Trustworthy protein-ligand binding affinity prediction using large protein language model

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

[1]Department of Computer Science, Hunter College, The City University of New York
New York City, United States of America
[2]Ph.D. Programs in Computer Science, Biology & Biochemistry
The Graduate Center, The City University of New York
New York City, United States of America
[3]Helen and Robert Appel Alzheimer's Disease Research Institute
Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University
ab11588@hunter.cuny.edu, lei.xie@hunter.cuny.edu

Abstract

Accurate, reliable and scalable predictions of protein-ligand binding affinity have a great potential to accelerate drug discovery. Despite considerable efforts, three challenges remain: out-of-distribution (OOD) generalizations for understudied proteins or compounds from unlabeled protein families or chemical scaffolds, uncertainty quantification of individual predictions, and scalability to billions of compounds. We propose a sequence-based deep learning framework, TrustAffinity, to address aforementioned challenges. TrustAffinity synthesizes a structure-informed large protein language model and 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). We extensively validate TrustAffinity in multiple OOD settings. TrustAffinity significantly outperforms state-of-the-art computational methods by a large margin and at least three orders of magnitude of faster than protein-ligand docking, highlighting its potential in real-world drug discovery. We further demonstrate TrustAffinity's practicality through an Opioid Use Disorder lead discovery case study.

Introduction

Drug discovery is very complex process, taking up to 15 years and costing billions of dollars (Hughes et al. 2011). The advent of increased available protein structure and chemical genomics data and ever-improving deep learning algorithms has inspired the application of computational science and Artificial Intelligence (AI) to drug discovery (Paul et al. 2021; Burley et al. 2023; Liu et al. 2015), speculating its potential to accelerate discovering new therapeutics for unmet medical needs (Mishra 2018). Screening a library of billions of compounds against a drug target to identify lead compounds and subsequently optimizing their binding affinities via medicinal chemistry for drug candidates are critical steps in a predominant target-based drug discovery process. In the paradigm of target-based drug discovery, an ideal drug should have a high binding affinity towards a specific target protein to ensure lower concentration usage, but not bind to other proteins to reduce side effects from off-targets. Thus, accurate, reliable, and scalable prediction of protein-ligand binding affinities across the human proteome is a central task of computer-aided drug discovery.

Despite considerable efforts, the performance of existing protein-ligand binding affinity prediction methods remains poor in terms of accuracy, scalability, and reliability. The generalization power of deep learning methods for protein-ligand interaction predictions is weak. Current works mainly focus on well-studied drugs and their analogs and pharmaceutically characterized targets(Xu, Ru, and Song 2021; Bagherian et al. 2021). Few machine learning methods can reliably predict protein-ligand interactions between understudied proteins whose functions are not well characterized and chemicals with novel scaffold. Unfortunately, the pharmaceutical characterization of the human proteome is highly biased (Haynes, Tomczak, and Khatri 2018; Wood et al. 2019; Stoeger et al. 2018; Oprea et al. 2018). More than 95% of human proteins do not have known small molecule ligands (Cai et al. 2023; Gerlt et al. 2011; Carvalho-Silva et al. 2019; Kustatscher et al. 2022). Additionally, the chemical space of small organic molecules is astronomical vast. Although the number of possible small organic molecules is approximate 10^{60} (Lemonick 2020), only around 10^6 compounds have annotated protein targets (Mendez et al. 2019; Kim et al. 2023; Gilson et al. 2016). The scarcity of ligand information for the majority of proteins and 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 understudied proteins and unexplored chemical space(Nguyen et al. 2019; Karimi et al. 2019), in which unseen testing data (proteins or 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 predic-

tion comes from either a data distribution shift or model bias and variance. The application of uncertainty quantification to the field of biology is relatively limited. Gaussian Process (GP) is one of popular approach to the uncertainty quantification. Several works (Hie, Bryson, and Berger 2020)(Qiu, Meyerson, and Miikkulainen 2019) 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 (Zeng and Gifford 2019) 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.

