# Multimodal Out-of-Distribution Individual Uncertainty Quantification Enhances Binding Affinity Prediction for Polypharmacology

Amitesh Badkul[1], Li Xie[2], Shuo Zhang[2,3], Lei Xie[1,3,4*]

[1]Ph.D. Programs in Computer Science, The Graduate Center, The City University of New York, New York City, NY, 10016, U.S.A.
[2]Department of Computer Science, Hunter College, The City University of New York, New York City, NY, 10065, U.S.A.
[3]Helen and Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, New York City, NY, 02115, U.S.A.
[4]School of Pharmacy and Pharmaceutical Sciences & Center for Drug Discovery, Northeastern University, Boston, MA, 10065, U.S.A.

*Corresponding author(s). E-mail(s): lxie@iscb.org;

**Abstract**

Polypharmacology, a single drug that targets multiple proteins, holds promise for addressing unmet medical needs. Achieving accurate, reliable, and scalable predictions of protein-ligand binding affinity across multiple proteins is crucial to realizing the potential of polypharmacology. Machine learning offers a powerful tool for multi-target binding affinity prediction. However, three major challenges remain: generalizing predictions to out-of-distribution compounds that are structurally different from those in the training data, quantifying the uncertainty of predictions in out-of-distribution scenarios where the assumption underlying existing methods does not hold, and scaling to billions of compounds, which remains unattainable for current structure-based methods. To overcome these challenges, we propose a model-agnostic anomaly detection-based individual uncertainty quantification method: **e**mbedding **M**ahalanobis **O**utlier **S**coring and **A**nomaly **I**dentification via **C**lustering (eMOSAIC). eMOSAIC features the divergence between the multi-modal representations of known cases and unseen instances and quantifies individual prediction uncertainty on a compound-by-compound basis. We integrate eMOSAIC with a multi-modal

deep neural network for multi-target ligand binding affinity predictions, leveraging a structure-informed large protein language model. Comprehensive validation in out-of-distribution settings demonstrates that eMOSAIC significantly outperforms state-of-the-art sequence-based and structure-based methods as well as existing uncertainty quantification approaches. These findings underscore eMOSAIC's potential to advance real-world polypharmacology and other applications that require robust predictions and scalable solutions.

**Keywords:** machine learning, deep learning, transfer learning, trustworthy ML, protein language model, protein-ligand docking, compound screening, drug discovery

# 1 Introduction

Drug discovery is a highly complex process that takes up to 15 years and costs billions of dollars [1], but its failure rate is extremely high due to a lack of clinical efficacy or safety [2]. Many complex diseases such as neurological and mental disorders are multi-genic, multi-factorial diseases. Thus, an effective therapy needs to modulate multiple genes [3, 4]. Polypharmacology that uses a single chemical to selectively bind to multiple drug targets has emerged as a new paradigm in drug discovery [5–10]. Furthermore, a drug molecule rarely interacts exclusively with its intended target. Off-target binding is common and can lead to undesirable side effects and drug-drug interactions [10]. Thus, elucidating selective ligand binding profiles on a genome-wide scale is essential for developing effective and safe polypharmacology therapeutics.

Screening a library of billions of compounds against multiple drug targets to identify hit compounds and subsequently optimizing their binding affinity selectivity profiles for drug leads are critical steps in polypharmacology [11]. A giga-scale compound screening can increase the likelihood of identifying more potent or selective ligands [12–15], streamline lead optimization, and expand the chemical diversity, chemical novelty, and patentability of drug leads [16]. Although DNA-encoded libraries are an effective approach to generate and screen billions of compounds for a single target [17], they have limited chemistries and a high rate of false positives. The computational approaches are expected to facilitate large-scale polypharmacology compound screening [11].

Protein-ligand docking is commonly used for compound screening [18]. However, protein-ligand docking suffers from a high rate of false positives due to poor modeling of protein dynamics, solvation effects, and other challenges [19]. The reliability of protein-ligand docking significantly deteriorates when using predicted structures [20, 21], and is not reliable for virtual screening [22]. Despite the success of AlphaFold2 [23], it can only reliably model half of understudied human proteins whose small-molecule ligands are unknown [24]. Thus, the application of protein-ligand docking in polypharmacology is hindered by the frequent lack of critical structural information, such as the experimentally determined holo structures and binding sites of many target proteins. Moreover, protein-ligand docking is computationally demanding. As a result,

it is impractical to use this approach to screen billions of compounds across multiple targets in the context of polypharmacology.

Recent advances in artificial intelligence encourage increasing interest in applying deep learning to drug discovery [25, 26]. Sequence-based deep learning offers significant advantages. Sequence-based drug-target interaction prediction, which only uses 3D structure information implicitly, enables fast drug-target interaction predictions when inputs consist solely of a molecular description of the chemical and the amino acid sequence of the target protein [27, 28]. By leveraging protein sequences instead of full 3D structures, we can reduce the computational burden and allow for the rapid evaluation of vast libraries of compounds, making it a more practical and scalable strategy for polypharmacology. However, the generalization power of deep learning methods for protein-ligand interaction predictions remains poor. Furthermore, the chemical space of small organic molecules is astronomically vast. Although the number of possible small organic molecules is approximate $10^{60}$ [29], only around $10^6$ compounds have annotated protein targets [30–32]. The limited coverage of chemical genomics space makes it challenging to train generalizable deep learning models for binding affinity predictions in an out-of-distribution (OOD) scenario [33, 34], in which unseen testing chemicals are significantly different from training data.

Since drug discovery is a high-stakes process, making decisions based on incorrect predictions can waste time and resources. Knowing the confidence level of an individual prediction will facilitate prioritizing lead compounds more effectively and efficiently. This requires an estimation of prediction errors on a compound-by-compound basis, not just a uniform confidence interval for all cases [35]. The uncertainty of prediction comes from either the lack of labeled data or data noisiness. Gaussian Process (GP) is one of the popular approaches for uncertainty quantification. Several works [36, 37] propose a combined GP and multi-layer perceptron (MLP) approach for various biological tasks. However, the proposed GP+MLP algorithm is computationally intensive and requires modification of the architecture of predictive models. Many studies have leveraged ensembles of neural networks for uncertainty quantification [38, 39]. Similar to the ensemble of neural networks, Bayesian-regularized neural networks generate a distribution of network weights and variations on predictions [40]. However, ensemble-based methods are time-consuming during both the training and inference stages, posing challenges for their application to large-scale polypharmacology. Along with these methods, Conformal Prediction aims to provide distribution-free uncertainty quantification under data exchangeability assumptions, a weaker version of independent and identically distribution (IID) [41, 42]. Many works have attempted to utilize conformal prediction for various aspects in drug discovery [43–45]. A vanilla conformal prediction offers uniform confidence intervals for all cases, but does not quantify the prediction error of a specific individual prediction. In practice, however, different cases warrant different levels of confidence: for example, OOD cases should naturally have larger intervals than in-distribution cases. Moreover, the exchangeability assumption underlying the conformer prediction does not hold in OOD cases.

To overcome the aforementioned challenges in large-scale polypharmacology compound screening, we propose a model-agnostic anomaly detection-based individual uncertainty quantification method, **e**mbedding **M**ahalanobis **O**utlier **S**coring and

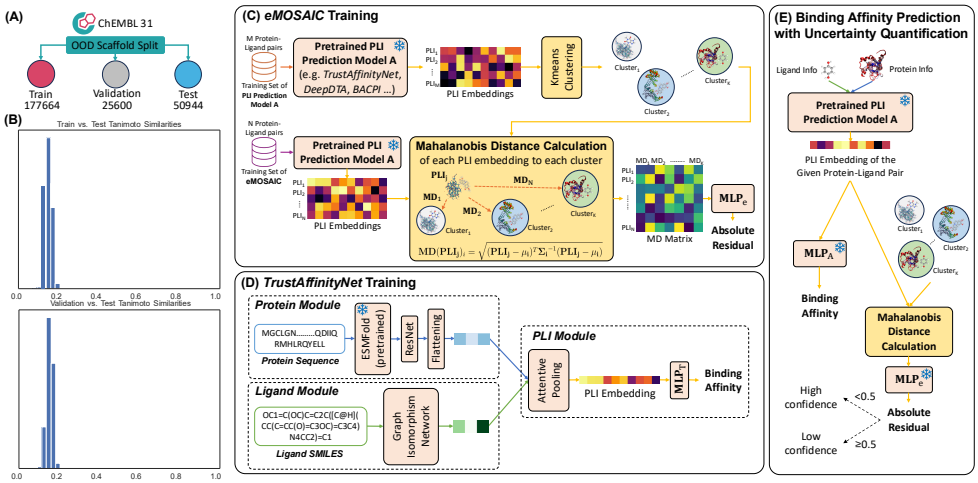


**Fig. 1** Dataset used for the model development and overall architecture of the pipeline. (A) ChEMBL31 dataset with OOD scaffold split to ensure no overlap of scaffolds among the training, validation, and test sets. (B) Tanimoto similarity distribution of the training vs. test and validation vs. test set, respectively. (C) Training process of the model-agnostic individual uncertainty quantification module, eMOSAIC. (D) Training process of the binding affinity prediction module, TrustAffinityNet. (E) Binding affinity prediction with uncertainty quantification for new protein-ligand pair. PLI: Protein-ligand interaction, which represents a pair of interacting protein and chemical.

