *w*-Pearson | *w*-CI | *w*-Pearson | *w*-CI |
| DeepDTA | 45.82±1.64 | 66.01±1.22 | 44.28±2.31 | 64.09±0.72 | 58.87±1.63 | 72.92±1.68 | 45.03±2.38 | 66.89±1.12 |
| DeepAffinity | 40.61±2.53 | 62.58±1.21 | 40.19±1.64 | 62.41±0.83 | 61.97±2.09 | 74.03±1.92 | 48.02±1.67 | 67.74±1.79 |
| TransformerCPI | 46.18±2.47 | 65.53±1.18 | 44.02±1.91 | 64.36±1.56 | 64.21±1.58 | 76.39±1.71 | 50.68±1.92 | 69.22±2.07 |
| GraphDTA | 40.67±1.61 | 63.71±0.88 | 39.02±0.71 | 61.72±2.34 | 60.28±0.83 | 73.87±1.46 | 46.64±1.39 | 68.26±0.51 |
| mGraphDTA | 51.61±2.12 | 67.19±0.29 | 49.42±1.43 | 65.33±0.58 | 68.48±0.72 | 78.27±0.57 | 52.87±0.86 | 70.46±0.81 |
| MMDG-DTI | 51.32±1.43 | 66.93±0.69 | 48.82±1.76 | 66.12±0.44 | 68.45±0.83 | 78.23±0.72 | 54.43±1.08 | 70.43±1.09 |
| **ASAP** | 54.52±1.63 | 68.51±0.67 | 51.18±1.73 | 66.96±0.67 | 68.82±1.28 | 78.39±0.92 | 54.41±1.46 | 70.94±1.03 |

molecules and apply quality filters to ensure data reliability (details in Appendix B.1). To prevent data leakage between pre-training and downstream evaluation, we remove ChEMBL molecules present in Davis or KIBA and use MMseqs2 (Steinegger & Söding, 2017) to exclude assays whose targets share $\geq 30\%$ sequence identity with a benchmark target at $\geq 80\%$ target-sequence coverage. The curated ChEMBL data are split by assay ID into pre-training and test partitions at an 8:2 ratio. The pre-training partition is further divided into training and validation subsets at a 9:1 ratio, with Table 1 reporting statistics before this internal split. Only training assays are used to fit assay-specific models, construct collaborative supervision, and optimize model parameters; validation assays are reserved for model selection. For Davis and KIBA, we group measurements by target protein to define assays. Following prior work (Yang et al., 2022; Hua et al., 2025), we exclude Davis entries assigned the $10\,\mu$M threshold, which represent weak binding or interactions not detected in the primary screen rather than precisely measured affinities (Davis et al., 2011).

**Baselines.** For prediction using molecules and targets, we retain DeepDTA (Öztürk et al., 2018), TransformerCPI (Chen et al., 2020), DeepAffinity (Karimi et al., 2019; 2020), GraphDTA (Nguyen et al., 2021), mGraphDTA (Yang et al., 2022), and MMDG-DTI (Hua et al., 2025). To ensure a unified evaluation, we exclude methods that rely on the three-dimensional structure of the target protein, because the target protein information in our study is only available in the form of a sequence. For models originally designed for DTI prediction (like TransformerCPI and MMDG-DTI), we modify their prediction heads to output a single value with MSE loss for regression.

For prediction using molecules and assay descriptions, we compare against Galactica (Taylor et al., 2022), KV-PLM (Zeng et al., 2022), CLAMP (Seidl et al., 2023), BioT5+ (Pei et al., 2024), and TwinBooster (Schuh et al., 2024), with random prediction and nearest-assay transfer (1-NN and Soft 10-NN) as additional baselines. Few-shot evaluation additionally includes ActFound (Feng et al., 2024), FS-CAP (Eckmann et al., 2024), and AdaMBind (Wan et al., 2026) to compare methods designed for assays with limited measurements. Baseline details are provided in Appendix C.

**Metrics.** Considering the **non-standardized measurement units** across assays, we evaluate methods based on their ability to *rank drug molecules within each assay*, using two widely adopted metrics: (1) Pearson correlation coefficient and (2) Concordance Index (CI). Given the substantial variation in the number of drug molecules across assays, we compute weighted versions of both metrics, with weights proportional to the number of drug molecules in each assay. These weighted metrics are denoted as *w*-Pearson and *w*-CI, respectively. More implementation details are provided in Appendix D.

### 4.2 Zero-Shot Transfer (RQ1)

To evaluate our model's generalization ability in real-world scenarios where experimentally measured drug molecules may be unavailable, we investigate two zero-shot prediction settings: (1) Target-drug setting and (2) Text-drug setting.

**Target-drug Setting.** In this setting, we evaluate model performance when only target protein information is available for new assays. where $A_{\text{test}} = \{p_{\text{new}}, \mathcal{M}_{\text{test}}\}$. We evaluate on Davis and KIBA datasets using two 8:2 train-test target splits: (1) "random" split, where target proteins are randomly divided, and (2) "target sim" split, where target proteins are clustered by sequence similarity using the easy_cluster algorithm from MMseqs2 with commonly adopted values for minimum sequence identity (0.3) and minimum coverage (0.5) (Steinegger & Söding, 2018; 2017). For each split, 10% of the training partition is reserved for validation. All methods are first pre-trained in $\text{ChEMBL}_{\text{pre-train}}$.