To overcome the aforementioned limitations, we propose a new deep learning framework, TrustAffinity, which uses a pre-trained structure-informed protein language model (Lin et al. 2023) for exploring new chemical genomics space. Furthermore, we develop a new uncertainty quantification methods eMOSAIC that utilizes embedding clustering and Mahalanobis distance for anomaly detection. Under a rigorous benchmark study, our proposed method significantly outperforms state-of-the-art deep learning model for binding affinity prediction in the OOD scenario by a large margin. Interestingly, TrustAffinity also demonstrates superiority over protein-ligand docking in terms of both accuracy and scalability. We further demonstrate the applicability of TrustAffinity to real-world drug discovery in a case study on lead discovery for Opioid Use Disorder (OUD). Thus, TrustAffinity represents a significant advance in deep learning applications to drug discovery.

In summary, our contributions include the following key points:

1. We introduce TrustAffinity, a novel trustworthy deep learning framework for accurate, reliable and scalable binding affinity prediction in the OOD scenario.
2. Through rigorous benchmark studies, we demonstrate the superior performance of TrustAffinity compared to other state-of-the-art methods.
3. We apply TrustAffinity to a drug design case, and showed that efficient sequence-based TrustAffinity significantly outperforms structure-based protein-ligand docking.

Related Works

Protein-ligand docking (PLD) The prediction of binding affinity of a protein-ligand complex can be categorized into two major approaches: structure-based methods and structure-free methods. Structure-based methods rely on three-dimensional (3D) information about the protein-ligand complex to predict their binding affinity and often the underlying mechanism associated with their interactions. PLD is the basis of structure-based methods and provides substantial biological interpretability, as the results provide insights into spatial configurations of protein-ligand complex and information about the various binding sites present on the protein. However, PLD is computationally expensive and relies heavily on the availability of 3D structural data (Leach, Shoichet, and Peishoff 2006). Furthermore, PLD is highly sensitive to the conformational state of structures, which can often lead to inaccurate predictions (Grinter and Zou 2014).

Machine learning methods for binding affinity predictions In the past decade, an increasing number of machine learning and deep learning algorithms have been incorporated into the prediction of protein-ligand interactions (PLIs) and their binding affinity from 3D information (Jiménez et al. 2018; Rezaei et al. 2020; Wallach, Dzamba, and Heifets 2015; Feinberg et al. 2018). These structure-based machine learning methods use a variety of methods to incorporate 3D information into their models for predicting binding affinity. K_{DEEP} (Jiménez et al. 2018) and DeepAtom (Rezaei et al. 2020) represent both the protein and the ligand as 3D voxels and use deep convolutional neural networks (DCNNs) to account for molecular interactions within the protein-ligand complex. AtomNet (Wallach, Dzamba, and Heifets 2015) also uses a form of voxel representation. Instead of representing the entire protein, it represents only the binding site of the target protein. PotentialNet (Feinberg et al. 2018) utilizes the power of graph neural networks (GNNs) to learn powerful representations for both proteins and ligands for predicting binding affinity. While these methods have demonstrated their effectiveness in the prediction of binding affinity, they rely on extensive 3D data and are computationally expensive.

Recent studies have demonstrated the success when using recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) to directly utilize protein sequences and ligand simplified molecular-input line-entry system (SMILES) (Karimi et al. 2019; Li et al. 2022; Wang et al. 2021; Öztürk, Özgür, and Ozkirimli 2018). DeepDTA (Öztürk, Özgür, and Ozkirimli 2018) performs training on label encoded SMILES and protein sequences after using 1D CNNs to obtain the representations. DeepAffinity (Karimi et al. 2019) follows a similar approach but instead of using protein sequence and secondary structure as inputs. Its architecture combines RNNs, attention mechanism, and 1D CNNs. BACPI (Li et al. 2022) deploys the bi-directional attention mechanism which facilitates the interaction between the ligand and protein representations, and applied graph attention networks for learning ligand representations. DeepDTAF (Wang et al. 2021) is similar to DeepAffinity, it also encodes secondary structural information along with one-hot encoded protein sequence representation based on amino acids and utilizes the protein pocket information as a set of features. There are two main limitations with both the structure-based and structure-free deep learning models. Firstly, they don't perform rigorous OOD testing, which is the reason why they fail to perform well on real-world unseen data. While some efforts have been made to assess the generalizability of models. Secondly, these models lack the capability to provide confidence estimates for their predictions. It is well-known that machine learning models are not perfectly accurate, therefore having knowledge of confidence associated with predictions in the sensitive field of drug discovery is crucial.