**A**nomaly **I**dentification via **C**lustering (eMOSAIC). eMOSAIC is specifically designed to meet the following requirements: (i) accuracy for proteins lacking structural, holo-conformational, or binding site information, (ii) reliability when applied to compounds with novel scaffolds (OOD scenarios), and (iii) scalability to millions of compounds across thousands of proteins for large-scale screening. eMOSAIC explicitly models clusters of multiple local distributions within a multi-modal embedding space and utilizes the divergence between each cluster distribution and the embedding of an unseen test case as features for predicting residuals of test cases. The residual is the difference between a predicted value and the actual observed value. This approach differs from stochastic OOD detection methods [46], which rely on the global distribution of the embedding space to classify cases as OOD or non-OOD. In contrast, eMOSAIC not only detects OOD cases but also quantifies the prediction error of an OOD case. We apply eMOSAIC to various base models, including but not limited to TrustAffinityNet developed by us [47]. Under rigorous OOD benchmark studies, eMOSAIC significantly improves performance for binding affinity predictions when applied to multiple base models. The combination of eMOSAIC with TrustAffinityNet is superior to both state-of-the-art deep learning baselines and structure-based methods in terms of both accuracy and scalability for understudied proteins. Furthermore, eMOSAIC outperforms existing uncertainty quantification methods, demonstrating its potential for other machine learning applications. Thus, eMOSAIC represents a significant advance in deep learning applications to drug discovery.

# 2 Results

## 2.1 Overview of eMOSAIC

We developed eMOSAIC using ChEMBL31 database [30]. The distribution of training, validation, and test data is shown in Figure 1A. We evaluated the performance of eMOSAIC in a scaffold-based OOD setting, where chemicals in the testing set have different chemical scaffolds from those in the training/validation set. This setup mimics real-life scenarios, where we are likely to encounter new potential chemicals that have not been used in the training process of the deep learning model. In addition, the chemicals in the testing set are structurally dissimilar to those in the training/validation set, as shown in Figure 1B. Our OOD split is almost identical to UMAP-based split [48] in terms of chemical similarity distributions between testing and training data (Supplementary Figure 1). For the purpose of comparison, we also evaluate eMOSAIC in the in-distribution setting of random split.

Figure 1C-E provides an overview of eMOSAIC. eMOSAIC trains a secondary model for uncertainty estimation alongside the task-specific base models. However, unlike conventional methods, eMOSAIC generalizes well to OOD data. The training of eMOSAIC involves three main steps. First, the embeddings of training examples that are used to train a task-specific deep learning model are clustered. Second, given a case in the training set of eMOSAIC, the Mahalanobis distances between its embedding and each cluster of trained embeddings from the base model are calculated, and its absolute residual, i.e., the absolute value of the difference between predicted value and actual value, is obtained using the base model. Finally, eMOSAIC is trained using these Mahalanobis distances as features and the absolute residuals as labels. During the inference stage (Figure 1E), given a new case, in addition to the task-specific label $\hat{y}$ (e.g., binding affinity) predicted by the task-specific base model, eMOSAIC will predict the absolute residual $|\hat{y} - y|$, where $y$ is the actual binding affinity, using Mahalanobis distance features derived from the trained based model. The prediction is classified as high-confidence (low uncertainty) if the predicted absolute residual is less than 0.5, otherwise, low-confidence.

eMOSAIC is model-agnostic and can be applied to any trained task-specific deep learning models with embeddings. In this paper, we apply it to several protein-ligand binding affinity prediction models, including TrustAffinityNet developed by us [47] and other state-of-the-art models such as BIND [49], BACPI [50], DeepDTA [51], and DeepPurpose [52]. As shown in Figure 1D, TrustAffinityNet takes a protein sequence and a ligand SMILES as inputs. Leveraging a pre-trained protein language model, ESMFold [53], and the graph isomorphism network (GIN) [54], the embeddings of protein sequences and chemical structures are combined via attention pooling. The generated vectors from the attention pooling are used to predict binding affinities and train eMOSAIC for uncertainty quantification. For simplicity, whenever eMOSAIC$_T$ is mentioned, it specifically refers to the application of eMOSAIC to the TrustAffinityNet model.

We compare eMOSAIC with state-of-the-art methods in three tasks, (1) binding affinity prediction, (2) polypharmacology compound screening, and (3) uncertainty quantification. For binding affinity prediction and polypharmacology screening tasks,

we compare eMOSAIC$_T$ with sequence-based deep learning models, including BIND [49], BACPI [50], DeepDTA [51], and DeepPurpose [52], as well as a typical protein-ligand docking method, AutoDock Vina [55], a deep learning-based docking model, KarmaDock [56], and a structure-based deep learning model, EHIGN [57]. Along with this, we also compare the performance of eMOSAIC with other uncertainty quantification methods, including GP-based RIO framework [37], Monte Carlo Dropout [58], and Conformal Prediction [41].

We evaluate the performance of eMOSAIC and baseline models for various tasks and corresponding metrics, as detailed in Results and Methods. For a fair comparison, the same training, validation, and testing sets, training and evaluation procedures, and binding affinity thresholds are applied to all baseline models in the comparisons for binding affinity predictions and uncertainty quantification.

## 2.2 eMOSAIC improves OOD Binding Affinity Prediction

When we apply eMOSAIC to base models, including TrustAffinityNet, BIND, BACPI, DeepDTA, and DeepPurpose, in the OOD setting, as seen in Figure 2A, we observe significant performance improvements across *all* base models and *all* four regression metrics: RMSE, MAE, Pearson Correlation, and Spearman Correlation. These results demonstrate that eMOSAIC is model-agnostic and has the potential to enhance any deep learning models.

In the OOD setting, using structure-informed ESMFold model [53] for the protein sequence pre-training, TrustAffinityNet already outperforms state-of-the-art methods for the protein-ligand binding affinity predictions, as shown in Figure 2A. eMOSAIC further boosts the performance of TrustAffinityNet by quantifying the uncertainty of predictions from it. eMOSAIC$_T$ significantly outperforms all baseline models for the binding affinity predictions, as shown in Figure 2A and B. Although BIND, BACPI, DeepPurpose, and DeepDTA have acceptable performance in the random split setting (Supplementary Table 1), their performances significantly drop in the scaffold split setting. In contrast, the correlation between predicted binding affinities by eMOSAIC$_T$ and actual binding affinities remains high when testing chemicals have different scaffolds from those in the training set. Besides, the performance difference is not significant when using scaffold and random splitting for eMOSAIC$_T$ (Supplementary Figure 2). These findings clearly demonstrate the superior generalization power of eMOSAIC$_T$ when predicting the binding affinity in the OOD setting. The predictions by eMOSAIC$_T$ not only have higher correlation but also have significantly lower deviation as recorded by the RMSE on average 19.39%, 16.92%, 21.54%, and 18.97% lower, and MAE on average 20.03%, 16.98%, 22.55%, and 19.95% lower when compared to BACPI, BIND, DeepPurpose, and DeepDTA, respectively. As shown in Figure 2B, the prediction errors mainly come from the low- and high-affinity regions where there is little training data. Although eMOSAIC$_T$ can alleviate the problem compared to other methods, further improvement is needed.