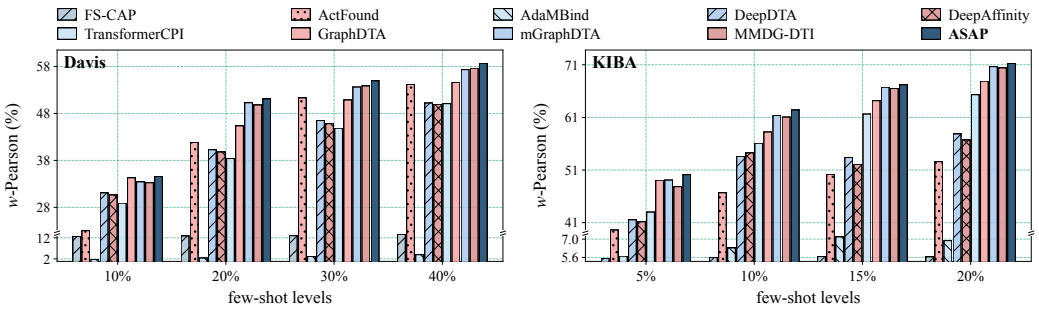



Figure 3: Few-shot results w.r.t *w*-Pearson on Davis (left) and KIBA (right) datasets.

As shown in Table 2, (1) ASAP achieves best performance w.r.t *w*-Pearson and *w*-CI across most scenarios, except against MMDG-DTI on KIBA's "target sim" split. This exception may be attributed to a higher average number of drug molecule pairs per assay in KIBA (avg. $\frac{\#pair}{\#assay}$), which allows models to learn activity patterns more thoroughly. (2) All methods exhibit performance degradation on "target sim" split compared to the "random" split. This decline is expected as the "target sim" split evaluates the generalization to proteins with distinct sequence patterns from training data, presenting a more realistic but challenging scenario. Notably, our model shows stronger relative performance improvements in the "target sim" setting compared to the "random" split on KIBA, demonstrating the effectiveness of our approach for capturing the collaborative information and utilizing this information to help zero-shot prediction for new assays with "unfamiliar" target proteins.

**Text-drug Setting.** In this setting, we assess the model's capability when only assay descriptions are provided. We conduct experiments on the ChEMBL test dataset since Davis and KIBA do not provide textual assay descriptions. For comparison, alongside the pre-trained large models introduced earlier, wefollow Seidl et al. (2023) to implement two other baselines: 1-NN and Soft 10-NN. These methods leverage textual assay descriptions to identify the most similar assay (1-NN) or top 10 similar assays (Soft 10-NN) from the training set, and use their predictions for the given test assay. Detailed implementation specifications are provided in Appendix C. Based on the results presented in Table 3, we observe that: (1) Our ASAP outperforms all baselines on both *w*-Pearson and *w*-CI metrics, achieving 8.50% and 1.89% performance improvement over the second-best method. These results demonstrate our model's superior ability to leverage textual assay descriptions for identifying collaborative assays to assist zero-shot transfer. (2) Scientific language models (GAL 125M, KV-PLM, and BioT5+) remain below the nearest-assay baselines, CLAMP and TwinBooster. This suggests that current scientific language models have yet to fully address the challenges in handling complex regression tasks like activity prediction, which require a precise understanding of molecular structures and chemical properties.

Table 3: Zero-shot text-drug results (in %) of different methods on ChEMBL dataset.

| Dataset | ChEMBL | |
|---|---|---|
| Metric | *w*-Pearson | *w*-CI |
| Random | -0.06±0.12 | 50.03±0.04 |
| 1-NN | 20.86±0.78 | 57.15±0.31 |
| Soft 10-NN | 22.03±0.64 | 57.10±0.27 |
| GAL 125M | 16.65±0.13 | 55.91±0.04 |
| KV-PLM | 16.27±0.24 | 56.35±0.05 |
| CLAMP | 26.64±0.73 | 59.89±0.30 |
| BioT5+ | 12.23±0.14 | 54.04±0.06 |
| TwinBooster | 27.77±0.16 | 59.21±0.08 |
| **ASAP** | 30.13±0.18 | 61.02±0.08 |

### 4.3 Few-Shot Transfer (RQ2)

To evaluate our model's few-shot transferability, we conduct experiments on Davis and KIBA datasets. Given the smaller number of measurements per assay in Davis, we use 10-40% few-shot levels for Davis and 5-20% for KIBA. Due to space constraints, we present the *w*-Pearson results in Figure 3, while the CI results are provided in Appendix Figure 6. From the results, we observe: (1) Our model achieves strong overall performance across the evaluated few-shot levels, supporting its ability to adapt using limited target-assay measurements. (2) The baselines designed for few-shot learning, including FS-CAP, ActFound, and AdaMBind, remain less competitive under our cross-dataset transfer protocol. This may partly reflect the difficulty of adapting to shifts in assay distributions and activity scales with limited support data. (3) While MMDG-DTI using pre-trained components (Prot-BERT (Elnaggar et al., 2022) and SMILES-BERT (Wang et al., 2019)) shows benefits, our method surpasses them by capturing collaborative information between assays, leveraging shared knowledge for better cross-assay transfer.