Out-Of-Distribution generalization The out-of-distribution (OOD) generalization problem arises when the distribution of test data significantly deviates from that of the training data. Notably, this deviation remains undisclosed and uncharacterized during the training phase of the model (Quinonero-Candela et al. 2008). Existing methods in machine learning commonly assumed that both training and testing data are independent and identically distributed (iid) (Shen et al. 2021). This assumption does not hold in many real-world scenarios, especially when dealing with new PLIs in new environments. The main aim is to evaluate how well the model would adapt and perform to these deviations when confronted with real-world unseen data. deep learning is susceptible to performance degradation under these deviations. It has inspired researchers to focus on tackling the issue of OOD generalization (Cai et al. 2023; Hendrycks and Gimpel 2016; Liang, Li, and Srikant 2017; Yang et al. 2022; He, Shen, and Cui 2021). PortalCG (Cai et al. 2023) is one of few sequence-based PLI prediction methods that address the OOD generalization problem. It utilizes sequence pre-training, structure-based fine-tuning, and meta-learning to improve the model performance under an OOD setting. However, PortalCG focuses on binary classification (binding or non-binding) of protein-ligand pairs rather than predicting binding affinity.

Uncertainty quantification in biology The knowledge of uncertainty is highly crucial in the safety-sensitive applications that involve human lives. Thus, uncertainty quantification has become more common in various fields such as computer vision (Kendall and Gal 2017; Kendall and Cipolla 2016; Kendall, Badrinarayanan, and Cipolla 2015) and natural language processing (Xiao and Wang 2019; Hu et al. 2023; Lin, Hilton, and Evans 2022). However, the application of uncertainty quantification to the field of biology is relatively limited. Zeng and Gilford (Zeng and Gifford 2019) implement an ensemble of NNs to obtain the uncertainty associated with the predictions for peptide-MHC binding for an improved therapeutic drug design process. But, these ensemble-based techniques aren't as accurate as the Gaussian Process (GP) algorithms for prediction of uncertainty. GPs provide a distribution over functions, enabling them to capture the inherent uncertainty in predictions and make more reliable inferences. By providing this distribution, GPs are able to capture the possible outcomes and likelihood facilitating more informed decision making process. Hie et al. (Hie, Bryson, and Berger 2020) propose a combined GP and multi-layer perceptron (MLP) approach for various biological tasks, including predicting protein-kinase binding affinity, generative compound design with protein kinase B activity, and protein fluorescence prediction. For binding affinity prediction, they observe that GP-based models provide accurate, low-uncertainty predictions, enhancing the selection of promising compound-kinase pairs for validation. In generative compound design, GP-based models outperform MLP-based models in terms of binding affinity. Along with simply using GP, they use it alongside MLP (GP+MLP) as proposed by Qiu et al. (Qiu, Meyerson, and Miikkulainen 2019) and notice similar or better results when compared to the GP.

