TrustAffinity: accurate, reliable and scalable out-of-distribution protein-ligand binding affinity prediction using trustworthy deep learning

Amitesh Badkul[1], Li Xie[1], Shuo Zhang[1, 3], Lei Xie[1, 2, 3]
[1]Department of Computer Science, Hunter College, The City University of New York
[2] Ph.D. Programs in Computer Science, Biology & Biochemistry, The Graduate Center
The City University of New York
[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, lixie9508@gmail.com
szhang4@gradcenter.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 protein language model, efficient uncertainty quantification based on residue-estimation and novel uncertainty regularized optimization. We extensively validate TrustAffinity in multiple OOD settings. TrustAffinity significantly outperforms state-of-the-art computational methods by a large margin. It achieves a Pearson's correlation between predicted and actual binding affinities above 0.9 with a high confidence 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.

1 Introduction

Drug discovery is very complex process, taking up to 15 years and costing billions of dollars [1]. 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 [2, 3, 4], speculating their potential to accelerate discovering new therapeutics for unmet medical needs [5]. 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

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

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[6, 7]. 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 [8, 9, 10, 11]. More than 95% of human proteins do not have known small molecule ligands [12, 13, 14, 15]. Additionally, the chemical space of small organic molecules is astronomically vast. Although the number of possible small organic molecules is approximate 10^{60} [16], only around 10^6 compounds have annotated protein targets [17, 18, 19]. 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 OOD scenario for understudied proteins and unexplored chemical space[20, 21], in which unseen testing data (proteins or chemicals) are significantly different from training data.

Biophysics-based protein-ligand docking (PLD) may endure the OOD problem when the reliable 3-dimensional (3D) structure of drug target is available [22]. However, PLD suffers from a high rate of false positives due to poor modeling of protein dynamics, solvation effects, crystallized water, and other challenges [23]. The reliability of PLD significantly deteriorates when using predicted protein structures [24, 25, 26]. Despite the success of AlphaFold2 [27], it can only reliably model approximately half of understudied human proteins whose small molecule ligands are unknown [28]. Moreover, PLD is computationally intensive, taking several seconds to score a protein-chemical pair.

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 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 [29][30] 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 [31] 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 [32] for exploring new chemical genomics space, incorporates an uncertainty quantification module inspired by Residual Estimation with an I/O Kernel (RIO) [30], and proposes a new NLPD-based uncertainty score. Under a rigorous benchmark study, our proposed method significantly outperforms state-of-the-art deep learning models 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.

2 Method

Our method, TrustAffinity, consists of two main modules, the binding affinity prediction module and the uncertainty quantification module. They are used together to improve each other's performance.