### 4.4 Pre-training Study and Ablation (RQ3)

To evaluate the effectiveness of our collaborative-based pre-training strategy, we conduct comprehensive comparison and ablation studies on the pre-training phase using the ChEMBL test dataset. (1) In model comparison (upper part in Table 4): our ASAP outperforms all DTA prediction methods on both metrics, showing 52.61% and 8.99% improvements over second-best MMDG-DTI. These results demonstrate that our ranking-based activity training (*stage 2*) effectively leverages collaborative information to enhance performance. (2) For ablation studies (lower part in Table 4), we analyze several variants: "ASAP w/o $c$" omits collaborative-based pre-training in *stage 1*; "ASAP$_{\text{mse}}$" optimizes the collaborative-based pre-training by MSE loss instead of contrastive margin loss. We can see that collaborative-based pre-training significantly contributes to the final results.

Table 4: Performance comparison and ablation study results (in %) on ChEMBL test dataset.

| | Dataset | ChEMBL | |
|---|---|---|---|
| | Metric | $w$-Pearson | $w$-CI |
| Method | DeepDTA | 21.52±0.42 | 55.90±0.31 |
| | DeepAffinity | 19.37±0.53 | 55.40±0.25 |
| | TransformerCPI | 20.94±0.36 | 56.83±0.21 |
| | GraphDTA | 21.33±0.66 | 56.88±0.35 |
| | mGraphDTA | 21.88±0.28 | 56.86±0.17 |
| | AttentionMGT-DTA | 21.85±0.19 | 56.38±0.11 |
| | MMDG-DTI | 22.17±0.34 | 57.04±0.19 |
| Variant | ASAP w/o $c$ | 27.48±0.27 | 58.48±0.16 |
| | ASAP$_{\text{mse}}$ | 30.26±0.27 | 61.03±0.24 |
| | ASAP | **33.83±0.21** | **62.17±0.17** |

## 5 Related Works

**Drug-target Affinity Prediction.** ASAP is closely related to DTA methods, which usually evolve into two main categories: sequence-based methods that process raw sequences and graph-based approaches that utilize molecular structural information. Sequence-based methods like DeepDTA (Öztürk et al., 2018), DeepAffinity (Karimi et al., 2019; 2020), and TransformerCPI (Chen et al., 2020) employ CNNs, RNNs, and Transformers to extract features from drug SMILES and protein sequences. However, sequence-based methods often fail to capture complex structural information and spatial relationships between molecules. Graph-based methods like GraphDTA (Nguyen et al., 2021), mGraphDTA (Yang et al., 2022), and MMDG-DTI (Hua et al., 2025) address this by using GNNs to model molecular topology while preserving atom-level information. Li & Yao (2025) further improves feature extraction by introducing hypergraph representations and employing cross-attention mechanisms. Recent approaches improve cold-start generalization through pretrained representations and task-adaptive meta-learning (Ouyang et al., 2025; Wan et al., 2026). These methods focus on target-drug pairs; ASAP also predicts activity from assay descriptions without target inputs.

**Assay-aware and Low-data Activity Prediction.** CLAMP (Seidl et al., 2023) and TwinBooster (Schuh et al., 2024) combine molecular representations with assay descriptions for zero-shot prediction, with CLAMP using activity-supervised contrastive alignment. Assay-aware proteochemometric models further account for measurement variation associated with experimental context (Schoenmaker et al., 2025). For few-shot prediction, ActFound combines within-assay activity differences with meta-learning to handle heterogeneous measurements (Feng et al., 2024). Closely related, AssayMatch learns assay compatibility from data attribution and uses assay text to select training data for unlabeled test assays (Fan & Barzilay, 2026). ASAP learns collaborative similarity from cross-assay predictive transfer without requiring shared molecules, and uses it to pre-train embeddings of assay descriptions and protein sequences. Listwise ranking within each assay then reduces sensitivity to measurement scales, supporting zero- and few-shot transfer from either modality.

## 6 Conclusion

In this work, we introduce ASAP, a novel framework for bioactivity prediction that addresses key challenges in drug discovery, including inductive generalization, poorly characterized assays, and effective utilization of multi-assay data. By leveraging model-driven assay similarity and ranking loss, our method learns assay embeddings that capture intrinsic binding patterns, enabling accurate predictions for new assays. The proposed model-driven similarity metric provides a flexible and generalizable way to model assay relationships, even without target information. ASAP shows its effectiveness and generalizability through extensive experiments in zero-shot and few-shot settings.

## AI USE STATEMENT

We used generative AI tools for manuscript revision, literature search, figure preparation, and experimental result analysis. The authors are responsible for verifying AI-assisted content against original sources and experimental evidence and take full responsibility for the final text, claims, figures, and artifacts.

## ETHICS STATEMENT

This work studies computational bioactivity prediction using existing data from ChEMBL, Davis, and KIBA. The reported experiments evaluate predictive performance rather than establishing the safety or clinical efficacy of any compound. Predictions may be affected by biases in assay coverage, experimental conditions, and the distribution of measured compounds and targets. They should therefore be used to prioritize hypotheses for further experimental validation, rather than as a substitute for biological experiments or clinical assessment. Potential misuse of bioactivity prediction to support harmful compound selection should also be considered in downstream applications.

## REPRODUCIBILITY STATEMENT

The two-stage training procedure and optimization objectives are described in Sections 3.1 and 3.2. Section 4.1 specifies the datasets, data splits, pre-training–evaluation overlap filtering, and evaluation metrics. Data curation and baseline implementations are described in Appendices B.1 and C, respectively. Appendix D reports the training configuration, hardware, repeated runs, and validation-based checkpoint selection. Additional robustness analyses and experimental results are provided in Appendix E. To facilitate reproduction, we provide an anonymous code repository at `https://anonymous.4open.science/r/ASAP-04C8/`.