LLMs for Biology The application of Large Language Models (LLMs) in domains like natural language processing (NLP) and computer vision (CV) is well-established. LLMs have significantly evolved their ability to comprehend human language and context, which has led to improved performance on multiple downstream NLP tasks (Devlin et al. 2018; Brown et al. 2020). However, the exploration of their potential and the integration of LLMs in solving complex biological problems are only just beginning. The vast protein sequence data has motivated using LLMs, specifically protein language models (PLMs) such as ESM-2 (Lin et al. 2023), AlphaFold (Jumper et al. 2021), and ProtTrans (Elnaggar et al. 2021). They have successfully demonstrated their utility in predicting protein structures from their sequence by capturing secondary and tertiary information. PLMs, such as DISAE (Cai et al. 2021a), built upon the ALBERT (Lan et al. 2019) architecture, showcase that direct application of LLMs from NLP may not always yield robust performance in different domains. Therefore, including domain knowledge through pretraining and fine-tuning is essential in adapting LLMs. DISAE achieves SOTA performance on CPI classification by including domain knowledge, notably in challenging scenarios such as predicting ligands for orphan proteins, which have structures significantly dissimilar from known proteins. PortalCG (Cai et al. 2021b, 2023) is another work that incorporates domain knowledge through structure-regularized pretraining to improve the DISAE PLM protein descriptor and finally uses meta-learning for accurate dark protein chemical interaction prediction. Wu et al. (Wu et al. 2023) illustrate the integration of PLMs with geometric neural networks for downstream tasks such as binding affinity prediction, protein-protein interface prediction, protein-protein rigid-body docking, and model quality assessment. They obtained over 20% performance improvement over the baseline method, indicating that PLM knowledge strongly improves the capabilities of the geometric neural networks. ProGen (Madani et al. 2020), an LLM capable of generating protein sequences with specific functions. They show that fine-tuning ProGen leads to more controlled protein sequence generation. DG-Affinity (Yuan et al. 2023), for predicting antigen-antibody affinity while adapting pre-trained PLMs TAPE for antigen sequence and Ablang for antibody sequence and utilizing it for their specific task. However, PLMs have not been applied to address OOD protein-ligand binding affinity predictions.

Method

Our method, TrustAffinity, consists of two main modules, a binding affinity prediction module and an uncertainty quantification module eMOSAIC. They are used together to improve each other's performance. The following subsections provide motivations and details for each of these modules in TrustAffinity framework. Figure 1 provides an overview of TrustAffinity.

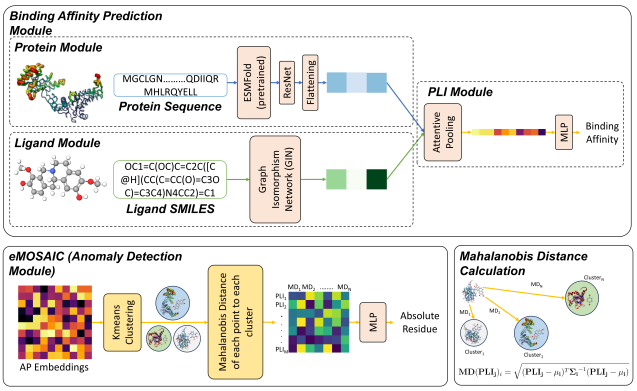

Figure 1: Overview of TrustAffinity. TrustAffinity consists of two main modules, binding affinity prediction module and anomaly detection module (eMOSAIC), which work in sync to provide the binding affinity and detect the anomalies within the predictions.

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.

Protein Sequence Module Protein sequence representation is one of the most vital components in the machine learning frameworks for predicting not only PLIs (Karimi et al. 2019; Li et al. 2022; Wang et al. 2021; Öztürk, Özgür, and Ozkirimli 2018), but also their 3D structure (Jumper et al. 2021; Lin et al. 2023). Protein sequences contain information that can be used to infer protein structure, function, and family (Lin et al. 2023), making them a rich source of data for machine learning models (Jumper et al. 2021; Lin et al. 2023). Large datasets of protein sequences are available (Jumper et al. 2021; Mitchell et al. 2020; Suzek et al. 2015), enabling machine learning frameworks to learn high-level, general representations of proteins. We utilize ESMFold (Lin et al. 2023) 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 (He et al. 2016) 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.

