eMOSAIC: Multi-modal Out-of-distribution Uncertainty Quantification Streamlines Large-scale Polypharmacology

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

[1]Ph.D. Programs in Computer Science, Biology Biochemistry, 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, 10065, U.S.A.

*Corresponding author(s). E-mail(s): lei.xie@hunter.cuny.edu;

Abstract

Polypharmacology has emerged as a new paradigm to discover novel therapeutics for unmet medical needs. Accurate, reliable and scalable predictions of protein-ligand binding affinity across multiple proteins are essential for polypharmacology. Machine learning is a promising tool for multi-target binding affinity predictions, often formulated as a multi-modal regression problem. Despite considerable efforts, three challenges remain: out-of-distribution (OOD) generalizations for compounds with new chemical scaffolds, uncertainty quantification of OOD predictions, and scalability to billions of compounds, which structure-based methods fail to achieve. To address aforementioned challenges, we propose a new model-agnostic anomaly detection-based uncertainty quantification method, **e**mbedding **M**ahalanobis **O**utlier **S**coring and **A**nomaly **I**dentification via **C**lustering (eMOSAIC). eMOSAIC uniquely quantifies distribution similarities or differences between the multi-modal representation of known cases and that of a new unseen one. We apply eMOSAIC to a multi-modal deep neural network model for multi-target ligand binding affinity predictions, leveraging a pre-trained strucrture-informed large protein language model. We extensively validate eMOSAIC in OOD settings, showing that it significantly outperforms state-of-the-art sequence-based deep learning and structure-based protein-ligand

docking (PLD) methods by a large margin as well as existing uncertainty quantification methods. This finding highlights eMOSAIC's potential for real-world polypharmacology and other applications.

Keywords: machine learning, deep learning, transfer learning, protein language model, compound screening, drug discovery

1 Introduction

Drug discovery is highly complex process, taking up to 15 years and costing billions of dollars [1], but its failure rate is extremely high due to lack of clinical efficacy or safety [2]. Many complex diseases such as neurological and mental disorders are multigenic, multi-factorial diseases. Thus, an effective therapy needs to modulate multiple genes [3, 4]. Polypharmacology that uses a single chemical to selectively modulate multiple drug targets has emerged as a new paradigm in drug discovery [5–10]. Screening a library of billions of compounds against multiple drug targets to identify lead compounds and subsequently optimizing their binding affinity selectivity profiles via medicinal chemistry for drug candidates are critical steps in polypharmacology [11]. A giga-scale compound screening is needed for several reasons. First, it can proportionally increase the likelihood of identifying more potent or selective ligands [12–15]. Secondly, the accessibility of analogues of lead compounds in the same chemical spaces can streamline the generation of meaningful structure–activity relationship (SAR) for lead optimization. Finally, large-scale libraries can expand chemical diversity, chemical novelty and patentability of the drug leads [16]. Although DNA-encoded libraries are an effective approach to generate and screen billions of compounds for a single target[17], it has limited chemistries due to DNA conjugation and a high rate of false positives by nonspecific binding of DNA labels. It is envisioned that computational approaches will facilitate large-scale compound screening for polypharmacology[11].

Protein-ligand docking (PLD) is commonly used for polypharmacology screening when the 3D structure of drug target is experimentally determined or computationally predicted [18]. However, PLD suffers from a high rate of false positives due to poor modeling of protein dynamics, solvation effects, crystallized water, and other challenges [19]. The reliability of PLD 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]. Moreover, PLD is computationally intensive, taking several seconds to score a protein-chemical pair. Thus, it is not feasible for PLD to screen billions of compounds against multiple targets for polypharmacology.

Recent advances in deep learning (DL) encourage increasing interest in applying DL to drug discovery [25, 26]. Sequence-based DL, an alternative to traditional docking methods, offers significant advantages. Sequence-based drug-target interaction (DTI) prediction, which uses 3D structure information only implicitly, enables fast DTI predictions when the inputs consist solely of a molecular description of the drug represented by a SMILES string or a 2D graph 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. The chemical space of small organic molecules is astronomical 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 make it challenging to train generalizable deep learning models for binding affinity predictions in an out-of-distribution (OOD) scenario for unexplored chemical space [33, 34], in which unseen testing chemicals are significantly different from training data.