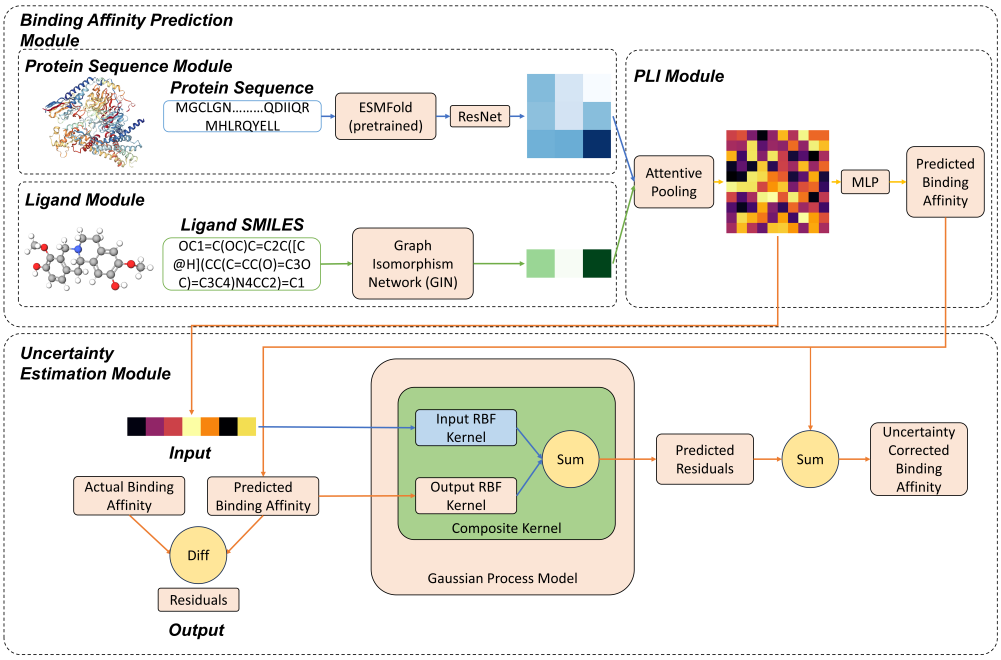

Figure 1: Overview of TrustAffinity. TrustAffinity consists of two main modules, binding affinity prediction module and uncertainty estimation module, which work in sync to provide the binding affinity and the uncertainty associated with the prediction.

The following subsections provide motivations and details for each of these modules in TrustAffinity framework. Figure 1 provides an overview of TrustAffinity.

2.1 Binding Affinity Module

The binding affinity module consists of three sub-modules - protein sequence module, ligand processing module, and lastly protein-ligand interaction (PLI) module. All of these modules work together to predict the binding affinity associated with the PLI.

2.1.1 Protein Sequence Module

Protein sequence representation is one of the most vital components in the machine learning frameworks for predicting not only PLIs [21, 33, 34, 35], but also their 3D structure [27, 32]. Protein sequences contain information that can be used to infer protein structure, function, and family [32], making them a rich source of data for machine learning models [27, 32]. Large datasets of protein sequences are available [27, 36, 37], enabling machine learning frameworks to learn high-level, general representations of proteins. We utilize ESMFold [32] 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 local sequence representations, we use the ResNet model [38] 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.

2.1.2 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) [39]. 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.

2.1.3 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.

2.2 Uncertainty Quantification Module

In our uncertainty quantification module, we integrate a refined RIO framework [30] to enhance uncertainty quantification. Our refined GP uses the proposed I/O kernel from the RIO work. Since our dataset is very large, the time complexity for using the exact GP would be $\mathcal{O}(n^3)$, where n is the total number of training samples. Therefore, we use the approximate GP, which considers a certain number of inducing points learnable through the training process. We randomly take 50 points, reducing the time complexity down to $\mathcal{O}(mn^2)$, where m is the total number of inducing points. Unlike the original RIO framework, our approach involves the adaptation of the simultaneous training of GP and binding affinity prediction. So that we can actively utilize the uncertainty associated with the predictions to make the model aware of its own uncertainties and thereby improve its performance. Our uncertainty quantification module uses the output obtained from the attentive pooling layer concatenated with the predicted residues as input to predict the residuals (the difference between the predicted and the true binding affinity). By learning to model the residuals, which represent the variance associated with the model's predictions, the model aims to become aware of its variability and uncertainty, thereby obtaining a more trustworthy model. We assess uncertainty through the average Natural Logarithm Predictive Density (NLPD), which prefers conservative models while penalizing overly confident and underconfident predictions [40]. A lower NLPD value is better, and it evaluates both the prediction and its associated uncertainty. We implement a novel loss function inspired by [41], that directly uses the average uncertainty in form of NLPD to combine with the mean squared error (MSE) loss. Our NLPD-MSE combined loss function is defined as, where the NLPD_{mean} is the summation of the individual NLPDs:

$$\text{loss} = \frac{1}{2}\left(e^{-\log(\text{NLPD}_\mu)} \cdot (\text{MSE} + \log(\text{NLPD}_\mu)\right). \tag{1}$$