## REFERENCES

Ryuto Abe, Tomoyuki Miyao, and Jürgen Bajorath. Accurate prediction of activity cliff compounds based on bioactivity profiles depends on assay nearest neighbor relationships. *Journal of Cheminformatics*, 18(1):72, 2026. doi: 10.1186/s13321-026-01210-9.

Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 2020.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

John G Cumming, Andrew M Davis, Sorel Muresan, Markus Haeberlein, and Hongming Chen. Chemical predictive modelling to improve compound quality. *Nature reviews Drug discovery*, 12 (12):948–962, 2013.

Alperen Dalkıran, Ahmet Atakan, Ahmet S Rifaioğlu, Maria J Martin, Rengül Çetin Atalay, Aybar C Acar, Tunca Doğan, and Volkan Atalay. Transfer learning for drug–target interaction prediction. *Bioinformatics*, 39(Supplement_1):i103–i110, 2023.

Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.

Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics*, 47:20–33, 2016.

Peter Eckmann, Jake Anderson, Rose Yu, and Michael K Gilson. Ligand-based compound activity prediction via few-shot learning. *Journal of Chemical Information and Modeling*, 64(14):5492–5499, 2024.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022.

Vincent Fan and Regina Barzilay. Assaymatch: Learning to select data for molecular activity models. *Journal of Chemical Information and Modeling*, 66(7):3540–3549, 2026.

Bin Feng, Zequn Liu, Nanlan Huang, Zhiping Xiao, Haomiao Zhang, Srbuhi Mirzoyan, Hanwen Xu, Jiaran Hao, Yinghui Xu, Ming Zhang, et al. A bioactivity foundation model using pairwise meta-learning. *Nature Machine Intelligence*, 6(8):962–974, 2024.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

Yang Hua, Zhenhua Feng, Xiaoning Song, Xiao-Jun Wu, and Josef Kittler. Mmdg-dti: Drug–target interaction prediction via multimodal feature fusion and domain generalization. *Pattern Recognition*, 157:110887, 2025.

José Jiménez, Miha Skalic, Gerard Martinez-Rosell, and Gianni De Fabritiis. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling*, 2018.

Yogesh Kalakoti, Shashank Yadav, and Durai Sundar. Deep neural network-assisted drug recommendation systems for identifying potential drug–target interactions. *ACS omega*, 7(14):12138–12146, 2022.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Explainable deep relational networks for predicting compound–protein affinities and contacts. *Journal of chemical information and modeling*, 61(1):46–66, 2020.

Michael J Keiser, Bryan L Roth, Blaine N Armbruster, Paul Ernsberger, John J Irwin, and Brian K Shoichet. Relating protein pharmacology by ligand chemistry. *Nature biotechnology*, 25(2): 197–206, 2007.

Jiannuo Li and Lan Yao. Hcaf-dta: drug-target binding affinity prediction with cross-attention fused hypergraph neural networks. *arXiv preprint arXiv:2504.02014*, 2025.

Mei Li, Xiangrui Cai, Sihan Xu, and Hua Ji. Metapath-aggregated heterogeneous graph neural network for drug–target interaction prediction. *Briefings in Bioinformatics*, 24(1):bbac578, 2023.

Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In *ACM SIGKDD*, pp. 975–985, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Zijing Liu, Xianbin Ye, Xiaoming Fang, Fan Wang, Hua Wu, and Haifeng Wang. Docking-based virtual screening with multi-task learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 381–385. IEEE, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications*, 2017.

Chanin Nantasenamat, Chartchalerm Isarankura-Na-Ayudhya, and Virapong Prachayasittikul. Advances in computational methods to predict the biological activity of compounds. *Expert opinion on drug discovery*, 5(7):633–654, 2010.

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37 (8):1140–1147, 2021.

Xike Ouyang, Yannuo Feng, Chen Cui, Yunhe Li, Li Zhang, and Han Wang. Improving generalizability of drug–target binding prediction by pre-trained multi-view molecular representations. *Bioinformatics*, 41(1):btaf002, 2025. doi: 10.1093/bioinformatics/btaf002.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *arXiv preprint arXiv:2402.17810*, 2024.

Roger Perkins, Hong Fang, and Weida Tong. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environmental Toxicology and Chemistry: An International Journal*, pp. 1666–1679, 2003.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

RN Sahoo, S Pattanaik, G Pattnaik, S Mallick, and R Mohapatra. Review on the use of molecular docking as the first line tool in drug discovery and development. *Indian Journal of Pharmaceutical Sciences*, 84(5), 2022.

Johannes Schimunek, Sohvi Luukkonen, and Günter Klambauer. MHNfs: Prompting in-context bioactivity predictions for low-data drug discovery. *Journal of Chemical Information and Modeling*, 65(9):4243–4250, 2025. doi: 10.1021/acs.jcim.4c02373.

Linde Schoenmaker, Enzo G Sastrokarijo, Laura H Heitman, Joost B Beltman, Willem Jespers, and Gerard JP van Westen. Toward assay-aware bioactivity model (er) s: getting a grip on biological context. *Journal of Chemical Information and Modeling*, 65(13):7013, 2025.

Maximilian G. Schuh, Davide Boldini, and Stephan A. Sieber. Synergizing chemical structures and bioassay descriptions for enhanced molecular property prediction in drug discovery. *Journal of Chemical Information and Modeling*, 64(12):4640–4650, 2024. doi: 10.1021/acs.jcim.4c00765.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International Conference on Machine Learning*, pp. 30458–30490, 2023.