Since drug discovery is a high stake process, making decisions based on incorrect predictions can lead to time and resource wastage. Knowing the confidence level of a prediction is crucial, as it allows researchers to make informed decisions about whether to consider or disregard specific drug leads. This necessitates an estimation of a reliability measure for individual predictions. The uncertainty of prediction comes from either the lack of labeled data or data noisiness. Gaussian Process (GP) is one of popular approach to the uncertainty quantification. Several works [35, 36] 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 the modification of the architecture of predictive models. Zeng and Gilford [37] implement an ensemble of NNs to obtain the uncertainty associated with the predictions for peptide-MHC binding. However, the ensemble-based technique is not as accurate as the GP algorithm for quantifying uncertainty. 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) [38, 39]. Many works have attempted to utilize conformal prediction for various aspects in drug discovery ranging from Quantitative Structure-Activity Relationships (QSAR), Quantitative Structure-Property Relationships (QSPR), pIC50, and toxicity prediction [40–42]. While conformal prediction in the case of classification provides accurate uncertainty quantification, these methods often struggle with efficiency in regression tasks by estimating large confidence intervals. Moreover, the exchangeablility 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 new model-agnostic anomaly detection-based uncertainty quantification method, **e**mbedding **M**ahalanobis **O**utlier **S**coring and **A**nomaly **I**dentification via **C**lustering (eMOSAIC). eMOSAIC explicitly models multi-modal distributions of embedding space and quantifies the similarity between embeddings of known cases and that of a new one. We apply eMOSAIC to a multi-modal deep learning network model for multi-target ligand binding affinity predictions, leveraging a pre-trained structure-informed large protein language model[43]. Under rigorous OOD benchmark studies, eMOSAIC significantly outperforms state-of-the-art deep learning models for binding affinity prediction. Interestingly, eMOSAIC also demonstrates superiority over PLD in terms of both accuracy and scalability. Furthermore,

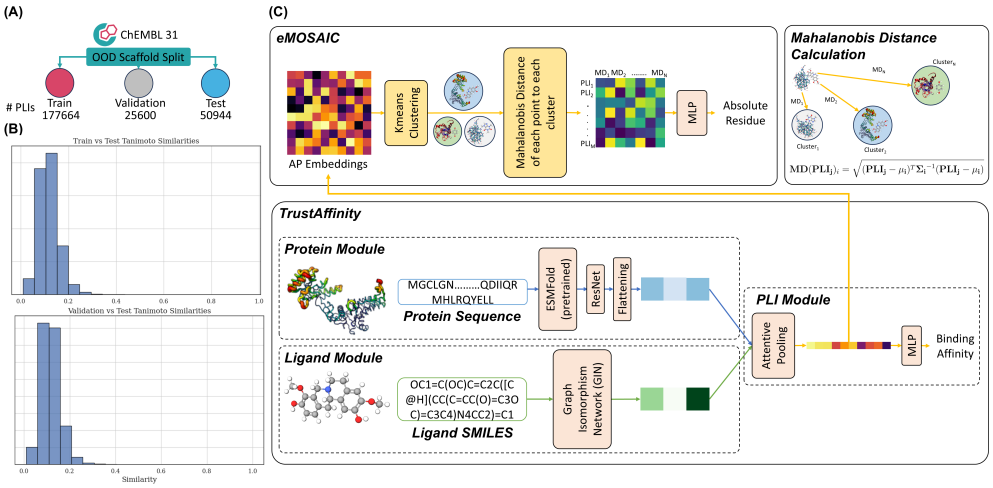

Fig. 1 Training data and architecture of eMOSAIC. (A) ChEMBL 31 dataset OOD Scaffold Split to ensure no overlap of scaffolds among the train, test, and validation set. (B) Tanimoto similarity histogram of the test vs validation and test vs train set, respectively. (C) Overall architecture of the pipeline with the binding affinity prediction module, TrustAffinity, and the model agnostic uncertainty quantification module, eMOSAIC.