NLPD is defined as follows for a point x_i from a distribution with X values:

$$\text{NLPD}(x_i, \mu, \sigma) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{(x_i - \mu)^2}{2\sigma^2} \tag{2}$$

TrustAffinity utilizes two types of NLPD-based scores. NLPD (Y_{true}) is calculated using Y_{true}, $Y_{corrected}$, and σ as the X, μ, and σ respectively in the Eq. 2. We calculate the corrected/adjusted prediction ($Y_{corrected}$) values using the residue value obtained by the uncertainty quantification module by adding the mean value (μ) of the residue obtained and the predicted neural network prediction (Ynn) as shown in the Eq. 3. We define $NLPD(Y_{true})$ as shown in the Eq. 4. It evaluates how well the TrustAffinity's corrected predictions align with the true values or the ground truth, and provides an upper bound of TrustAffinity performance when evaluated by the NLPD.

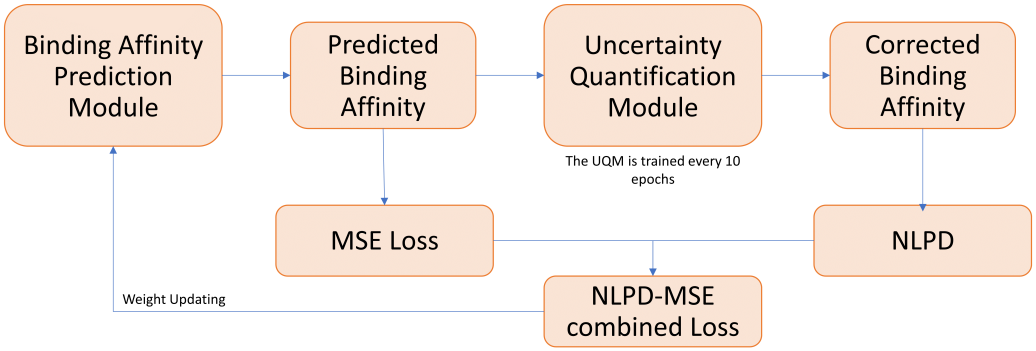

Figure 2: TrustAffinity's training process.

For NLPD (Y_{pred}), we compute it using $Y_{corrected}$, μ, and σ as the X, μ, and σ respectively in the Eq. 2. NLPD (Y_{pred}), as shown in Eq. 5, is a measure of how significant the corrected value is in relation to the TrustAffinity's Binding Affinity prediction model's uncertainty for unseen data. It is useful in practical scenarios for novel prediction uncertainty estimation.

$$Y_{corrected} = Y_{nn} + \mu \tag{3}$$

$$\text{NLPD}(Y_{true}) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{(Y_{true} - Y_{corrected})^2}{2\sigma^2} \tag{4}$$

$$\text{NLPD}(Y_{pred}) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{(Y_{corrected} - \mu)^2}{2\sigma^2} \tag{5}$$

2.3 Training of TrustAffinity

We train our binding affinity module for 50 epochs with a batch size of 256. As shown in Figure 2, we train the uncertainty quantification module or the GP every 10 epochs to ensure it is in sync with the latest binding affinity prediction module, leading to continual improvement in its ability to estimate uncertainty. More details about the hyperparameters and configuration of the TrustAffinity model are present in the appendix.

3 Experiments

3.1 Experimental Settings

Dataset: We train TrustAffinity on the ChEMBL31 database[17], which consists of 350,400 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 two scaffold splitting [42] methods - 1) Random Scaffold Split - random selection of scaffolds, 2) Standardized Scaffold Split - ordered selection of scaffolds, 3) Pfam Split - random selection of Pfam protein families. Scaffold split ensures that there was no overlap of scaffold in the training, testing and validation set. Pfam split ensures that there was no overlap of protein families 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 [33], on the OOD test set. BACPI [33] outperforms other state-of-the-art models DeepAffinity [21], DeepPurpose [43], MONN [44]. Thus, we do not directly compare TrustAffinity with these models. We also compared TrustAffinity with

a typical PLD method, AutoDock Vina [45], on an external Dopamine Receptor antagonist data set [12], 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).