Aylin Sertkaya, Trinidad Beleche, Amber Jessup, and Benjamin D Sommers. Costs of drug development and research and development intensity in the US, 2000-2018. *JAMA network open*, 7(6): e2415445–e2415445, 2024.

Jeongtae Son and Dongsup Kim. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. *PloS one*, 16(4):e0249404, 2021.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):2542, 2018.

Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062, 2022.

Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

Xiwei Tang, Yiqiang Zhou, Mengyun Yang, and Wenjun Li. Tc-dta: predicting drug–target binding affinity with transformer and convolutional neural networks. *IEEE Transactions on NanoBioscience*, 2024.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Wen Torng and Russ B Altman. Graph convolutional neural networks for predicting drug-target interactions. *Journal of chemical information and modeling*, 59(10):4131–4149, 2019.

Fabien Vincent, Arsenio Nueda, Jonathan Lee, Monica Schenone, Marco Prunotto, and Mark Mercola. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, pp. 899–914, 2022.

Mengxuan Wan, Yanpeng Zhao, Yixin Zhang, Huiyan Xu, Duoyun Yi, Peng Zan, Song He, and Xiaochen Bo. A meta learning and task adaptive approach for drug target affinity prediction. *Nature Communications*, 17(1):3734, 2026. doi: 10.1038/s41467-026-70554-5.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 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*, pp. 429–436, 2019.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *ICML*, pp. 1192–1199, 2008.

Junhao Xue, Jun Liu, and Kai Chen. Context-informed few-shot molecular property prediction via heterogeneous meta-learning. *Computational and Structural Biotechnology Journal*, 27: 4173–4182, 2025. doi: 10.1016/j.csbj.2025.08.016.

Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical science*, 13 (3):816–833, 2022.

Qing Ye, Chang-Yu Hsieh, Yu Kang, Shibo He, and Tingjun Hou. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature communications*, 12(1):6775, 2021.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022.

Xiaodong Zheng, Hao Ding, Hiroshi Mamitsuka, and Shanfeng Zhu. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *ACM SIGKDD*, pp. 1025–1033, 2013.

# A More Related Works

## A.1 Pre-training and Additional Low-data Approaches

Galactica (Taylor et al., 2022) and KV-PLM (Zeng et al., 2022) learn representations from scientific text and molecular data through self-supervised pre-training. MHNfs (Schimunek et al., 2025) and CFS-HML (Xue et al., 2025) use labeled support sets for few-shot bioactivity and molecular property prediction, respectively. Complementary evidence from Abe et al. (2026) shows that accurate activity-cliff prediction from bioactivity profiles can depend on one or a few assays nearest to the test assay. This finding supports examining assay relationships, while its specific setting is activity-cliff classification using observed bioactivity profiles.

## A.2 Recommender Systems for Activity Prediction.

Research incorporating collaborative information for activity prediction has evolved. DTINet (Luo et al., 2017) pioneered heterogeneous networks integrating diverse drug information. Building upon this, KGE_NFM (Ye et al., 2021) combined knowledge graphs with recommendation systems to address cold start problems. dNNDR (Kalakoti et al., 2022) developed multiclass classification incorporating molecular descriptors through interpretable embeddings, while MHGNN (Li et al., 2023) introduced metapath-aggregated neural networks capturing complex biological relationships. These works demonstrate the promising potential (Zheng et al., 2013) of assay-aware interaction modeling in bioactivity prediction, particularly in effectively handling complex heterogeneous data and capturing intricate relationships between drug molecules and target proteins.

# B Datasets

## B.1 Data Curation.

We curated our pre-training dataset from ChEMBL, which is a comprehensive database of bioactivity information derived from sources such as scientific articles, patents, and other datasets. It serves as a key resource for drug discovery and development, encompassing 2.5 million distinct compounds and 19 million measured activity records. Due to variations in experimental conditions, affinities from different assays are not directly comparable. To address this, we organized the experimental affinity values into distinct assays. Specifically, we grouped the corresponding data by assay ID. Only data with molar or density concentration units (e.g., nmol, $\mu g/ml$) were retained to ensure consistency, as other units were ambiguous and excluded. Additionally, we excluded entries where the standard relation was not "=" since such values lacked specific quantitative activity information. Inorganic compounds and large molecules (molecular weights exceeding 1,000) were also removed. We retain assays with molecular counts between 20 and 10,000.

## B.2 Assay Description.

The assay descriptions usually use very compact language and contain terminology from molecular biology or chemistry and, thus, are very dissimilar from the language on which language models are usually trained. Figure 4 shows a Wordcloud of assay descriptions for ChEMBL.

# C Baselines

**DeepDTA (Öztürk et al., 2018).** DeepDTA predicts protein–ligand binding affinities using one-dimensional sequence representations of proteins and ligands, specifically raw protein sequences and SMILES strings. The model employs two separate CNN blocks, each containing multiple 1D convolutional layers followed by max-pooling, to extract features from these inputs. The extracted features are concatenated and passed through fully connected layers to perform binding affinity prediction as a regression task.

**DeepAffinity (Karimi et al., 2019; 2020).** DeepAffinity uses bidirectional RNNs combined with CNNs to predict protein–ligand interactions. Proteins are encoded with structural and physicochem-




Figure 4: Wordcloud of assay descriptions of ChEMBL.

ical alphabets, while compounds are represented as SMILES strings. Pre-trained RNNs capture long-term sequence dependencies, and attention mechanisms enhance interpretability by identifying key molecular fragments contributing to binding interactions.