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.

In summary, our contributions encompass both methodology development and translational applications:

1. We introduce eMOSAIC, a novel OOD multi-modal uncertainty quantification method that outperforms state-of-the-art methods and may have broad applications in addressing the OOD challenge.
2. By leveraging a pre-trained structure-informed protein language model, we demonstrate that eMOSAIC outperforms state-of-the-art deep learning methods in protein-ligand binding affinity prediction and polypharmacology compound screening.
3. Through rigorous benchmark studies, we show that sequence-based eMOSAIC significantly surpasses the existing structure-based compound screening methods dominant in drug discovery in terms of accuracy, reliability, and scalability.

2 Results

2.1 Overview of eMOSAIC

Figure 1 provides an overview of eMOSAIC. Similar to state-of-the-art uncertainty quantification methods such as conformal prediction, eMOSAIC utilizes the training data to train 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 from a trained task-specific deep learning model are clustered. Second, for a given validation case, the Mahalanobis distances between its embedding and each cluster of training embeddings are calculated, and the absolute residual of its prediction is obtained using the trained model. Finally, eMOSAIC is trained using these Mahalanobis distances as features and the absolute residuals as labels.

eMOSAIC is model-agnostic and can be applied to any trained task-specific DL models with embeddings. In this paper, we apply it to a protein-ligand binding affinity prediction model, called as TrustAffinity. The affinity prediction model takes a protein sequence and a ligand SMILES as inputs. Leveraging a pre-trained protein language model ESMFold [43] and a graph isomorphism network (GIN) [44], 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 is mentioned, it always refers to its application to the TrustAffinity model.

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 mimicking real-life scenario, where we are likely to encounter new potential chemicals which would have not been used for training process of the DL model. For the purpose of comparison, we also evaluate eMOSAIC in the in-distribution setting of random split.

We compared 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 with both sequence-based deep learning models including BACPI [45], DeepDTA [46], DeepPurpose [47] as well as a typical PLD method, AutoDock Vina [48], and deep learning-based docking model KarmaDock [49]. Along with this, we also compare the performance of eMOSAIC with other uncertainty quantification methods including GP-based RIO framework [36], Monte Carlo Dropout [50], and Conformal Prediction [38].

2.2 eMOSAIC improves OOD Binding Affinity Prediction

In the OOD setting, using structure-informed ESMFold model[43] for the protein sequence pre-training, TrustAffinity already outperforms state-of-the-art methods for the protein-ligand binding affinity predictions, as shown in Figure 2. eMOSAIC further boosts the performance of TrustAffinity by quantifying the uncertainty of predictions from it. eMOSAIC consistently outperforms the current state-of-the-art models regarding *all* four metrics in a regression task: **RMSE**, **MAE**, **Pearson Correlation**, and**Spearman Correlation**, as shown in Figure 2 A and B. eMOSAIC could successfully detect the anomalies and predict the uncertainty of predictions for improving the model performance. Although BACPI, DeepPurpose, and DeepDTA have an acceptable performance in the random split setting (Supplemental Table S1), their performances significantly drop in the scaffold split setting. In contrast, the correlation between predicted binding affinities by eMOSAIC and actual binding affinities