When using a binding affinity threshold of pKi $> -2$, eMOSAIC$_T$ also significantly outperforms all baseline models in a classification task. As shown in Figure 2C and D, the ROC AUC and PR AUC of eMOSAIC$_T$ improve 4.6% and 4.7% over TrustAffinityNet, respectively. The improvements over other state-of-the-art models are more

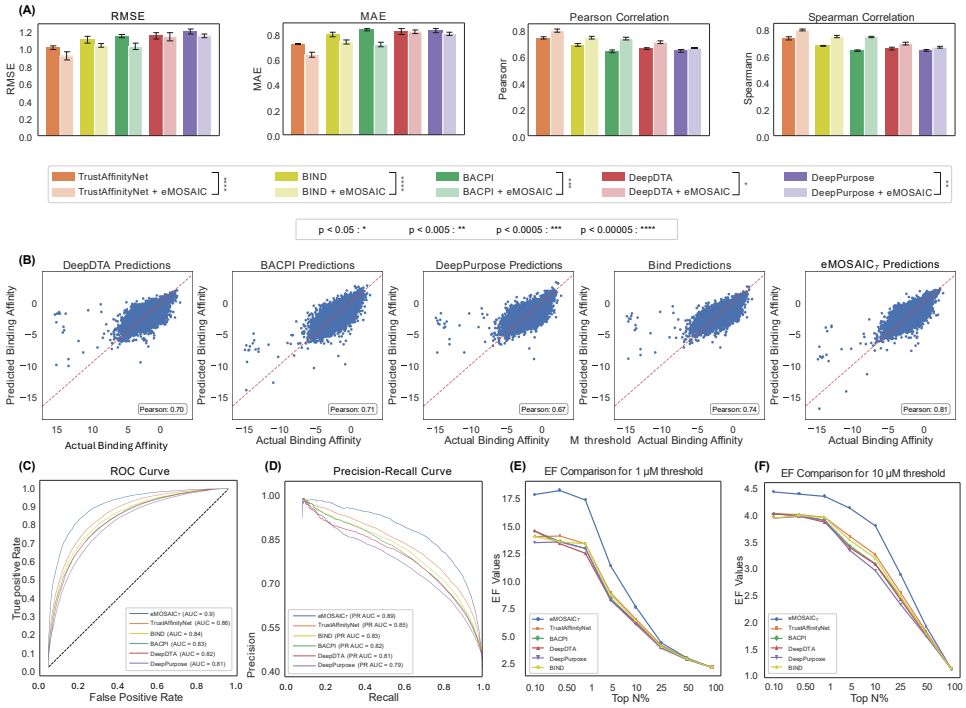


**Fig. 2** Performance comparison of the eMOSAIC-enhanced models (eMOSAIC+TrustAffinityNet (eMOSAIC$_T$), eMOSAIC+BIND, eMOSAIC+DeepDTA, eMOSAIC+DeepPurpose, and eMOSAIC+BACPI) against the base models (TrustAffinityNet, BIND, DeepDTA, DeepPurpose, and BACPI) in the OOD setting. (A) Barplots comparing RMSE, MAE, Pearson Correlation, and Spearman Correlation between actual and predicted binding affinities. Darker and lighter shade bars indicate the results obtained from base models and applying eMOSAIC to the base model. The bars represent the mean values across all test pairs (n=50,944 protein-ligand samples, unit of study = individual protein-ligand pairs). Error bars indicate the standard deviation (SD) across 3 independent splits (obtained via resampling replicates). Statistical significance of the differences between baseline and eMOSAIC-augmented versions was determined by p-values from the Mann-Whitney U-test is shown in the bottom of the barplot and the exact p-values are present in source data. (B) Scatterplot of actual binding affinities vs. predicted binding affinities. (C) Receiver operating characteristic (ROC) curves for binding classification. (D) Precision-recall (PR) curves for binding classification. (E) Enrichment factor (EF) for virtual screening based on the binding affinity threshold of $1\mu M$. (F) EF for virtual screening based on the binding affinity threshold of $10\mu M$

significant. They are at least 7.1% and 7.2%, respectively. The performance improvement in the low false positive rate is more obvious. When the false positive rate is 0.05, the positive rate of eMOSAIC$_T$ is over 50% higher than the existing methods, demonstrating the potential of eMOSAIC$_T$ for compound screening. As seen in Figure 2E and F, the enrichment factor (EF) of eMOSAIC$_T$ is indeed significantly higher than baseline models in the early ranking for virtual screening.

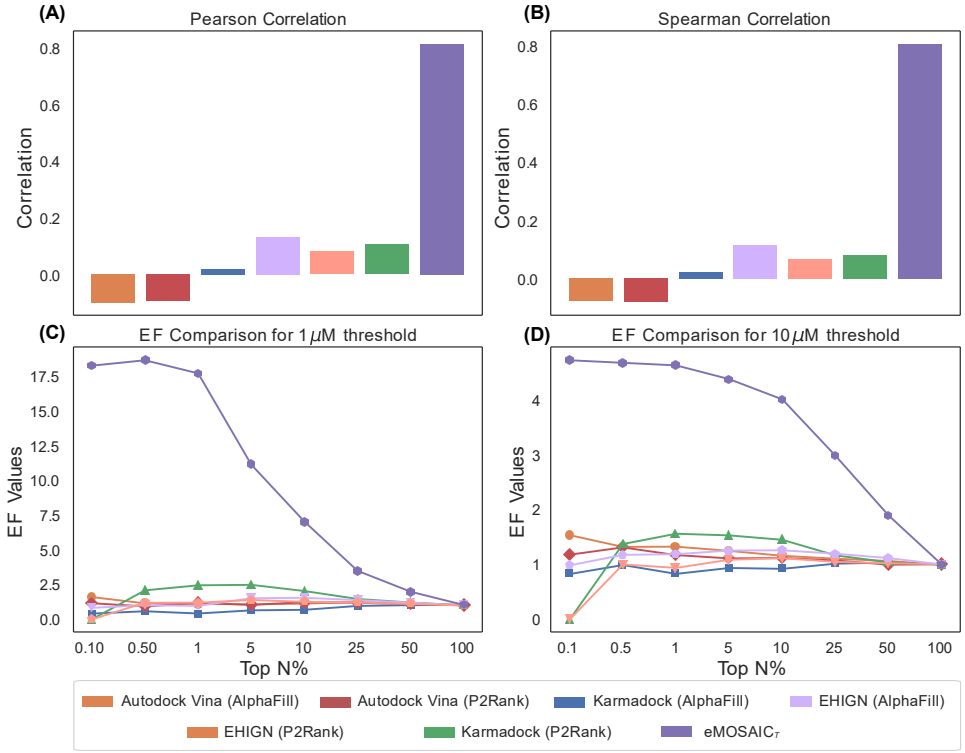


**Fig. 3** Performance comparison of eMOSAIC$_T$ with structure-based methods on the (A) Pearson Correlation, (B) Spearman Correlation, (C) Enrichment Factor with 1$\mu$M threshold, and (D) Enrichment Factor with 10$\mu$M threshold.

## 2.3 eMOSAIC significantly outperforms structure-based methods for binding affinity predictions and compound screening

We further compare the performance of eMOSAIC$_T$ with structure-based methods widely applied in compound virtual screening. Besides the commonly used baseline Autodock Vina, state-of-the-art method KarmaDock [56] was included because it has been reported to outperform widely used docking packages such as GLIDE [59] on standard docking and screening benchmarks, while also offering substantial speed advantages, thereby serving as a rigorous and practically relevant baseline. We also include EHIGN [57], a structure-based binding affinity prediction method that can be applied to virtual screening when binding conformations are computed via docking. When tested on the scaffold split set, structure-based methods show a notably poor correlation between predicted and actual binding affinities, as demonstrated in Figure 3A and B (also Supplementary Table 2). The structure-based deep learning models, EHIGN and KarmaDock, offer modest improvements over AutoDock Vina. In

contrast, eMOSAIC$_T$ significantly enhances the correlation of predicted binding affinities, showing an eight-fold improvement over KarmaDock, as shown in Figure 3. For the structure-based experiment, the definition of binding pockets is critical. However, in a real-world polypharmacology application, the binding pocket is often not clearly defined due to a lack of experimentally determined ligand-bound structures. Two methods are applied to determining binding pockets: the AlphaFill approach based on the alignment with co-crystallized complex structures [60] and predicted binding pockets from a machine learning method, P2Rank [61]. Although both AutoDock Vina and EHIGN perform better on these proteins with AlphaFill-predicted binding pockets, KarmaDock works better on the proteins with P2rank-predicted binding pockets.

For high-throughput screening applications, we compare the enrichment factor (EF) for thresholds of 1.0 $\mu$M and 10.0 $\mu$M. In this context, eMOSAIC$_T$ significantly outperforms all structure-based baselines, as shown in Figure 3C and D. The EF from eMOSAIC$_T$ is approximately eight times higher than that of structure-based methods for the threshold of 1.0 $\mu$M and three and a half times higher for the threshold of 10.0 $\mu$M for top 0.1% ranked compounds. Note that a transformation is applied to align the docking score distribution with the mean and standard deviation of the ground truth pKi before computing the performance metrics. Our findings suggest that for a large-scale polypharmacology screening where understudied proteins are often involved, eMOSAIC has a clear advantage over structure-based methods.