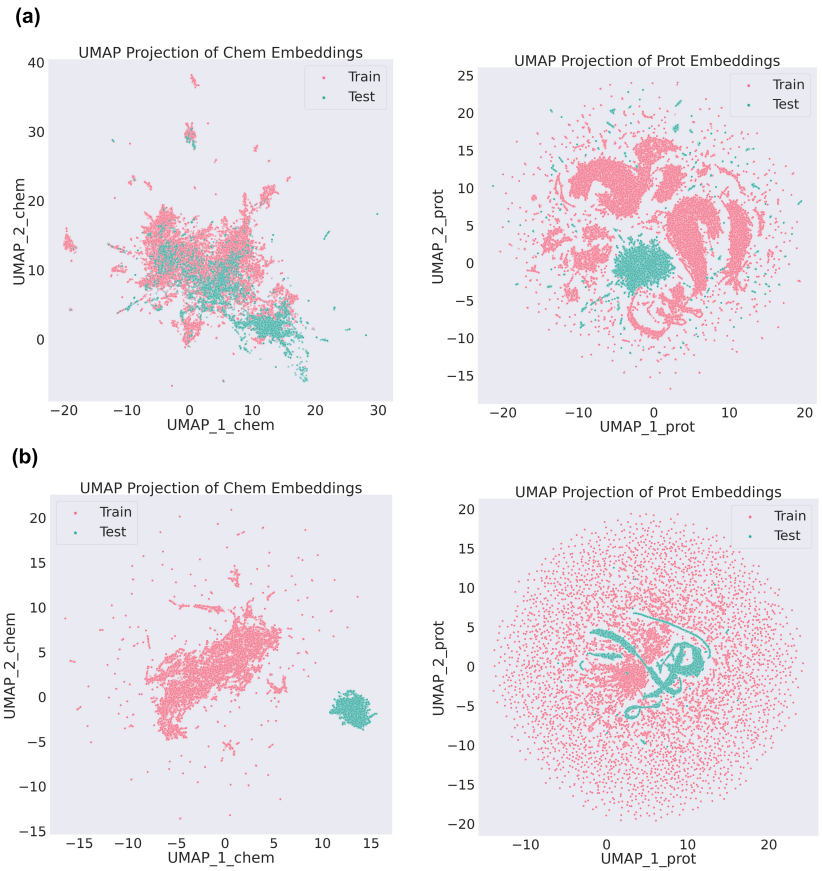

Figure 3: UMAP plot of chemical and protein embeddings of TrustAffinity. (a) UMAP plot for Pfam split, (b) UMAP plot for chemical scaffold split.

4 Results and Discussions

Improved OOD binding affinity prediction: We evaluated the performance of TrustAffinity in several OOD settings. They include standardized and random scaffold splits where chemicals in the testing set have different chemical scaffolds from those in the training/validation set. The difference between two types of splits is that the scaffold classes in testing data is selected following the order sorted by the number of chemicals in each scaffold class in the standardized split, but randomly in the random split. For the purpose of comparison, we also evaluate TrustAffinity in the in-distribution setting of random split. In addition, we assess the generalization power of TrustAffinity in a Pfam split benchmark where proteins in the testing data belong to different Pfam families from those in the training and validation data, a challenging OOD setting. Fig. 3, depicts the Uniform Manifold Approximation and Projection (UMAP) plot of protein and chemical embeddings for both the Pfam split and the scaffold split, where training and testing data form distinctly different clusters for protein embeddings and chemical embeddings, respectively. They clearly demonstrate OOD scenarios.