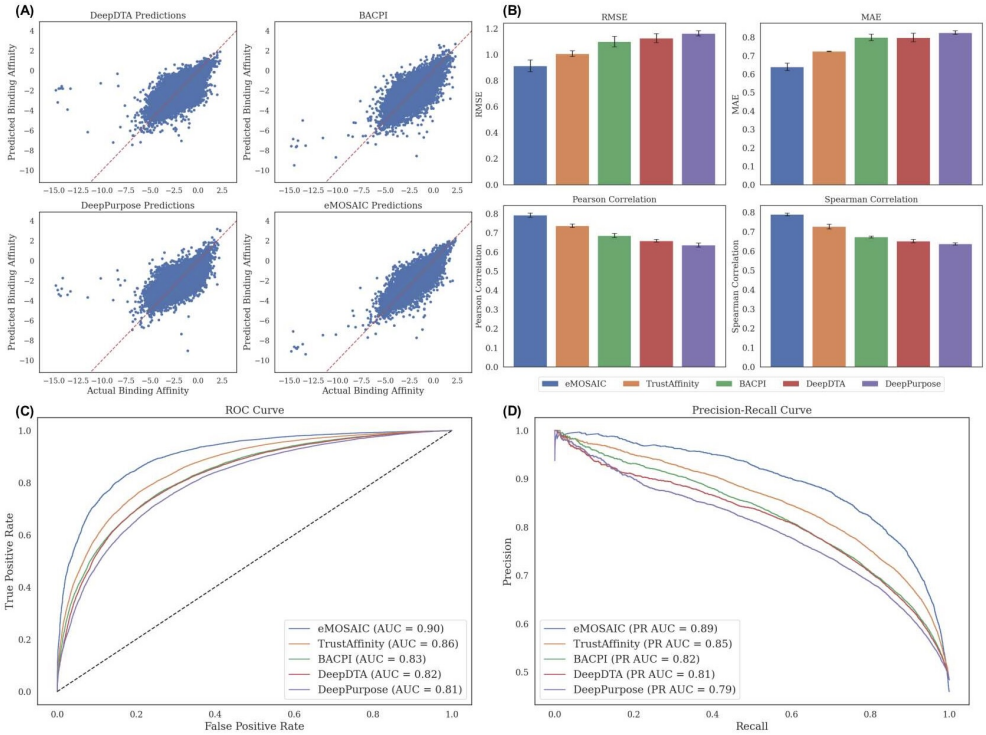

Fig. 2 Performance comparison of eMOSAIC with baselines TrustAffinity, DeepDTA, DeepPurpose, and BACPI in the OOD setting. (A) Scatterplot of actual binding affinities vs predicted binding affinities. (B) Barplots comparing the deviations and correlations between actual binding affinities and predicted binding affinities. (C) Receiver operating characteristic (ROC) curves for the binding classification. (D) Precision-recall (PR) curves for the binding classification.

remains high when testing chemicals have different scaffold from those in the training set. These findings clearly demonstrate the superior generalization power of eMOSAIC when predicting the binding affinity in the OOD setting. The predictions by eMOSAIC not only have higher correlation but also have significantly lower deviation as recorded by the RMSE on average 16.92%, 21.54%, and 18.97% lower, and MAE on average 20.03%, 22.55%, and 19.95% lower when compared to BACPI, DeepPurpose, and DeepDTA, respectively.

When using binding affinity greater than pKi of -2 as a threshold, eMOSAIC also significantly outperforms all baseline models in a classification task. As shown in Figure 2 C and D, the ROC AUC and PR AUC of eMOSAIC improve 4.6% and 4.7% over TrustAffinity, respectively. The improvement over other state-of-the-art models are more significant. They are at least 9.7% and 8.5%, respectively. The performance improvement in the low false positive rate is more significant. When the false positive rate is 0.05, the positive rate of eMOSAIC is over 50% higher than the existing methods, demonstrating the potential of eMOSAIC for compound screening.

As shown in Figure 2 A, the prediction errors mainly come from the low and high affinity regions where there are few training data. Although eMOSAIC can alleviate the problem compared to other methods, further improvement is needed.