Furthermore, eMOSAIC$_T$ is several orders of magnitude faster than structure-based methods (Supplementary Figure 3). It takes less than 0.01 seconds for eMOSAIC$_T$, around 1 second for KarmaDock, and 30 seconds for Autodock Vina to predict the binding affinity of a protein-ligand pair. As a result, screening one million compounds against a target with eMOSAIC$_T$ can be completed in just a few days.

We also evaluated eMOSAIC$_T$ along with sequence-based and structure-based deep learning baselines on LIT-PCBA [62], a benchmark curated for compound virtual screening on a limited number of well-characterized ligand-bound conformations with well-defined binding pockets, an unrealistic scenario for polypharmacology. Since LIT-PCBA provides only separate ligand information and protein structures, an additional *time-consuming* docking step is required to generate protein-ligand binding complexes before structure-based scoring can be applied [57]. As shown in Supplementary Figure 4, and consistent with previous results, eMOSAIC$_T$ outperforms all sequence-based models and achieves performance comparable to the structure-based method EHIGN that is trained using holo structures.

## 2.4 eMOSAIC outperforms state-of-the-art uncertainty quantification methods

We compare eMOSAIC with three state-of-the-art methods for uncertainty quantification: RIO, which is based on Gaussian process, Monte Carlo dropout, and conformal prediction. Figure 4A presents sparsification curves, which assess how effectively uncertainty estimates identify unreliable predictions. Each curve is generated by progressively removing the most uncertain predictions and plotting the model's performance (RMSE in this case) on the remaining data as a function of the fraction removed. eMOSAIC achieves the lowest Area Under the Sparsification Error (AUSE),

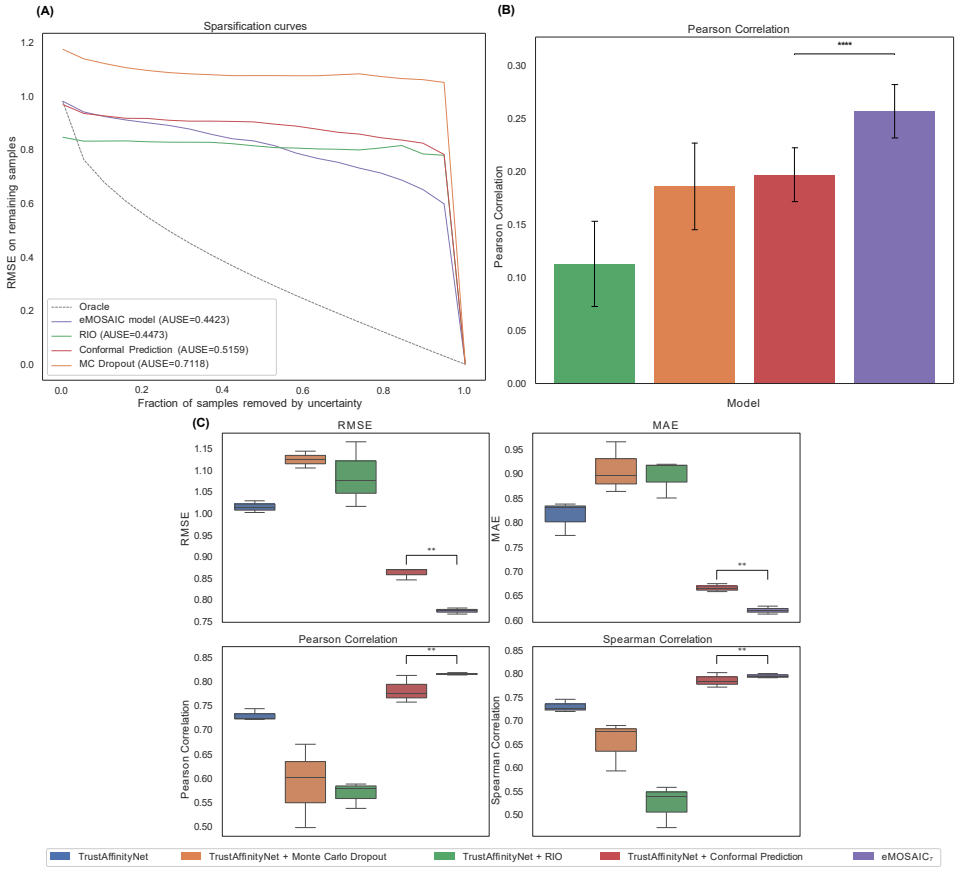


**Fig. 4** Performance comparison of eMOSAIC with state-of-the-art uncertainty quantification methods. (A) Sparsification curves: RMSE on remaining samples as an increasing fraction of highest-uncertainty cases is removed. Area under the sparsification error (AUSE) is reported in the legend (lower is better). (B) Barplot of Pearson correlation between predicted and true residuals across three independent random train/test splits of the scaffold-OOD dataset. Bars indicate mean values over independent protein-ligand pairs ($n$ = 50,944 per split; unit of study = independent protein-ligand pairs), with SD across the three splits shown as error bars. Statistical significance of correlation improvement was evaluated between eMOSAIC and CP using a two-sided Mann–Whitney U-test. (C) Errors and correlations of actual vs. predicted binding affinities for high-confidence predictions selected by each method. Statistical significance between eMOSAIC and CP was assessed using a two-sided Mann–Whitney U-test exact p-values for all comparisons are provided in the source data.

indicating the most effective uncertainty ranking. Its RMSE decreases sharply as uncertain samples are excluded, a desirable property for uncertainty quantification. In contrast, the RMSE of all baseline models shows little improvement. In addition, eMOSAIC has the highest Pearson Correlation between predicted residual and actual residual, as shown in Figure 4B.

When evaluating methods to enhance absolute binding affinity predictions, applying either eMOSAIC or conformal prediction to the base model TrustAffinityNet leads to a statistically significant performance improvement, as shown in Figure 4C. In contrast, RIO and Monte Carlo dropout do not provide noticeable gains. Between the two best methods, eMOSAIC outperforms conformal prediction across all evaluation metrics. A key distinction is that eMOSAIC offers instance-specific error estimation, while conformal prediction provides only global confidence intervals without per-sample error estimates, as demonstrated by examples in Supplemental Table 3. Instance-specific uncertainty quantification is particularly critical for high-stakes applications such as drug discovery, where decisions must be made on a compound-by-compound basis. Interestingly, the high-confidence predictions generated by eMOSAIC and those identified by conformal prediction are largely complementary, as illustrated in Supplementary Figure 5. The Pearson correlation between their predicted values is modest ($r = 0.38$), suggesting that the two methods capture different aspects of the predictive uncertainty. These observations point to the potential value of integrating eMOSAIC and conformal prediction to develop a more robust and comprehensive framework for individual uncertainty quantification.

The Monte Carlo dropout method is a crude approximation of uncertainty estimation. Its performance is worse than the base model without uncertainty quantification because the Bayesian posterior distribution estimated by dropout might be more complex [63]. Secondly, the Monte Carlo dropout explores limited configurations of potential weights, not all of them, resulting in incomplete uncertainty estimation [64]. The GP-based RIO framework performs even worse than the Monte Carlo dropout, due to the limited capability of the current kernels that fail to effectively model the dependency between the protein-ligand interaction embeddings and residuals. It results in less accurate residual correction and uncertainty estimation, and hence the poorest performance among all the methods.

The number of clusters in the embedding space may affect the performance of eMOSAIC. We evaluated the impact of this parameter (Supplementary Table 4). Our findings indicate that fine-grained clustering improves the capture of detailed information about the model's uncertainty. However, beyond 50 clusters, the performance plateaus, suggesting that further increasing the number of clusters does not yield significant improvements.

## 2.5 eMOSAIC enhances polypharmacology screening

GPCRs are compromised of targets for approximately 35% of all approved drugs, highlighting their therapeutic value. GPCRs' druggability and accessibility make them central targets for therapeutic interventions [65]. Along with these, protein kinase inhibitors have become a critical class of drugs, especially in oncology. Up to 33% of the drug development process targets these kinases [66]. Furthermore, many clinically successful therapeutics targeting GPCRs and kinases are proven to be polypharmacological drugs [10]. Therefore, we assess the models' performance in a polypharmacological context by evaluating the ability of ligands to interact with multiple GPCRs and kinases. Predicted pKi values for protein-ligand pairs are used to identify meaningful interactions. We apply a threshold of less than 100 $\mu$M of the

predicted pKi values because weak bindings may contribute to polypharmacology [10]. The evaluation is based on a multi-label classification to determine the potential of ligands to selectively act across different therapeutic targets. We evaluate the models using multi-label accuracy, micro-precision, micro-recall, micro-F1 score, and Hamming loss. For the docking scores, a transformation was applied to align their distribution with the mean and standard deviation of the ground truth pKi before computing the performance metrics.