In the OOD setting, TrustAffinity consistently outperforms the current state-of-the-art BACPI model regarding *all* four metrics in *all* settings, as shown in Figure 4. In both the standardized scaffold

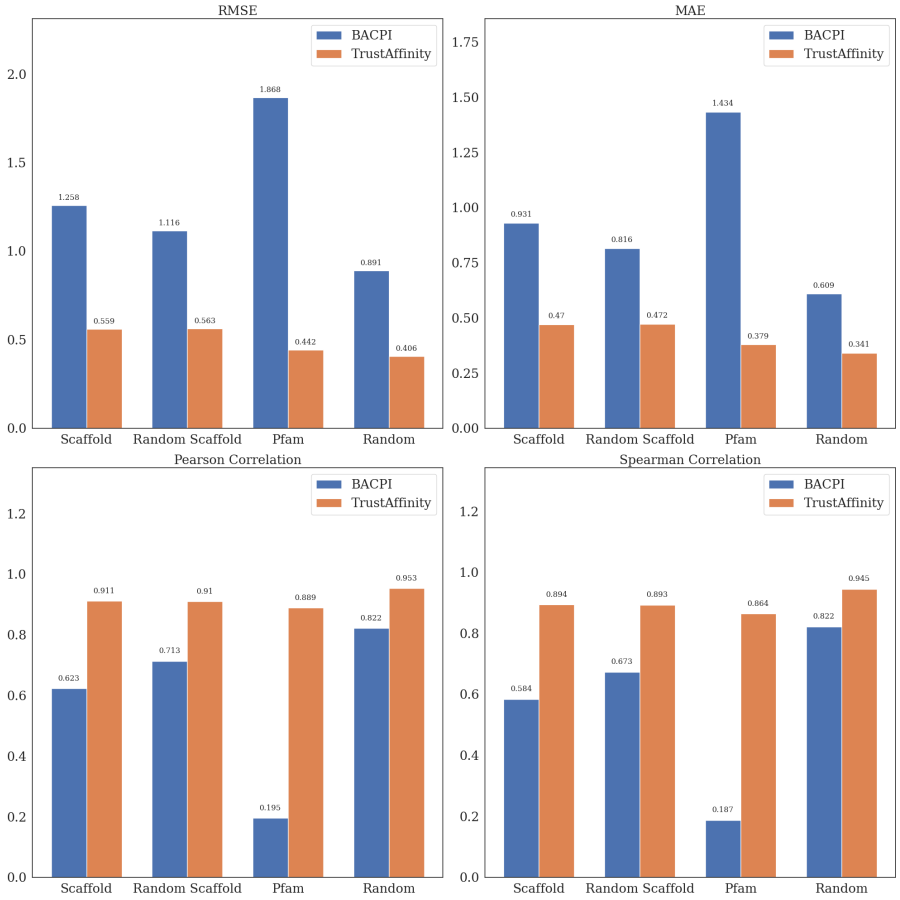

Figure 4: Performance comparison of TrustAffinity with BACPI on test set for various settings including standard scaffold split, random scaffold split, Pfam OOD split, and random split.

split and the random scaffold split settings, our model could successfully utilize 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. Moreover, in case of Pfam based split, we see that BACPI generalizes even worse as compared to scaffold based splits. But TrustAffinity is still able to obtain consistent performance similar to the other splits across all the metrics. 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 61.62% lower), and MAE (on average 56.15% lower) when compared to BACPI.

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 (Y_{true})	**0.384**	**0.312**	**0.856**	**0.820**
TrustAffinity (Y_{pred})	0.846	0.65	0.612	0.667