2.3 eMOSAIC significantly outperforms structure-based protein-ligand docking for binding affinity predictions and compound screening

We further compare the performance of eMOSAIC with PLD, which has been widely applied in compound virtual screening. When tested on the scaffold split set, PLD shows a notably poor correlation between predicted and actual binding affinities, as demonstrated in Supplemental Figure S1 and Table S2. While the structure-based deep learning model, KarmaDock, offers moderate improvement over Autodock Vina, its performance remains inadequate. In contrast, eMOSAIC significantly enhances the correlation of predicted binding affinities, showing an eight-fold improvement over KarmaDock, as shown in Figure 3. For the docking experiment, the definition of binding pockets is critical. Two methods are applied to determining binding pockets: AlphaFill approach based on the alignment with co-crystallized complex structures [51] and predicted binding pockets from a machine learning method P2Rank [52]. Although AutoDock Vina performs better with AlphaFill-predicted binding pockets compared to P2Rank's, P2Rank outperforms AlphaFill when used with the machine learning-based method, KarmaDock.

For high-throughput screening applications, the ratio of high binding affinity hits in the top-n ranked predictions might be more interesting than the actual binding affinity. In this setting, eMOSAIC again significantly outperforms the PLD methods, as shown in Figure 3. The hit rate from eMOSAIC is approximately eight times higher than that of PLD methods for the threshold of $1\,\mu$M and three and a half times higher for the threshold of $10\,\mu$M. 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 the PLD.

Furthermore, eMOSAIC is several orders of magnitude faster than structure-based methods (Supplemental Figure S3). It takes less than 0.01 seconds for eMOSAIC, 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 can be completed in just a few days.

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 that is based on Gaussian process, Monte Carlo dropout, and conformal prediction. Figure 4A shows the predicted residue distribution for all predictions with actual absolute residue less than 0.5. It is clear to see that eMOSAIC has lowest upper bound of around 2.45 as opposed to the rest of the models and lowest variation. In

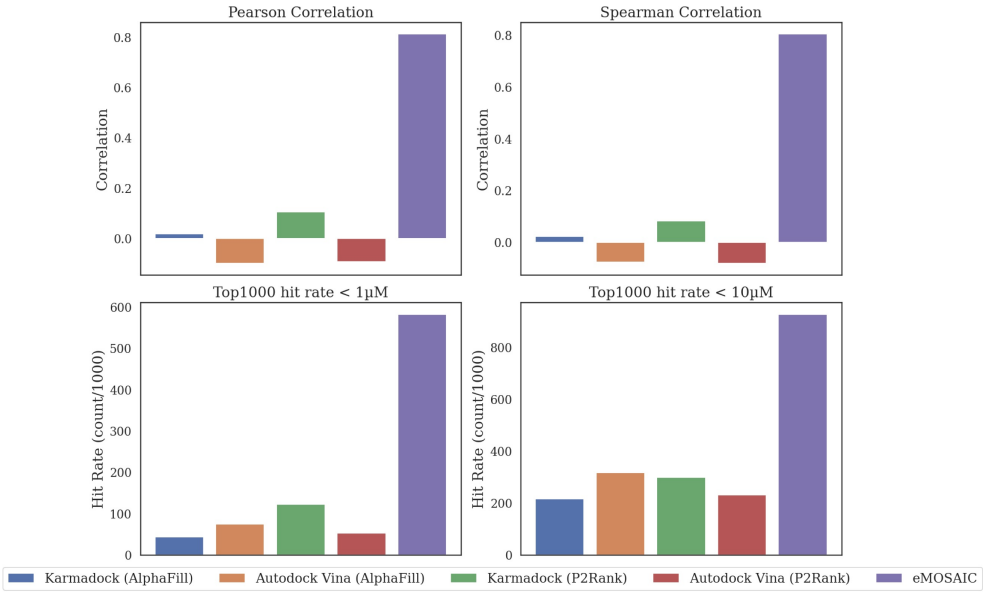

Fig. 3 Performance comparison of eMOSAIC with structure-based methods on thePearson Correlation, Spearman Correlation, and hit rates.

addition, eMOSAIC has the highest Pearson Correlation between predicted residues and actual residues, as shown in Figure 4B

When evaluated for their contributions to absolute binding affinity predictions, eMOSAIC's OOD uncertainty quantification module not only successfully identifies high-confidence, low-uncertainty PLIs but also outperforms other methods on average, as shown in Figure 4C. eMOSAIC has a 21.21% lower RMSE and 23.72% higher Pearson correlation compared to the second best method, conformal prediction. eMOSAIC can correctly identify high-confidence PLI interactions by calculating the Mahalanobis distance as an anomaly measure from the embeddings learned in the binding affinity detection module.