**TransformerCPI (Chen et al., 2020).** TransformerCPI leverages a transformer architecture to predict compound–protein interactions. Protein sequences are embedded with Word2Vec, while drug molecules are modeled as graphs and processed using graph convolutional networks (GCNs) to extract atom-level features. The transformer decoder integrates these representations using self-attention layers, adapting the transformer for interaction-specific feature learning.

**GraphDTA (Nguyen et al., 2021).** GraphDTA combines GNNs and CNNs for drug–target affinity prediction. Drugs are modeled as molecular graphs derived from SMILES strings, and proteins are represented as one-hot-encoded sequences. A GNN processes the drug graphs to capture structural features, while 1D CNNs extract sequence-level features from proteins. These representations are integrated to predict binding affinities.

**mGraphDTA (Yang et al., 2022).** mGraphDTA uses a multiscale graph neural network (MGNN) with 27 graph convolutional layers to extract multiscale drug features and a multiscale convolutional neural network (MCNN) to process protein sequences. The extracted features are fused for each modality, concatenated, and passed through a multilayer perceptron (MLP) for affinity prediction.

**ActFound (Feng et al., 2024).** ActFound is a target-free few-shot model for bioactivity prediction across heterogeneous assays. Compounds are represented by Morgan fingerprints, and a Siamese network predicts pairwise activity differences between compounds measured within the same assay, avoiding direct comparison of incompatible assay values. A MAML-based meta-learning procedure enables rapid adaptation to a new assay using a small set of measured compounds. The predicted differences are combined with these reference measurements to estimate the absolute activities of unseen compounds. We retrained the model on our ChEMBL using 2048-dimensional fingerprints, two hidden layers of width 2048 with batch normalization and LeakyReLU, and a bias-free scalar output layer. Native evaluation used up to five task-specific updates; assays with at most five support records were evaluated without local gradient updates. One-shot assays used a single-reference extension with no gradient updates.

**FS-CAP (Eckmann et al., 2024).** FS-CAP performs few-shot compound activity prediction for unseen assays using a small set of measured compounds as context. Each context compound is represented by its Morgan fingerprint multiplied by its activity value, encoded independently, and aggregated by mean pooling. The resulting context representation is concatenated with a separately encoded query compound and passed to an MLP to predict its continuous activity. We retrained the released architecture on ChEMBL using 2048-bit Morgan fingerprints with radius 3 and two independent MLP encoders with 512-dimensional outputs. Each training query was paired with eight distinct support records from the same assay, excluding the query record.

Table 5: Hyperparameter settings for our experiments.

| Hyperparameter | Value |
|---|---|
| Seed | 1,2,3 (three runs) |
| Learning rate | 0.0001 |
| Epoch in Stage1 | 32 |
| Epoch in Stage2 | 32 |
| Batch size in Stage1 | 64 |
| Batch size in Stage2 | 1024 |
| Dimension of ECFP features | 1024 |
| Layer of assay encoder (MLP) | 2 |
| Layer of molecule encoder (MLP) | 2 |
| Output dim. of assay encoder (MLP) | 512 |
| Output dim. of molecule encoder (MLP) | 512 |
| Margin $\delta$ in contrastive margin loss | 1.0 |
| Steepness $\beta$ in sigmoid-based weighting | 5.0 |

**MMDG-DTI (Hua et al., 2025).** MMDG-DTI combines semantic and structural features for drug–target interaction prediction. Semantic features are extracted using two BERT-based models, Prot-Bert for proteins and Smiles-Bert for drugs, trained on large biological datasets. Structural features are captured using a hybrid GNN, which models relationships between atoms in drugs and residues in proteins. The features are fused using a classifier that incorporates domain adversarial training and contrastive learning to enhance prediction accuracy and generalization.

**AdaMBind (Wan et al., 2026).** AdaMBind addresses few-shot drug-target affinity prediction through task-adaptive meta-learning. A molecular graph encoder and a protein sequence encoder provide representations for an affinity predictor, whose shared initialization is adapted to new targets using labeled support examples. A learned scheduler uses query losses and support-query gradient agreement to select meta-training tasks, while uniform label noise provides regularization. We implemented a first-order version adapted to our assay-level transfer protocol, treating each ChEMBL assay as a task and using a GAT-GCN molecular encoder with the released single-convolution protein-encoding architecture. Training used 1000 meta-updates, with 15 candidate tasks, eight task selections with replacement, and five full-support Adam updates per selected task.

**Galactica (Taylor et al., 2022).** The large-scale scientific language model Galactica (GAL) is specifically designed to process scientific text, equations, chemical formulas, and references within one unified framework. We initialize the model from the public 125M checkpoint and adapt it to continuous activity prediction by adding a linear scalar head to the 768-dimensional hidden state of the final non-padding input token. The backbone and prediction head are jointly fine-tuned on $\text{ChEMBL}_{pre-train}$. The input contains the assay description and molecular SMILES, formatted using the template below and the authors' scientific-sequence preprocessing.

**Prompt**: `"Question: Here is a SMILES formula: [START_SMILES]" + SMILES + "[END_SMILES]. What will be the binding affinity of this chemical molecule in the assay:" + ASSAY_DESCRIPTION + "?\n\nAnswer:".`

**KV-PLM (Zeng et al., 2022).** The KV-PLM is a unified pretrained language model that is able to process both molecule structures and biomedical text, enabling for knowledgeable and versatile machine reading. It adopts the popular pretrained language model BERT (Devlin et al., 2019) as its backbone and is pretrained using the masked language modeling task to learn contextual representations of input sequences. In this work, we use the cosine similarity between SMILES inputs and assay descriptions to rank molecules within an assay.