As shown in Table 1, eMOSAIC$_T$ demonstrates significantly improved performance over other methods across all evaluation metrics for GPCR polypharmacology. Specifically, accuracy, precision, recall, and F1 score are improved by 18.4%, 15.5%, 16% and 15.6%, respectively, while the Hamming loss is reduced by 166.7% compared with the second-best method. For kinase polypharmacology, docking-based and structure-based methods including AutoDock, KarmaDock, and EHIGN, have the best sensitivity but poor specificity, aligning with the observations that protein-ligand docking has a high false positive rate in general. In contrast, eMOSAIC$_T$ has a balanced sensitivity and specificity, thus it has the best accuracy and F-1 score. Among all evaluation metrics, the Hamming loss is the best for evaluating multi-label classification performance. eMOSAIC$_T$ reduces the Hamming loss by 8.3% compared with the second-best method.

**Table 1** Performance comparison of polypharmacology compound screening on GPCR and Kinase proteins. ***Bold**: represents the best method, and Underlined: represents the second best model.

| Protein Type | Model | Accuracy ↑ | Precision ↑ | Recall ↑ | F-1 Score ↑ | Hamming Loss ↓ |
|---|---|---|---|---|---|---|
| GPCR | AutoDock | 0.50 | 0.45 | 0.62 | 0.47 | 0.50 |
| | KarmaDock | 0.52 | 0.52 | 0.40 | 0.35 | 0.48 |
| | EHIGN | 0.50 | 0.50 | 0.57 | 0.44 | 0.50 |
| | DeepPurpose | 0.72 | 0.70 | 0.68 | 0.67 | 0.28 |
| | DeepDTA | 0.74 | 0.72 | 0.70 | 0.69 | 0.26 |
| | BIND | 0.76 | 0.77 | 0.75 | 0.76 | 0.24 |
| | BACPI | 0.74 | 0.74 | 0.72 | 0.65 | 0.26 |
| | **eMOSAIC$_T$** | **0.90** | **0.89** | **0.87** | **0.87** | **0.09** |
| Kinase | AutoDock | 0.42 | 0.42 | **0.79** | 0.41 | 0.58 |
| | KarmaDock | 0.51 | 0.51 | 0.45 | 0.28 | 0.48 |
| | EHIGN | 0.46 | 0.46 | 0.67 | 0.38 | 0.54 |
| | DeepPurpose | 0.70 | 0.61 | 0.64 | 0.59 | 0.30 |
| | DeepDTA | 0.71 | 0.63 | 0.63 | 0.61 | 0.29 |
| | BIND | 0.72 | 0.63 | 0.63 | **0.63** | 0.26 |
| | BACPI | 0.68 | **0.68** | 0.41 | 0.31 | 0.32 |
| | **eMOSAIC$_T$** | **0.76** | 0.66 | 0.65 | 0.62 | **0.24** |

# 3 Discussion

In this work, we propose eMOSAIC$_T$, a framework for accurate, reliable, and scalable prediction of binding affinity along with an estimation of the associated uncertainty. We have successfully demonstrated the robust OOD generalization capabilities of

eMOSAIC$_T$, yielding reliable (high-confidence) binding affinity with high accuracy. Furthermore, we highlight the framework's notable advantage in terms of rapid inference speed, in contrast to protein-ligand docking, thereby rendering it well-suited for deployment in automated polypharmacology processes to leverage uncertainty-based methodologies.

Despite eMOSAIC$_T$'s superior performance, it has certain limitations that can be further addressed. Firstly, the partitioning of the embedding space affects eMOSAIC's performance. Using improved clustering methods, such as supervised contrastive learning, could yield better results. Secondly, in classification-based single-target compound screening, decoys are commonly used to augment the dataset with inactive compounds. However, in regression tasks aimed at predicting binding affinities, it is not straightforward to assign "pseudo" binding affinity values to decoys. Incorporating semi-supervised techniques [67] during training may facilitate data augmentation and boost its generalization to OOD data. Thirdly, the available bioactive chemicals used for training and testing are biased toward the existing target space and established medicinal chemistry practices. Further studies are needed to assess the generalization ability of machine learning methods in unexplored chemical spaces. Finally, the interpretability of trained machine learning models in OOD scenarios remains an open challenge. This issue warrants further investigation, such as by incorporating sequence-based ligand binding pose predictions [68] or exploring the combination of uncertainty quantification and interpretability within a transformer-based evidential learning framework [69].

# 4 Method

## 4.1 eMOSAIC for uncertainty quantification

### 4.1.1 eMOSAIC algorithm

Figure 1C-E provides an overview of eMOSAIC. We utilize embedding clustering and Mahalanobis distance to identify anomalies and quantify uncertainties. P.C. Mahalanobis introduced the metric known as Mahalanobis Distance (MD) in 1936 as a measure of anomaly and for detecting outliers [70]. For a given point in a distribution, the MD is defined as follows:

$$MD(x) = \sqrt{(x-\mu)\Sigma^{-1}(x-\mu)^T} \tag{1}$$

Here, $\mu$ and $\Sigma$ refer to the mean and variance of the distribution. It considers the variance between the various variables in a multivariate distribution since real-life data often contains many correlated variables. Mahalanobis distance has normalization through division by the covariance matrix, ensuring that variables with different scales are suitably handled. Because of these properties, it can effectively identify outliers or anomalies from the main distribution. Mahalanobis distance has previously been used in deep learning frameworks for anomaly, OOD, and adversarial detection in computer vision, time series, and natural language processing tasks [71–74]. These instances have

clearly motivated the usage of Mahalanobis distance for anomaly detection tasks to improve the reliability of predictions by deep learning frameworks.

After training a deep learning model, which serves as the protein-ligand binding affinity prediction module in this paper, we extract the protein-ligand interaction (PLI) embeddings that represent the protein-ligand pairs. We then perform k-means clustering on the PLI embeddings from the deep learning model's training set and obtain the mean ($\mu$) and variance matrix ($\mathbf{\Sigma}$) for each cluster to facilitate Mahalanobis distance calculation. Given the embedding of an unseen PLI, we calculate the Mahalanobis distance to each cluster and use these distances as features to train a simple multi-layer perceptron (MLP) to predict the absolute residual, defined as the absolute difference between the predicted and true binding affinity. This approach enables the MLP to identify the protein-ligand interaction anomalies predicted by the model. For outlier detection, we select protein-ligand pairs with predicted residuals below 0.5, ensuring that only high-confidence predictions are retained.

### 4.1.2 Training of eMOSAIC$_T$

As shown in Figure 1C, we first train the protein-ligand binding affinity prediction module. After training, we use the best model, selected based on validation set performance, to extract PLI embeddings and generate clusters. Finally, we train eMOSAIC, the OOD uncertainty quantification module, to identify outliers. More details about the hyperparameters and configuration of the eMOSAIC$_T$ and TrustAffinityNet model are present in Supplementary Table 5 and 6.

## 4.2 Binding Affinity Module TrustAffinityNet

The binding affinity module, TrustAffinityNet, consists of three sub-modules: the protein sequence module, the ligand processing module, and the protein-ligand interaction module. All of these modules collaboratively predict the binding affinity associated with the PLI.

### 4.2.1 Protein sequence module

Protein sequence representation is one of the most vital components in the machine learning frameworks for predicting not only PLIs [34, 50, 51, 75], but also their 3D structure [23, 53]. Protein sequences contain information that can be used to infer protein structure, function, and family [53], making them a rich source of data for machine learning models [23, 53]. Large datasets of protein sequences are available [23, 76, 77], enabling machine learning frameworks to learn high-level, general representations of proteins. We utilize ESMFold [53] to obtain the protein sequence embeddings, which deploys a large language model (LLM), ESM-2, alongside a folding module and a structure module for modeling the protein structure. The ESM-2 protein language model, which is able to capture the protein structures at the fine resolution of the atomic level, consists of variable parameters ranging from 8M to 15B. We use the 650M parameter model to obtain the refined protein sequence representation. We observed that the sequence representations obtained from the structure module of the ESMFold model performed better than the protein embeddings obtained from the ESM-2 model

directly as well as the sequence embeddings obtained from the folding block, possibly because the structure block refines the protein sequence obtained from the ESM-2 model. We remove protein sequences greater than 700 in length as they are very low in number, and due to constraints on time and memory. Since the embeddings obtained are variable in size corresponding to the protein sequence length, to make them consistent for the next steps, we perform padding to pad the sequences with lengths less than 700, and define masks associated with the sequences that track the padding. Since CNNs are known to work well with processing sequence representations, we use the ResNet model [78] with 5 layers, and each layer has 4 convolutional layers to obtain a refined protein sequence embedding. Finally, adaptive masking (using interpolation) is used based on the changes to the embeddings to avoid the loss of information.