The GP-based RIO framework fails, due to the limited capability of the current kernels, which fail to effectively model the dependency between the PLI embeddings and residuals, resulting in less accurate residual correction and uncertainty estimation, and hence the poorest performance among all the methods.

The Monte Carlo dropout method is a crude approximation of uncertainty estimation, performing better than RIO. However, it is worse than the estimation of no uncertainty because the Bayesian posterior distribution estimated by dropout might be more complex [53]. Secondly, the Monte Carlo dropout explores limited configurations of potential weights, not all of them, resulting in incomplete uncertainty estimation [54].

Lastly, eMOSAIC performs slightly better than conformal prediction, an extensively studied method for uncertainty quantification. Interestingly, the high confident

predictions from eMOSAIC and the conformal prediction is complementary, as shown in supplemental Figure S2. The Pearson Correlation between their predicted values is 0.38. It would be interesting to integrate the conformal prediction and eMOSAIC to provide a more robust statistics test for the uncertainty.

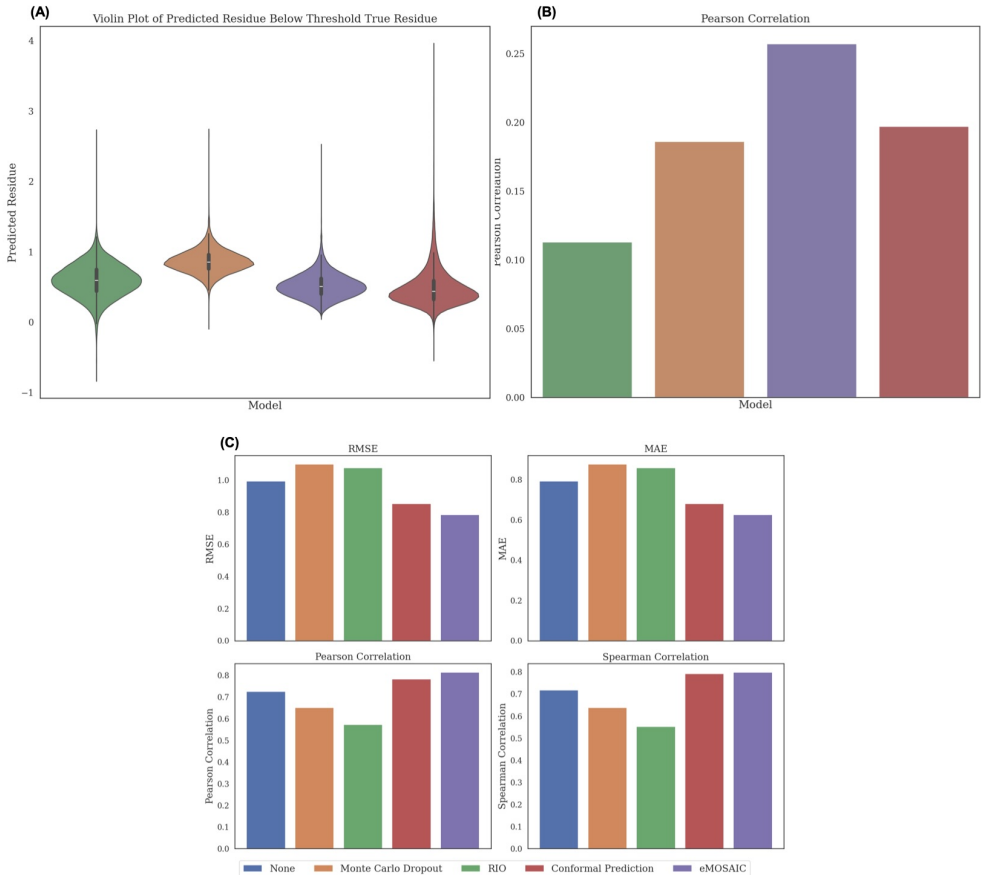

Fig. 4 Performance comparison of eMOSAIC with state-of-the-art uncertainty quantification methods. (A) Violin Plot of predicted residue points below the high confidence threshold of 0.5. (B)Pearson Correlation of predicted residue and true residue. (C) Errors and correlations of actual and predicted binding affinities of the high confidence predictions as selected by the various uncertainty quantification models.

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 [55]. Along with these, protein kinase inhibitors have become a critical class of drugs, especially in oncology. Up

to 33% of the drug development process target these kinases [56]. 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 PLIs are used to identify meaningful interactions. We apply a threshold of less than 100 μM of the predicted pKi values, isolating strong interactions that are likely to be therapeutically significant. The evaluation is based on a multi-label classification (threshold is Ki = 100 μM) to determine the potential of ligands to selectively act across different therapeutic targets and ensures that only interactions with high clinical relevance are considered. We evaluate the models using multi-label accuracy, precision, recall, 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 the