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) (Xu et al. 2018). 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.

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.

Uncertainty Quantification Module

In our uncertainty quantification module, 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 (Mahalanobis 2018). For a given point in a distribution, the MD is defined as follows:

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

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 (Rippel, Mertens, and Merhof 2021; Gjorgiev and Gievska 2020; Anthony and Kamnitsas 2023; Prekopcsák and Lemire 2012). 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 the binding affinity module, 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.

Training of TrustAffinity

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 eMOSAIC to identify outliers. More details about the hyperparameters and configuration of the TrustAffinity Model are present in the appendix.

Experiments

Experimental Settings

Dataset: We train TrustAffinity on the ChEMBL31 database(Mendez et al. 2019), 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(Ramsundar et al. 2019) 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 multiple OOD settings. Moreover, considering the vast chemical space for drug discovery, it is very likely that the model will encounter unknown and new scaffolds.

Baseline models: We compare our model's high confidence, low uncertainty predictions with the current state-of-the-art model, BACPI, which uses a novel bi-directional attention mechanism for modeling interaction between the protein and ligand (Li et al. 2022), on the OOD test and validation set. BACPI (Li et al. 2022) outperforms other state-of-the-art models DeepAffinity (Karimi et al. 2019), DeepPurpose (Huang et al. 2020), MONN (Li et al. 2020). Thus, we do not directly compare TrustAffinity with these models. We also compared TrustAffinity with a typical PLD method, AutoDock Vina (Trott and Olson 2010), on an external Dopamine Receptor antagonist data set (Cai et al. 2023), which contains 65 new compounds screened for there sub-types of Dopamine Receptors including DRD1, DRD2, and DRD3.

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

Results and Discussions

Improved OOD binding affinity prediction:

We evaluated the performance of TrustAffinity in a scaffold-based OOD setting, where chemicals in the testing set have different chemical scaffolds from those in the training/validation set. For the purpose of comparison, we also evaluate TrustAffinity in the in-distribution setting of random split.

In the OOD setting, TrustAffinity consistently outperforms the current state-of-the-art BACPI model regarding *all* four metrics in both IID (random split) and OOD (scaffold split) settings, as shown in Figure 2. Our model could successfully detect the anomalies and predict the uncertainty of predictions for improving the model performance. Although BACPI has an acceptable performance in the random split setting, its performance significantly drops in the scaffold split setting. In contrast, the correlation between predicted binding affinities by TrustAffinity and actual binding affinities remains high when testing chemicals have different scaffold from those in the training set. These findings clearly demonstrate the superior generalization power of TrustAffinity when predicting the binding affinity in an OOD setting. The predictions by TrustAffinity not only have higher correlation but also have significantly lower deviation as recorded by the RMSE (on average 20.39% lower), and MAE (on average 25.19% lower) when compared to BACPI.

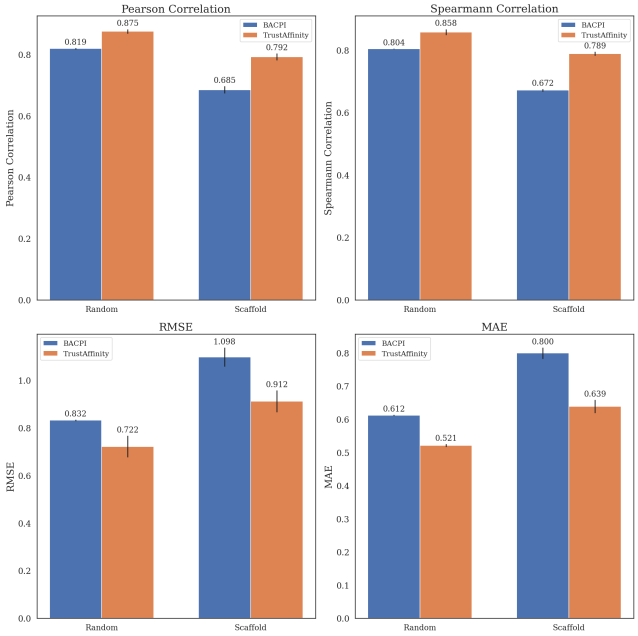

Figure 2: Performance comparison of TrustAffinity with BACPI on test set scaffold split and random split.