#### 4.2.2 Ligand module

We represent each ligand as a 2D graph, where the nodes symbolize atoms and the edges are bonds. Embeddings for both node and edge are learned using the graph isomorphism network (GIN) [79]. For atom or node attributes, we used atom types, hybridization types, atom degrees, atom chirality, atom formal charges, and atom aromatic, all converted to one-hot encoding before being utilized by GIN. We use a 5-layer GIN architecture, which aggregates and updates node embedding for each atom/node. To obtain a graph-level or a ligand-level embedding that remains permutation invariant, a final sum pooling operation is used.

#### 4.2.3 Protein-ligand interaction module

After obtaining both the protein and ligand embeddings, we use the attentive pooling network [80] such that the model is aware of both protein and ligand and that the interaction isn't solely dependent on either protein or ligand. This network gives us the attention-weighted embeddings for both which are then concatenated and fed to an MLP which predicts the final binding affinity.

### 4.3 Experiments

#### 4.3.1 Dataset

We train TrustAffinityNet and eMOSAIC on the ChEMBL31 database [30], which consists of 350400 protein-ligand interaction pairs, along with their binding affinity values in nanomolar ($nM$), denoted as Ki. ChEMBL has been extensively used as a benchmark to develop and evaluate molecular property prediction models [81, 82]. In the experiments, we split the dataset into training, testing, and validation sets by 7:2:1. Negative log (base 10) transformation was performed on Ki (binding affinity) to obtain pKi values. The data was split using the following methods: (1) Random Split: random selection of protein-ligand pairs, (2) Random Scaffold Split: random selection of scaffolds of chemical structures [83] such that the chemicals in the testing set have different scaffolds from those in the training/validation set. Scaffold split ensures that there is no overlap of the scaffold in the training, validation, and testing sets. This was done to validate the model's generalization power in a real-world OOD setting.

Moreover, considering the vast chemical space for drug discovery, it is very likely that the model will encounter unknown and new scaffolds. In addition, no chemicals in the testing set are similar to those in the training/validation set with a Tanimoto coefficient larger than 0.4. (3) UMAP-based clustering split: Introduced by Guo et al. [48], they perform dimensionality reduction of Morgan fingerprints using UMAP followed by clustering. In addition to ChEMBL31, we utilize LIT-PCBA [62], which is a benchmark curated for unbiased compound virtual screening. Specifically, we include data for seven targets: ALDH1 (PDB: 4x4l), FEN1 (PDB: 5fv7), GBA (PDB: 2v3e), KAT2A (PDB: 5h84), MAPK1 (PDB: 2ojg), PKM2 (PDB: 3gr4), and VDR (PDB: 3a2j). The protein-ligand complex structures of these seven targets were obtained from Chao et. al [84].

### 4.3.2 Baseline models

We test eMOSAIC against baselines in two different objectives: 1) Binding Affinity Prediction and 2) Uncertainty Quantification.

For binding affinity prediction task we compare eMOSAIC$_T$ with sequence-based deep learning models, including BIND [49], BACPI [50], DeepDTA [51], and DeepPurpose [52], as well as a typical protein-ligand docking method, AutoDock Vina [55], a deep learning-based docking model, KarmaDock [56], and a structure-based deep learning model, EHIGN [57] on the OOD test set. Specifically, when applying eMOSAIC to sequence-based deep learning models, for each case, we attempt to select the most meaningful embeddings. For BACPI, we select the embeddings after the bi-directional interaction of both the ligand and protein representations. For BIND, we apply eMOSAIC to Learnable Commutative Monoid (LCM) aggregated embeddings, which already have ligand-influenced protein representation. For DeepPurpose and DeepDTA, we extract the protein and ligand representation, once they have been encoded by their respective architectures, and then use that as input to eMOSAIC. Lastly, for TrustAffinityNet, we extract the attentive pooling-based representation. In order to compare with docking programs, AutoDock and KarmaDock were used to predict the docking scores for ligand-protein pairs for the scaffold splitting data set. Alphafill [60] annotated binding pockets were used to define the searching space for AutoDock and KarmaDock. After removing the binding pockets for 38 different irons, such as FE, NA, 1246 binding pockets on 429 proteins were used as the pre-defined binding pockets to set up docking. If there are multiple pockets on one protein, the ligand will be docked into all pockets, and the one with the best docking score will be selected for this ligand-protein pair. For comparison with EHIGN on polypharmacology screenings, we obtained protein–ligand complex structures for the pairs in the OOD test set using AutoDock Vina. Binding affinity predictions were then obtained using the pretrained EHIGN model provided by the original authors, which was trained on the PDBBind dataset based on experimentally determined protein-ligand complex. The same pretrained model was also used to predict binding affinities for the seven targets from the LIT-PCBA dataset with the corresponding protein-ligand complex structures obtained from additional docking [84]. For other proteins not in Alphafill dataset, P2Rank [61] was used to predict the binding pockets on their Alphafold predicted model structures. The ligands were then docked into these binding pockets by

AutoDock and KarmaDock, and the best docking scores were selected for these ligand-protein pairs. In order to evaluate how many of the top-ranked pairs have high binding affinities, top 1000 hit rates were calculated, which measure how many of the top 1000 ranked pairs have pKi > -log(1) or -log(10). A higher $top_{1000}$ hit rate correlates with better docking rank performances.

Along with this, we also compare the performance of our uncertainty quantification module with other uncertainty quantification methods, including GP-based RIO framework [37], Monte Carlo Dropout [58], and Conformal Prediction [41]. Conformal prediction relies on obtaining a nonconformity score to measure the confidence interval for the predictions. This nonconformity score is crucial, and several different methods exist to compute it [42], and several methods are available that describe various variants of Conformal Prediction [85]. The most common nonconformity measures for regression-based models are based on absolute error, including using the calibration set's absolute errors to provide intervals for the new predictions and using predicted absolute errors [42]. In our case, we use the attentive pooling embeddings to train an MLP for predicting the absolute error in the case of CP baseline. We then select the points that have a confidence interval length lower than the average on the test set. We do the same for obtaining the high-confidence points for the Monte Carlo Dropout method.

### 4.3.3 Evaluation

For binding affinity prediction, we evaluate model performance using both regression-based metrics, including root mean squared error (RMSE), mean absolute error (MAE), Pearson's correlation coefficient, and Spearman's correlation coefficient, and screening-based metrics, including enrichment factors (EF) at different top-ranked percentages. Compounds are labeled as active or inactive based on binding affinity thresholds of 1.0 $\mu$M and 10.0 $\mu$M.

For polypharmacology compound screening, formulated as a multi-label classification task, we report accuracy, micro-precision, micro-recall, micro-F1 score, and Hamming loss. A binding affinity threshold of 100 $\mu$M is used to determine active targets, as weak binding interactions can still contribute to polypharmacological effects [6, 10].

For uncertainty quantification, we evaluate the methods based on their impact on binding affinity prediction performance, the correlation between predicted and experimentally determined binding residues, and the distribution of predicted binding residues.

### 4.3.4 Hardware and Software

Models were implemented in Python 3.10.9 using PyTorch (version 1.12.1, CUDA 11.3.1 and cuDNN 8) and PyTorch Geometric 2.2.0, on a NVIDIA Tesla V100 32GB GPU. Additionally, analysis was performed using the following libraries: 1) NumPy (version 1.24.2), 2) Pandas (version 1.5.3), 3) SciPy (version 1.10.0) and 4) scikit-learn (version 2.1.1).

## Data Availability

The original binding affinity data were obtained from the publicly available ChEMBL database (https://www.ebi.ac.uk/chembl/). The LIT-PCBA dataset was downloaded from the official LIT-PCBA project website (https://drugdesign.unistra.fr/LIT-PCBA/), and the associated protein-complexes were obtained at https://zenodo.org/records/4291725 [86]. The scaffold-based split data used in our experiments, along with the UMAP-based split referenced in the supplementary material, is available on Zenodo (https://doi.org/10.5281/zenodo.17409085) [87].