performance of eMOSAIC is significantly improved over other methods regarding all evaluation metrics for the GPCR polypharmacology. The accuracy, precision, recall, F1 score improve 21.4%, 19.7%, 20.4%, and 26.6%, respectively, while the Hamming distance reduces 164.2% compared with the second best method. For kinase polypharmacology, docking methods have the best sensitivity but poor specificity in consistent with the observations that PLD has a high false positive rate in general. eMOSAIC has a balanced sensitivity and specificity, thus it has the best accuracy and F-1 score. Among all evaluation metrics, the Hamming distance is the best for evaluating multi-label classification performance. eMOSAIC reduces the Hamming distance by 18.2% compared with the second best method.

Table 1 Performance comparison of polypharmacology compound screening on GPCR and Kinase Proteins. The best and the second best performed methods are in bold and underline, respectively. (***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	0.502	0.501	0.484	0.387	0.499
	KarmaDock	0.51	0.51	0.143	0.124	0.49
	DeepPurpose	0.724	0.696	0.678	0.668	0.277
	DeepDTA	0.743	0.716	0.699	0.688	0.256
	BACPI	0.74	0.74	0.722	0.652	0.26
	eMOSAIC	**0.902**	**0.886**	**0.869**	**0.871**	**0.098**
Kinase	AutoDock	0.366	0.367	**0.858**	0.416	0.633
	KarmaDock	0.414	0.414	0.735	0.38	0.586
	DeepPurpose	0.698	0.609	0.637	0.595	0.302
	DeepDTA	0.714	0.635	0.627	0.605	0.286
	BACPI	0.681	**0.681**	0.407	0.312	0.319
	eMOSAIC	**0.758**	0.662	0.648	**0.622**	**0.242**

3 Discussion

In this work, we propose eMOSAIC, a novel 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, 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 PLD, thereby rendering it well-suited for deployment in automated drug discovery processes to leverage uncertainty-based methodologies.

Despite eMOSAIC's superior performance, it has certain limitations that can be further addressed. First, it lacks the ability to predict the binding pose of protein-ligand interactions (PLI), a crucial factor in the drug discovery process. eMOSAIC's performance could be enhanced through multi-task learning, which includes predicting binding poses, binding affinity, and classifying PLI interactions. Second, the partitioning of the embedding space affects eMOSAIC's performance. Using improved clustering methods, such as supervised contrastive learning, could yield better results. Finally, incorporating semi-supervised techniques[57] during training may boost its generalization to OOD data.