Table 1: OOD DRD set results (for AutoDock Vina, the predicted docking score was multiplied by a constant which is best suited for obtaining the final pKi value aligned with the actual pKi values)

Method	RMSE ↓	MAE ↓	r ↑	ρ ↑
AutoDock Vina	1.179	1.031	0.308	0.334
BACPI	2.523	2.181	0.103	0.122
TrustAffinity	**0.919**	**0.739**	**0.617**	**0.614**

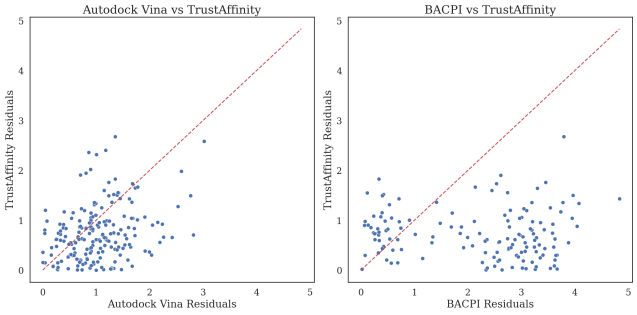

Figure 3: Comparative residual plot of TrustAffinity vs AutoDock Vina (PLD), and TrustAffinity vs BACPI.

Case study on lead discovery for OUD: Our method is able to achieve considerably higher Pearson and Spearman correlation as compared to AutoDock Vina and BACPI when tested on an external OOD DRD antagonist dataset, as shown in Table 1. Figure 3 provides more detailed performance comparisons between TrustAffinity, AutoDock Vina as well as BACPI. In general, more PLI pairs predicted from TrustAffinity have smaller residues (errors) than those from AutoDock Vina and BACPI (lower corner in Figure 3). Our method simply utilizes the protein sequence and chemical SMILES, while AutoDock Vina utilizes the 3D structures to obtain a docking score. Moreover, we find that TrustAffinity is able to predict the binding affinity as well as the uncertainty associated with it for a protein ligand pair in approximately 0.003 seconds, making it roughly three orders of magnitude of faster than AutoDock Vina, which is known to take several seconds to minutes (Huang 2018), even when considering the best case scenario for AutoDock Vina. Thus TrustAffinity could be a potentially powerful tool for screening novel compounds in the drug discovery pipeline due to its high accuracy, a reliable automated component based on the uncertainty predictions, and scalability.

Conclusion

In this work, we propose TrustAffinity, 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 TrustAffinity, 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. However, our method has certain limitations. It is unable to predict the binding pose of the PLI, which is also crucial for further drug discovery pipeline. Additionally, the performance of TrustAffinity can be further improved when it conducts multi-task learning including the prediction of binding poses, binding affinity and binary classification of PLI interactions. As part of future work, we would like to explore multi-task predictions and the incorporation of semi-supervised techniques such as student-teacher model training for even better OOD generalization.

References

Anthony, H.; and Kamnitsas, K. 2023. 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*, 136–146. Springer.

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

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V.; Dalenberg, K.; Duarte, J. M.; et al. 2023. RCSB Protein Data Bank (RCSB. org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic acids research*, 51(D1): D488–D508.

Cai, T.; Lim, H.; Abbu, K. A.; Qiu, Y.; Nussinov, R.; and Xie, L. 2021a. Msa-regularized protein sequence transformer toward predicting genome-wide chemical-protein interactions: Application to gpcrome deorphanization. *Journal of chemical information and modeling*, 61(4): 1570–1582.