## Code Availability

The code for this work is available on GitHub (https://github.com/XieResearchGroup/eMOSAIC) and Zenodo (https://doi.org/10.5281/zenodo.15313879) [88].

## Acknowledgement

This project has been funded with federal funds from the National Institute of General Medical Sciences of the National Institute of Health (R01GM122845, LX), the National Institute on Aging of the National Institute of Health (R01AG057555, LX; R21AG083302, LX), and the National Science Foundation (2226183, LX).

## Author Contributions

AB prepared data, implemented the algorithms, performed the experiments, analyzed data, and wrote the manuscript; Li X prepared data, performed the experiments, analyzed data, and wrote the manuscript; SZ implemented the algorithms and wrote the manuscript; Lei X conceived methods, planned the experiments, and wrote the manuscript.

## Competing Interests Statement

The authors declare no competing interests.

# References


[1] Hughes, J.P., Rees, S., Kalindjian, S.B., Philpott, K.L.: Principles of early drug discovery. British journal of pharmacology **162**(6), 1239–1249 (2011)

[2] Sun, D., Gao, W., Hu, H., Zhou, S.: Why 90% of clinical drug development fails and how to improve it? Acta Pharmaceutica Sinica B **12**(7), 3049–3062 (2022)

[3] Badano, J.L., Katsanis, N.: Beyond mendel: an evolving view of human genetic disease transmission. Nature Reviews Genetics **3**(10), 779–789 (2002)

[4] Hunter, D.J.: Gene–environment interactions in human diseases. Nature reviews genetics **6**(4), 287–298 (2005)

[5] Albertini, C., Salerno, A., Sena Murteira Pinheiro, P., Bolognesi, M.L.: From combinations to multitarget-directed ligands: A continuum in alzheimer's disease polypharmacology. Medicinal Research Reviews **41**(5), 2606–2633 (2021)

[6] Proschak, E., Stark, H., Merk, D.: Polypharmacology by design: a medicinal chemist's perspective on multitargeting compounds. Journal of medicinal chemistry **62**(2), 420–444 (2018)

[7] Ravikumar, B., Aittokallio, T.: Improving the efficacy-safety balance of polypharmacology in multi-target drug discovery. Expert opinion on drug discovery **13**(2), 179–192 (2018)

[8] Chaudhari, R., Tan, Z., Huang, B., Zhang, S.: Computational polypharmacology: a new paradigm for drug discovery. Expert opinion on drug discovery **12**(3), 279–291 (2017)

[9] Anighoro, A., Bajorath, J., Rastelli, G.: Polypharmacology: challenges and opportunities in drug discovery: miniperspective. Journal of medicinal chemistry **57**(19), 7874–7887 (2014)

[10] Xie, L., Xie, L., Kinnings, S.L., Bourne, P.E.: Novel computational approaches to polypharmacology as a means to define responses to individual drugs. Annual review of pharmacology and toxicology **52**, 361–379 (2012)

[11] Sadybekov, A.V., Katritch, V.: Computational approaches streamlining drug discovery. Nature **616**(7958), 673–685 (2023)

[12] Lyu, J., Wang, S., Balius, T.E., Singh, I., Levit, A., Moroz, Y.S., O'Meara, M.J., Che, T., Algaa, E., Tolmachova, K., *et al.*: Ultra-large library docking for discovering new chemotypes. Nature **566**(7743), 224–229 (2019)

[13] Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P.D., Coote, P.W., Padmanabha Das, K.M., Malets, Y.S., Radchenko, D.S., Moroz, Y.S., Scott, D.A., *et al.*: An open-source drug discovery platform enables ultra-large virtual screens.

Nature **580**(7805), 663–668 (2020)

[14] Sadybekov, A.A., Sadybekov, A.V., Liu, Y., Iliopoulos-Tsoutsouvas, C., Huang, X.-P., Pickett, J., Houser, B., Patel, N., Tran, N.K., Tong, F., *et al.*: Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. Nature **601**(7893), 452–459 (2022)

[15] Lyu, J., Irwin, J.J., Shoichet, B.K.: Modeling the expansion of virtual screening libraries. Nature Chemical Biology **19**(6), 712–718 (2023)

[16] Tomberg, A., Boström, J.: Can easy chemistry produce complex, diverse, and novel molecules? Drug Discovery Today **25**(12), 2174–2181 (2020)

[17] Neri, D., Lerner, R.A.: Dna-encoded chemical libraries: a selection system based on endowing organic compounds with amplifiable information. Annual review of biochemistry **87**, 479–502 (2018)

[18] Xie, L., Ge, X., Tan, H., Xie, L., Zhang, Y., Hart, T., Yang, X., Bourne, P.E.: Towards structural systems pharmacology to study complex diseases and personalized medicine. PLoS computational biology **10**(5), 1003554 (2014)

[19] Grinter, S.Z., Zou, X.: Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. Molecules **19**(7), 10150–10176 (2014)

[20] Jaiteh, M., Rodríguez-Espigares, I., Selent, J., Carlsson, J.: Performance of virtual screening against gpcr homology models: Impact of template selection and treatment of binding site plasticity. PLoS computational biology **16**(3), 1007680 (2020)

[21] Cai, T., Xie, L., Zhang, S., Chen, M., He, D., Badkul, A., Liu, Y., Namballa, H.K., Dorogan, M., Harding, W.W., *et al.*: End-to-end sequence-structure-function meta-learning predicts genome-wide chemical-protein interactions for dark proteins. PLOS Computational Biology **19**(1), 1010851 (2023)

[22] Scardino, V., Di Filippo, J.I., Cavasotto, C.N.: How good are alphafold models for docking-based virtual screening? Iscience **26**(1) (2023)

[23] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.*: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021)

[24] Binder, J.L., Berendzen, J., Stevens, A.O., He, Y., Wang, J., Dokholyan, N.V., Oprea, T.I.: Alphafold illuminates half of the dark human proteins. Current opinion in structural biology **74**, 102372 (2022)

[25] Jayatunga, M.K., Xie, W., Ruder, L., Schulze, U., Meier, C.: Ai in small-molecule

drug discovery: A coming wave. Nat. Rev. Drug Discov **21**, 175–176 (2022)

[26] Schneider, P., Walters, W.P., Plowright, A.T., Sieroka, N., Listgarten, J., Goodnow Jr, R.A., Fisher, J., Jansen, J.M., Duca, J.S., Rush, T.S., *et al.*: Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery **19**(5), 353–364 (2020)

[27] Xu, L., Ru, X., Song, R.: Application of machine learning for drug–target interaction prediction. Frontiers in Genetics **12**, 680117 (2021)

[28] Bagherian, M., Sabeti, E., Wang, K., Sartor, M.A., Nikolovska-Coleska, Z., Najarian, K.: Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Briefings in bioinformatics **22**(1), 247–269 (2021)

[29] Lemonick, S.: Exploring chemical space: can ai take us where no human has gone before? Chemical & Engineering News **98**(13), 30–35 (2020)

[30] Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., *et al.*: Chembl: towards direct deposition of bioassay data. Nucleic acids research **47**(D1), 930–940 (2019)

[31] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., *et al.*: Pubchem 2023 update. Nucleic acids research **51**(D1), 1373–1380 (2023)

[32] Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J.: Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research **44**(D1), 1045–1053 (2016)

[33] Nguyen, T., Le, H., Le, T., Venkatesh, S.: Prediction of drug–target binding affinity using graph neural networks. BioRxiv, 684662 (2019)

[34] Karimi, M., Wu, D., Wang, Z., Shen, Y.: Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics **35**(18), 3329–3338 (2019)

[35] Chakraborti, T., Banerji, C.R., Marandon, A., Hellon, V., Mitra, R., Lehmann, B., Bräuninger, L., McGough, S., Turkay, C., Frangi, A.F., *et al.*: Personalized uncertainty quantification in artificial intelligence. Nature Machine Intelligence **7**(4), 522–530 (2025)

[36] Hie, B., Bryson, B.D., Berger, B.: Leveraging uncertainty in machine learning accelerates biological discovery and design. Cell systems **11**(5), 461–477 (2020)

[37] Qiu, X., Meyerson, E., Miikkulainen, R.: Quantifying point-prediction uncertainty in neural networks via residual estimation with an i/o kernel. In: International Conference on Learning Representations (2020)

[38] Zeng, H., Gifford, D.K.: Quantification of uncertainty in peptide-mhc binding prediction improves high-affinity peptide selection for therapeutic design. Cell systems **9**(2), 159–166 (2019)