4 Method

4.1 eMOSAIC for uncertainty quantification

4.1.1 eMOSAIC algorithm

Figure 1 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 [58]. 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, μ and Σ 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 [59–62]. 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 of a deep learning model, a binding affinity module in this paper, we extract these attentive pooling embeddings which represent the PLI. Now we perform k-means clustering of the PLI embeddings from the training set, and for each cluster we obtain the mean (μ) and variance matrix ($\mathbf{\Sigma}$) for Mahalanobis distance calculation.

Given an embedding of a unseen PLI, we calculate the Mahalanobis distance to each cluster. Then we use these distances obtained as features to train a simple multi-layer perceptron (MLP) to predict the absolute residue, which is defined as the difference between the predicted binding affinity and the true binding affinity. This way the MLP learns to identify the PLI anomalies as predicted by the model. For outlier detection, we select points whose predicted residue is less than 0.5, ensuring that only high-confidence predictions are filtered.

4.1.2 Training of eMOSAIC

We first train the binding affinity prediction module. After training, we obtain the embeddings with the best model based on the validation set, and train the OOD uncertainty quantification module to identify outliers. More details about the hyperparameters and configuration of the eMOSAIC Model are present in the appendix.

4.2 Binding Affinity Module

The binding affinity module consists of three sub-modules - the protein sequence module, the ligand processing module, and lastly the PLI module. All of these modules work together to 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, 45, 46, 63], but also their 3D structure [23, 43]. Protein sequences contain information that can be used to infer protein structure, function, and family [43], making them a rich source of data for machine learning models [23, 43]. Large datasets of protein sequences are available [23, 64, 65], enabling machine learning frameworks to learn high-level, general representations of proteins. We utilize ESMFold [43] 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 numbers and due to constraints with 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 [66] 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 learnt using the graph isomorphism network (GIN) [67]. 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 PLI Module

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

4.3 Experiments

4.3.1 Dataset

We train TrustAffinity on the ChEMBL31 database [30], which consists of 350400 PLI pairs. In the experiments, we split the dataset into training, testing, and validation set by 7:2:1. Negative log 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[68] such that the chemicals in the testing set have different scaffolds from those in the training/validation set. Scaffold split ensures that there was no overlap of scaffold in the training, testing and validation set. This was done to validate the model's generalization power in 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.

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 with both sequence-based deep learning models such as BACPI [45], DeepDTA [46], DeepPurpose [47] as well as a typical PLD method, AutoDock Vina [48], and deep learning-based docking model KarmaDock [49] on the OOD test set. 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 [51] 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 other proteins not in Alphafill dataset, P2Rank [52] 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 measures 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 [36], Monte Carlo Dropout [50], and Conformal Prediction [38].

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 [39], and several methods are available that describe various variants of Conformal Prediction [69]. 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 [39]. In our case, we use the attentive pooling embeddings to train an MLP for predicting the absolute error in case of CP baseline. We then select the points which have a lower than average length of confidence interval in the test set. We do the same for obtaining the high confidence points in case of Monte-Carlo Dropout method.

4.3.3 Evaluation

We evaluate eMOSAIC's performance using root mean square (RMSE), mean absolute error (MAE), Pearson correlation coefficient (r) and Spearman's rank correlation coefficient (ρ).

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., Chen, M., Liu, Y., He, D., Zhang, S., Mura, C., Bourne, P.E., Xie, L.: Exploration of dark chemical genomics space via portal learning: applied to targeting the undruggable genome and covid-19 anti-infective polypharmacology. Research Square (2021)

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

[36] 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 (2019)

[37] 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)

[38] Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. arXiv preprint

arXiv:1301.7375 (2013)

[39] 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)

[40] 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)

[41] 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)

[42] 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)

[43] 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)

[44] 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)

[45] 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)

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

[47] 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)

[48] 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)

[49] 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)

[50] 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

[51] 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)

[52] 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)

[53] 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

[54] 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)

[55] 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)

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

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

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

[59] 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

[60] 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

[61] 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

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

[63] 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)

[64] 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)

[65] 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)

[66] 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)

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

[68] 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, Inc.", ??? (2019)

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