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

Cai, T.; Xie, L.; Zhang, S.; Chen, M.; He, D.; Badkul, A.; Liu, Y.; Namballa, H. K.; Dorogan, M.; Harding, W. W.; et al. 2023. End-to-end sequence-structure-function meta-learning predicts genome-wide chemical-protein interactions for dark proteins. *PLOS Computational Biology*, 19(1): e1010851.

Carvalho-Silva, D.; Pierleoni, A.; Pignatelli, M.; Ong, C.; Fumis, L.; Karamanis, N.; Carmona, M.; Faulconbridge, A.;

Hercules, A.; McAuley, E.; et al. 2019. Open Targets Platform: new developments and updates two years on. *Nucleic acids research*, 47(D1): D1056–D1065.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127.

Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; and Pande, V. S. 2018. PotentialNet for molecular property prediction. *ACS central science*, 4(11): 1520–1530.

Gerlt, J. A.; Allen, K. N.; Almo, S. C.; Armstrong, R. N.; Babbitt, P. C.; Cronan, J. E.; Dunaway-Mariano, D.; Imker, H. J.; Jacobson, M. P.; Minor, W.; et al. 2011. The enzyme function initiative. *Biochemistry*, 50(46): 9950–9962.

Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; and Chong, J. 2016. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1): D1045–D1053.

Gjorgiev, L.; and Gievska, S. 2020. 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*, 42–55. Springer.

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

Haynes, W. A.; Tomczak, A.; and Khatri, P. 2018. Gene annotation bias impedes biomedical research. *Scientific reports*, 8(1): 1362.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Y.; Shen, Z.; and Cui, P. 2021. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110: 107383.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

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

Hu, M.; Zhang, Z.; Zhao, S.; Huang, M.; and Wu, B. 2023. Uncertainty in Natural Language Processing: Sources, Quantification, and Applications. *arXiv preprint arXiv:2306.04459*.

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

Huang, S.-Y. 2018. Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges. *Briefings in bioinformatics*, 19(5): 982–994.

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

Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; and De Fabritiis, G. 2018. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling*, 58(2): 287–296.

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

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

Kendall, A.; Badrinarayanan, V.; and Cipolla, R. 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*.

Kendall, A.; and Cipolla, R. 2016. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, 4762–4769. IEEE.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. 2023. PubChem 2023 update. *Nucleic acids research*, 51(D1): D1373–D1380.

Kustatscher, G.; Collins, T.; Gingras, A.-C.; Guo, T.; Hermjakob, H.; Ideker, T.; Lilley, K. S.; Lundberg, E.; Marcotte, E. M.; Ralser, M.; et al. 2022. Understudied proteins: opportunities and challenges for functional proteomics. *Nature Methods*, 19(7): 774–779.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Leach, A. R.; Shoichet, B. K.; and Peishoff, C. E. 2006. Prediction of protein- ligand interactions. Docking and scoring: successes and gaps. *Journal of medicinal chemistry*, 49(20): 5851–5855.

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

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

Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; and Zeng, J. 2020. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems*, 10(4): 308–322.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

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

Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; and Wang, R. 2015. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3): 405–412.

Madani, A.; McCann, B.; Naik, N.; Keskar, N. S.; Anand, N.; Eguchi, R. R.; Huang, P.-S.; and Socher, R. 2020. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*.

Mahalanobis, P. C. 2018. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80: S1–S7.

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. 2019. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1): D930–D940.

Mishra, V. 2018. Artificial intelligence: the beginning of a new era in pharmacy profession. *Asian Journal of Pharmaceutics (AJP)*, 12(02).

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. 2020. MGnify: the microbiome analysis resource in 2020. *Nucleic acids research*, 48(D1): D570–D578.

Nguyen, T.; Le, H.; Le, T.; and Venkatesh, S. 2019. Prediction of drug–target binding affinity using graph neural networks. *BioRxiv*, 684662.