[39] Mathelin, A., Deheeger, F., Mougeot, M., Vayatis, N.: Deep anti-regularized ensembles provide reliable out-of-distribution uncertainty quantification. arXiv preprint arXiv:2304.04042 (2023)

[40] Pearce, T., Leibfried, F., Brintrup, A.: Uncertainty in neural networks: Approximately bayesian ensembling. In: International Conference on Artificial Intelligence and Statistics, pp. 234–244 (2020). PMLR

[41] Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. UAI'98, pp. 148–155. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998)

[42] Kato, Y., Tax, D.M., Loog, M.: A review of nonconformity measures for conformal prediction in regression. Conformal and Probabilistic Prediction with Applications, 369–383 (2023)

[43] Norinder, U., Carlsson, L., Boyer, S., Eklund, M.: Introducing conformal prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. Journal of chemical information and modeling **54**(6), 1596–1603 (2014)

[44] Svensson, F., Aniceto, N., Norinder, U., Cortes-Ciriano, I., Spjuth, O., Carlsson, L., Bender, A.: Conformal regression for quantitative structure–activity relationship modeling—quantifying prediction uncertainty. Journal of Chemical Information and Modeling **58**(5), 1132–1140 (2018)

[45] Alvarsson, J., McShane, S.A., Norinder, U., Spjuth, O.: Predicting with confidence: using conformal prediction in drug discovery. Journal of Pharmaceutical Sciences **110**(1), 42–49 (2021)

[46] Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024)

[47] Badkul, A., Xie, L., Zhang, S., Xie, L.: Trustaffinity: accurate, reliable and scalable out-of-distribution protein-ligand binding affinity prediction using trustworthy deep learning. bioRxiv, 2024–01 (2024)

[48] Guo, Q., Hernandez-Hernandez, S., Ballester, P.J.: Scaffold splits overestimate virtual screening performance. In: International Conference on Artificial Neural Networks, pp. 58–72 (2024). Springer

[49] Lam, H.Y.I., Guan, J.S., Ong, X.E., Pincket, R., Mu, Y.: Protein language models are performant in structure-free virtual screening. Briefings in Bioinformatics

**25**(6), 480 (2024)

[50] Li, M., Lu, Z., Wu, Y., Li, Y.: Bacpi: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. Bioinformatics **38**(7), 1995–2002 (2022)

[51] Öztürk, H., Özgür, A., Ozkirimli, E.: Deepdta: deep drug–target binding affinity prediction. Bioinformatics **34**(17), 821–829 (2018)

[52] Huang, K., Fu, T., Glass, L.M., Zitnik, M., Xiao, C., Sun, J.: Deeppurpose: a deep learning library for drug–target interaction prediction. Bioinformatics **36**(22-23), 5545–5547 (2020)

[53] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.*: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[54] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., Leskovec, J.: Strategies for pre-training graph neural networks. arXiv preprint arXiv:1905.12265 (2019)

[55] Trott, O., Olson, A.J.: Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry **31**(2), 455–461 (2010)

[56] Zhang, X., Zhang, O., Shen, C., Qu, W., Chen, S., Cao, H., Kang, Y., Wang, Z., Wang, E., Zhang, J., *et al.*: Efficient and accurate large library ligand docking with karmadock. Nature Computational Science **3**(9), 789–804 (2023)

[57] Yang, Z., Zhong, W., Lv, Q., Dong, T., Chen, G., Chen, C.Y.-C.: Interaction-based inductive bias in graph neural networks: enhancing protein-ligand binding affinity predictions from 3d structures. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

[58] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016). PMLR

[59] Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., *et al.*: Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. Journal of medicinal chemistry **47**(7), 1739–1749 (2004)

[60] Hekkelman, M.L., Vries, I., Joosten, R.P., Perrakis, A.: Alphafill: enriching alphafold models with ligands and cofactors. Nature Methods **20**(2), 205–213 (2023)

[61] Krivák, R., Hoksza, D.: P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of cheminformatics **10**, 1–12 (2018)

[62] Tran-Nguyen, V.-K., Jacquemard, C., Rognan, D.: Lit-pcba: an unbiased data set for machine learning and virtual screening. Journal of chemical information and modeling **60**(9), 4263–4273 (2020)

[63] Serpell, C., Araya, I., Valle, C., Allende, H.: Probabilistic forecasting using monte carlo dropout neural networks. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings 24, pp. 387–397 (2019). Springer

[64] Sicking, J., Akila, M., Wirtz, T., Houben, S., Fischer, A.: Characteristics of monte carlo dropout in wide neural networks. arXiv preprint arXiv:2007.05434 (2020)

[65] Sriram, K., Insel, P.A.: G protein-coupled receptors as targets for approved drugs: how many targets and how many drugs? Molecular pharmacology **93**(4), 251–258 (2018)

[66] Roskoski Jr, R.: Properties of fda-approved small molecule protein kinase inhibitors: A 2023 update. Pharmacological research **187**, 106552 (2023)

[67] Wu, Y., Xie, L., Liu, Y., Xie, L.: Semi-supervised meta-learning elucidates understudied molecular interactions. Communications Biology **7**(1), 1104 (2024)

[68] Zhang, S., Xie, L., Tiourine, D., Xie, L.: Sequence-based drug-target complex pretraining enhances protein-ligand binding process predictions tackling crypticity. bioRxiv, 2025–01 (2025)

[69] Wu, C., Zuo, M., Xie, L.: edoc: Explainable decoding out-of-domain cell types with evidential learning. arXiv preprint arXiv:2411.00054 (2024)

[70] Mahalanobis, P.C.: On the generalized distance in statistics. Sankhyā: The Indian Journal of Statistics, Series A (2008-) **80**, 1–7 (2018)

[71] Rippel, O., Mertens, P., Merhof, D.: Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 6726–6733 (2021). IEEE

[72] Gjorgiev, L., Gievska, S.: Time series anomaly detection with variational autoencoder using mahalanobis distance. In: ICT Innovations 2020. Machine Learning and Applications: 12th International Conference, ICT Innovations 2020, Skopje, North Macedonia, September 24–26, 2020, Proceedings 12, pp. 42–55 (2020). Springer

[73] Anthony, H., Kamnitsas, K.: On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, pp. 136–146 (2023). Springer

[74] Prekopcsák, Z., Lemire, D.: Time series classification by class-specific mahalanobis distance measures. Advances in Data Analysis and Classification **6**, 185–200 (2012)

[75] Wang, K., Zhou, R., Li, Y., Li, M.: Deepdtaf: a deep learning method to predict protein–ligand binding affinity. Briefings in Bioinformatics **22**(5), 072 (2021)

[76] Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., *et al.*: Mgnify: the microbiome analysis resource in 2020. Nucleic acids research **48**(D1), 570–578 (2020)

[77] Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., Consortium, U.: Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics **31**(6), 926–932 (2015)

[78] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[79] Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)

[80] Santos, C.d., Tan, M., Xiang, B., Zhou, B.: Attentive pooling networks. arXiv preprint arXiv:1602.03609 (2016)

[81] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., Leskovec, J.: Strategies for pre-training graph neural networks. In: International Conference on Learning Representations (2020)

[82] Ji, Y., Zhang, L., Wu, J., Wu, B., Li, L., Huang, L.-K., Xu, T., Rong, Y., Ren, J., Xue, D., *et al.*: Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 8023–8031 (2023)

[83] Ramsundar, B., Eastman, P., Walters, P., Pande, V.: Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More. O'Reilly Media, Sebastopol, Calif. (2019)

[84] Shen, C., Weng, G., Zhang, X., Leung, E.L.-H., Yao, X., Pang, J., Chai, X., Li,

D., Wang, E., Cao, D., *et al.*: Accuracy or novelty: what can we gain from target-specific machine-learning-based scoring functions in virtual screening? Briefings in Bioinformatics **22**(5), 410 (2021)

[85] Papadopoulos, H., Vovk, V., Gammerman, A.: Regression conformal prediction with nearest neighbours. Journal of Artificial Intelligence Research **40**, 815–840 (2011)

[86] Shen, C.: Accuracy or novelty: what can we gain from target-specific machine learning-based scoring functions in virtual screening? (2020) https://doi.org/10.5281/zenodo.4291725

[87] Badkul, A., Xie, L., Zhang, S., Xie, L.: Data splits for "multimodal out-of-distribution individual uncertainty quantification enhances binding affinity prediction for polypharmacology" (2025) https://doi.org/10.5281/zenodo.17409086

[88] Badkul, A., Xie, L., Zhang, S., Xie, L.: Code for "multimodal out-of-distribution individual uncertainty quantification enhances binding affinity prediction for polypharmacology" (2025) https://doi.org/10.5281/zenodo.15313879