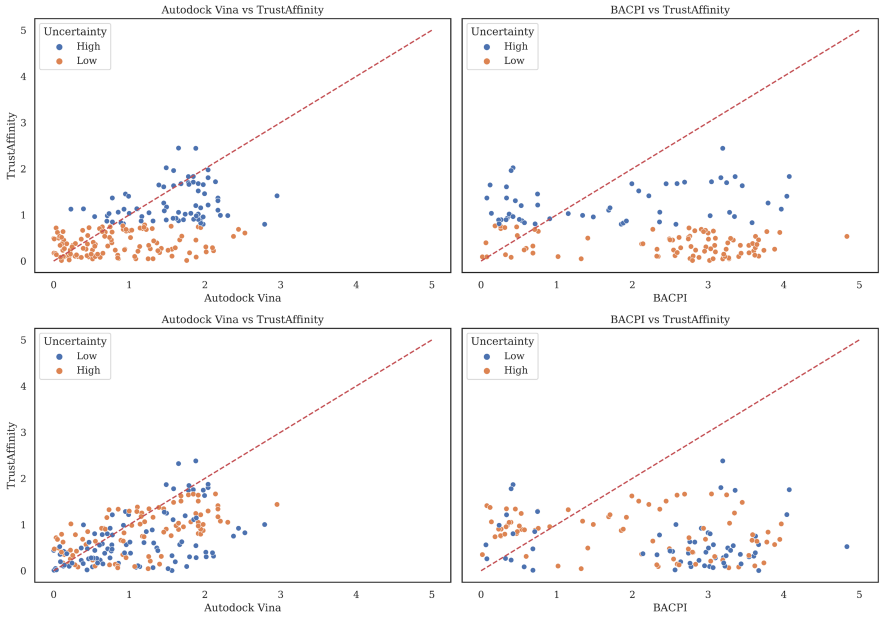

Figure 5: Comparative residual plot of TrustAffinity vs AutoDock Vina (PLD), and TrustAffinity vs BACPI. (a) Uses Y_{true} as part of uncertainty estimation, (b) Uses Y_{pred} as part of uncertainty estimation

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 5 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 5),which is particularly true for the low uncertainty predictions. 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 [46], 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.

5 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 demonstrated the robust OOD generalization capabilities of TrustAffinity, yielding reliable 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. However, our method has certain limitations. Firstly, there is a significant performance margin between the proposed NLPD score for the uncertainty quantification and its theoretical upper-bound. New methods are needed to push the limit of the uncertainty quantification. Secondly, it is unable to predict the binding pose of the PLI, which is also crucial for further drug discovery pipeline. Finally, 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

[1] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[2] Debleena Paul, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, and Rakesh K Tekade. Artificial intelligence in drug discovery and development. *Drug discovery today*, 26(1):80, 2021.

[3] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Henry Chao, Li Chen, Paul A Craig, Gregg V Crichlow, Kenneth Dalenberg, Jose M Duarte, et al. 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, 2023.

[4] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

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

[6] Lei Xu, Xiaoqing Ru, and Rong Song. Application of machine learning for drug–target interaction prediction. *Frontiers in Genetics*, 12:680117, 2021.

[7] Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska, and Kayvan Najarian. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in bioinformatics*, 22(1):247–269, 2021.

[8] Winston A Haynes, Aurelie Tomczak, and Purvesh Khatri. Gene annotation bias impedes biomedical research. *Scientific reports*, 8(1):1362, 2018.

[9] Valerie Wood, Antonia Lock, Midori A Harris, Kim Rutherford, Jürg Bähler, and Stephen G Oliver. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open biology*, 9(2):180241, 2019.

[10] Thomas Stoeger, Martin Gerlach, Richard I Morimoto, and Luís A Nunes Amaral. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS biology*, 16(9):e2006643, 2018.

[11] Tudor I Oprea, Cristian G Bologa, Søren Brunak, Allen Campbell, Gregory N Gan, Anna Gaulton, Shawn M Gomez, Rajarshi Guha, Anne Hersey, Jayme Holmes, et al. Unexplored therapeutic opportunities in the human genome. *Nature reviews Drug discovery*, 17(5):317–332, 2018.

[12] Tian Cai, Li Xie, Shuo Zhang, Muge Chen, Di He, Amitesh Badkul, Yang Liu, Hari Krishna Namballa, Michael Dorogan, Wayne W Harding, et al. End-to-end sequence-structure-function meta-learning predicts genome-wide chemical-protein interactions for dark proteins. *PLOS Computational Biology*, 19(1):e1010851, 2023.