Oprea, T. I.; Bologa, C. G.; Brunak, S.; Campbell, A.; Gan, G. N.; Gaulton, A.; Gomez, S. M.; Guha, R.; Hersey, A.; Holmes, J.; et al. 2018. Unexplored therapeutic opportunities in the human genome. *Nature reviews Drug discovery*, 17(5): 317–332.

Öztürk, H.; Özgür, A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; and Tekade, R. K. 2021. Artificial intelligence in drug discovery and development. *Drug discovery today*, 26(1): 80.

Prekopcsák, Z.; and Lemire, D. 2012. Time series classification by class-specific Mahalanobis distance measures. *Advances in Data Analysis and Classification*, 6: 185–200.

Qiu, X.; Meyerson, E.; and Miikkulainen, R. 2019. Quantifying Point-Prediction Uncertainty in Neural Networks via Residual Estimation with an I/O Kernel. In *International Conference on Learning Representations*.

Quinonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2008. *Dataset shift in machine learning*. Mit Press.

Ramsundar, B.; Eastman, P.; Walters, P.; and Pande, V. 2019. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. " O'Reilly Media, Inc.".

Rezaei, M. A.; Li, Y.; Wu, D.; Li, X.; and Li, C. 2020. Deep learning in drug design: protein-ligand binding affinity prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1): 407–417.

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

Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Stoeger, T.; Gerlach, M.; Morimoto, R. I.; and Nunes Amaral, L. A. 2018. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS biology*, 16(9): e2006643.

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

Trott, O.; and Olson, A. J. 2010. 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.

Wallach, I.; Dzamba, M.; and Heifets, A. 2015. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*.

Wang, K.; Zhou, R.; Li, Y.; and Li, M. 2021. DeepDTAF: a deep learning method to predict protein–ligand binding affinity. *Briefings in Bioinformatics*, 22(5): bbab072.

Wood, V.; Lock, A.; Harris, M. A.; Rutherford, K.; Bähler, J.; and Oliver, S. G. 2019. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open biology*, 9(2): 180241.

Wu, F.; Wu, L.; Radev, D.; Xu, J.; and Li, S. Z. 2023. Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology*, 6(1): 876.

Xiao, Y.; and Wang, W. Y. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 7322–7329.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Xu, L.; Ru, X.; and Song, R. 2021. Application of machine learning for drug–target interaction prediction. *Frontiers in Genetics*, 12: 680117.

Yang, N.; Zeng, K.; Wu, Q.; Jia, X.; and Yan, J. 2022. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35: 12964–12978.

Yuan, Y.; Chen, Q.; Mao, J.; Li, G.; and Pan, X. 2023. DG-Affinity: predicting antigen–antibody affinity with language models from sequences. *BMC bioinformatics*, 24(1): 430.

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

Appendix

Model Architecture & Hyperparameters

Table 2 depicts TrustAffinity's architecture and configuration in detail.

Table 2: Model Architecture Configuration

Modules	Component	Parameter	Value
Protein Sequence Module	ESMFold	Embedding Dimension	[700, 384]
	ResNet	Layers	5
		Embedding Dimension	704
Ligand Module	GNN	Embedding Dimension	300
		Layers	5
		Jump Knowledge	last
		Dropout	0.4
		Backbone	GIN
Protein Ligand Interaction Module	Attentive Pooling	Dropout	0.4
		Embedding Dimension	1004
	Multi-layer Perceptron	Layers	4
eMOSAIC	Multi-layer Perceptron	Embedding Dimension	50
		Layers	3

Table 3 depicts the hyperparameters involved in the training of TrustAffinity.

Table 3: Model and Training Hyperparameters

Module	Hyperparameter	Value
Binding Affinity Prediction Module	Learning Rate	0.0001
	Batch Size	256
	Epochs	50
	Loss	MSE Loss
	Optimizer	AdamW
eMOSAIC	Learning Rate	0.0001
	Batch Size	32
	Epochs	10
	Loss	MSE
	Optimizer	AdamW