**CLAMP (Seidl et al., 2023).** CLAMP is a model for text-based activity prediction with separate molecule and text encoders, trained with an activity-supervised cross-modal contrastive objective. It learns a shared embedding space from molecule–assay pairs labeled as active or inactive, aligning

active molecules with the corresponding assay descriptions. This design supports zero-shot activity prediction for unseen assays from their textual descriptions, without task-specific labeled examples or fine-tuning. For our experiments, we initialize from the released CLAMP checkpoint, which combines 8,192-dimensional Morgan/RDKit count fingerprints with 512-dimensional CLIP text features through two projection MLPs with hidden widths of 4,096 and 2,048 and an output dimension of 768. We freeze the CLIP text encoder and fine-tune both projection MLPs with an assay-wise ListMLE objective on continuous activity labels.

**1-NN and Soft 10-NN (Seidl et al., 2023).** These baselines transfer predictions from training assays to an unseen assay using the similarity of their textual descriptions. Both use predictions from a multi-task deep neural network (MT-DNN) trained on the training assays. Given a test assay, we compute the cosine similarity between its text embedding and those of all training assays. For 1-NN, the prediction for each query molecule is the MT-DNN output corresponding to the most similar training assay. For Soft 10-NN, we implement a top-10 variant of the soft-NN baseline: we select the ten training assays with the highest cosine similarities and apply softmax normalization over these ten similarities to obtain weights that sum to one. The final prediction for each query molecule is the weighted average of its MT-DNN predictions for the ten selected assays. We use PubMedBERT (Gu et al., 2021) to encode assay descriptions and ChemBERTa (Chithrananda et al., 2020) to encode molecules, matching the encoders used in our approach.

**BioT5+ (Pei et al., 2024).** BioT5+ is an encoder-decoder model based on the T5 (Raffel et al., 2020) architecture, explicitly tailored to bridge the gap between molecular representations and natural language in the biological domain. Unlike previous models that rely on SMILES or SELFIES, BioT5+ integrates IUPAC names into its input representation, fostering a deeper alignment between chemical structures and their textual descriptions found in scientific literature. Pre-trained on a vast corpus of biological literature and sequences, BioT5+ utilizes multi-task instruction tuning to generalize across diverse tasks. This approach allows the model to capture intricate relationships inherent in biological data, facilitating tasks such as molecule property prediction and description generation.

**TwinBooster (Schuh et al., 2024).** TwinBooster combines molecular fingerprints and assay descriptions through Barlow Twins representation learning, followed by gradient boosting for activity prediction. Its released encoders map 1,024-bit ECFP fingerprints with radius 2 and 768-dimensional PubChemDeBERTa text representations to separate 1,024-dimensional embeddings. For our continuous-label task, we freeze PubChemDeBERTa and both pretrained encoders, concatenate their L2-normalized outputs, and train a LightGBM regressor on the resulting 2,048-dimensional features using squared-error loss. Training labels are standardized separately within each training assay using its mean and population standard deviation; validation and test label statistics are not used for normalization or prediction rescaling.

## D Implementation Details

All experiments were conducted on a Linux server equipped with 8 NVIDIA A6000 GPUs. The model is optimized using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 0.0001. The hidden dimension of assay and molecule embeddings is set to 512. For the first stage collaborative pre-training, we train for 32 epochs with a batch size of 64 assays. For the second stage ranking-based bioactivity training, we train for 32 epochs with a batch size of 1024 (containing 64 assays, each including 16 drug molecules). Each experiment was run three times with different random seeds for robustness.

Regarding model selection, we did not perform extensive hyperparameter tuning for our model. Instead, we followed the architecture design principles from ESM2 (Lin et al., 2023) and adopted a fixed set of hyperparameters throughout our experiments. The checkpoint with the highest $w$-Pearson on the internal validation subset is selected. All trained baselines follow the same validation-based selection criterion.

## E ADDITIONAL RESULTS

**Robustness Analysis of Collaborative Similarity.** To evaluate the stability of our model-driven similarity metric, we conducted comprehensive robustness analyses examining its sensitivity to different fingerprint choices and SVR hyperparameters. We employed Mantel tests to compare our default similarity matrix (ECFP fingerprints with SVR regularization parameter $C = 1.0$) against matrices computed with alternative fingerprint representations (MACCS) and varying regularization parameters. Since the SVR serves as a proxy to mine *relative relationships* rather than achieve absolute precision, capturing general trends is sufficient for our collaborative learning framework.

Table 6: Mantel tests of default similarity matrix (ECFP, $C$=1.0) vs. varying SVR hyperparameters and fingerprint choices. $^*$ denotes statistically significant correlations ($p < 0.01$).

| | ECFP | | | | MACCS | | | | Random |
|---|---|---|---|---|---|---|---|---|---|
| $C$ value | 0.1 | 1.0 | 10 | 100 | 0.1 | 1.0 | 10 | 100 | |
| Spearman | 0.91* | 1.00* | 0.93* | 0.92* | 0.47* | 0.49* | 0.45* | 0.44* | 0.002 |

As shown in Table 6, the collaborative signal demonstrates remarkable robustness across different configurations. With ECFP fingerprints, the method maintains *high correlations* ($> 0.91$) across different $C$ values tested. Even when using MACCS fingerprints, as a simpler representation, correlations remain *statistically significant* ($p < 0.01$) and substantially higher than the random noise baseline. This consistency validates that our similarity extraction approach is stable and reliable across different hyperparameter choices and fingerprint representations.

**Interpretability of Learned Assay Embeddings.** To provide deeper insights into the learned assay embedding space beyond the single case study presented in Section 3.1.1, we conducted additional interpretability analyses. We performed t-SNE visualization of the learned assay embeddings from our collaborative pre-training stage, with assays colored by their L1 target classification in ChEMBL (e.g., Enzymes, Membrane receptors, Ion channels, etc.).

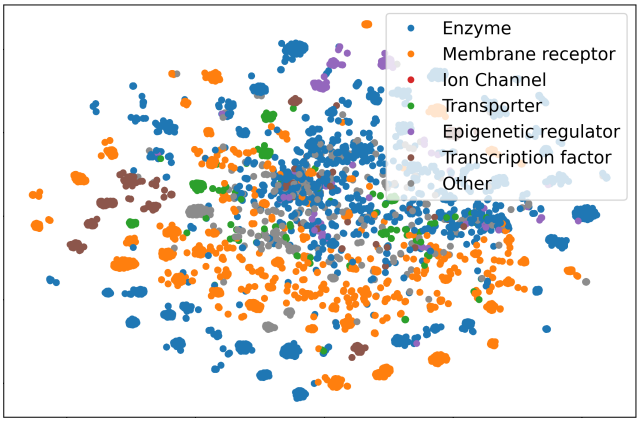



Figure 5: t-SNE visualization of learned assay embeddings colored by protein target classes.

Figure 5 exhibits clear local groupings aligned with protein classes, confirming that our embedding space captures broad biological commonalities while preserving fine-grained assay-specific characteristics (e.g., drug activity rankings) within each class. This provides confirmation that the collaborative pre-training successfully learns biologically meaningful representations that group functionally related assays together.

**Supplementary Few-shot CI Results.** Due to space constraints, the main body of the paper (Sec. 4.3) presented the $w$-Pearson correlation results for the few-shot learning experiments in Figure 3. As a supplement, here Figure 6 reports the $w$-CI scores for the same few-shot transferability experiments on Davis and KIBA. ASAP achieves the highest $w$-CI scores at all evaluated few-shot levels on Davis and at three of the four levels on KIBA. At the 15% level on KIBA, its score is slightly lower than those of mGraphDTA and MMDG-DTI. Overall, these results further support the effectiveness of ASAP in few-shot transfer.

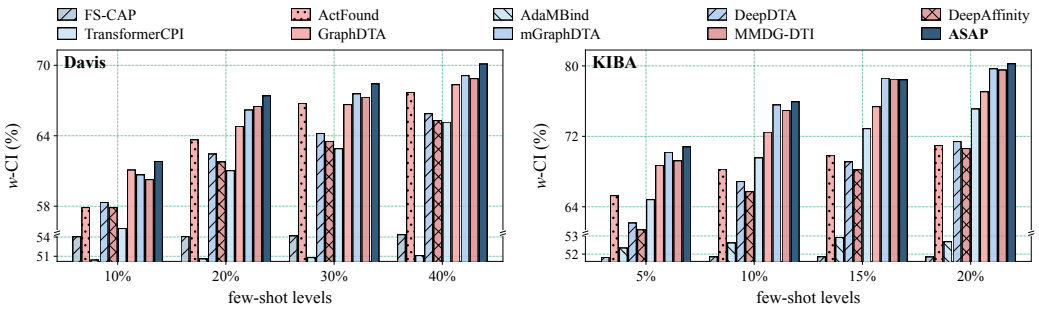



Figure 6: Few-shot results w.r.t *w*-CI on Davis (left) and KIBA (right) datasets.

**Support for Collaborative Assumption.** Our central assumption that assays with similar drug activity rankings share transferable predictive signals is supported both theoretically and empirically:

- **Theoretical Foundation:** Our hypothesis aligns with established principles from (1) the Similarity Ensemble Approach (SEA) (Keiser et al., 2007), which demonstrated that targets exhibiting similar ligand binding patterns share underlying pharmacological profiles, and (2) Collaborative Matrix Factorization (Zheng et al., 2013), which assumes that drug-target interaction matrices have low-rank structures, meaning assays with similar interaction patterns share latent features.
- **Empirical Validation:** Beyond our case study in Section 3.1.1 showing functional overlaps between 5-HT1a and Dopamine D2 receptors, the t-SNE visualization in Figure 5 provides additional evidence that our learned assay embeddings capture meaningful biological relationships, with clear clustering patterns corresponding to protein family classifications.

These analyses collectively demonstrate that our model-driven collaborative approach captures meaningful biological relationships and produces robust similarity signals that are stable across different parameter choices and fingerprint representations.

## F LIMITATIONS

While ASAP demonstrates strong performance in zero-shot and few-shot bioactivity prediction, several limitations remain. First, the proposed model-driven assay similarity metric assumes linear correlations (via Pearson coefficient) between predicted and actual bioactivities, which may not fully capture complex, non-linear relationships across heterogeneous experimental protocols.

Second, while the ranking-based loss mitigates batch effects, residual biases from divergent experimental conditions may still influence predictions. The current assay similarity matrix mostly captures the information of the target protein. Other modalities, such as cell types, species, or tissues, can be further incorporated to better model the assay conditions.