[13] John A Gerlt, Karen N Allen, Steven C Almo, Richard N Armstrong, Patricia C Babbitt, John E Cronan, Debra Dunaway-Mariano, Heidi J Imker, Matthew P Jacobson, Wladek Minor, et al. The enzyme function initiative. *Biochemistry*, 50(46):9950–9962, 2011.

[14] Denise Carvalho-Silva, Andrea Pierleoni, Miguel Pignatelli, ChuangKee Ong, Luca Fumis, Nikiforos Karamanis, Miguel Carmona, Adam Faulconbridge, Andrew Hercules, Elaine McAuley, et al. Open targets platform: new developments and updates two years on. *Nucleic acids research*, 47(D1):D1056–D1065, 2019.

[15] Georg Kustatscher, Tom Collins, Anne-Claude Gingras, Tiannan Guo, Henning Hermjakob, Trey Ideker, Kathryn S Lilley, Emma Lundberg, Edward M Marcotte, Markus Ralser, et al. Understudied proteins: opportunities and challenges for functional proteomics. *Nature Methods*, 19(7):774–779, 2022.

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

[17] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

[18] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, 2023.

[19] Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1):D1045–D1053, 2016.

[20] Thin Nguyen, Hang Le, Thuc Le, and Svetha Venkatesh. Prediction of drug–target binding affinity using graph neural networks. *BioRxiv*, page 684662, 2019.

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

[22] Lei Xie, Xiaoxia Ge, Hepan Tan, Li Xie, Yinliang Zhang, Thomas Hart, Xiaowei Yang, and Philip E Bourne. Towards structural systems pharmacology to study complex diseases and personalized medicine. *PLoS computational biology*, 10(5):e1003554, 2014.

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

[24] Mariama Jaiteh, Ismael Rodríguez-Espigares, Jana Selent, and Jens Carlsson. Performance of virtual screening against gpcr homology models: Impact of template selection and treatment of binding site plasticity. *PLoS computational biology*, 16(3):e1007680, 2020.

[25] Tian Cai, Li Xie, Muge Chen, Yang Liu, Di He, Shuo Zhang, Cameron Mura, Philip E Bourne, and Lei Xie. Exploration of dark chemical genomics space via portal learning: applied to targeting the undruggable genome and covid-19 anti-infective polypharmacology. *Research Square*, 2021.

[26] Valeria Scardino, Juan I Di Filippo, and Claudio N Cavasotto. How good are alphafold models for docking-based virtual screening? *Iscience*, 26(1), 2023.

[27] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[28] Carrie Arnold. Alphafold touted as next big thing for drug discovery-but is it? *Nature*, 2023.

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

[30] Xin Qiu, Elliot Meyerson, and Risto Miikkulainen. Quantifying point-prediction uncertainty in neural networks via residual estimation with an i/o kernel. In *International Conference on Learning Representations*, 2019.

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

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

[33] Min Li, Zhangli Lu, Yifan Wu, and YaoHang Li. Bacpi: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. *Bioinformatics*, 38(7):1995–2002, 2022.

[34] Kaili Wang, Renyi Zhou, Yaohang Li, and Min Li. Deepdtaf: a deep learning method to predict protein–ligand binding affinity. *Briefings in Bioinformatics*, 22(5):bbab072, 2021.

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

[36] Alex L Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R Crusoe, Varsha Kale, Simon C Potter, Lorna J Richardson, et al. Mgnify: the microbiome analysis resource in 2020. *Nucleic acids research*, 48(D1):D570–D578, 2020.

[37] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[39] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[40] Joaquin Quinonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop*, pages 1–27. Springer, 2005.

[41] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[42] Bharath Ramsundar, Peter Eastman, Pat Walters, and Vijay Pande. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. " O'Reilly Media, Inc.", 2019.

[43] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23):5545–5547, 2020.

[44] Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems*, 10(4):308–322, 2020.

[45] Oleg Trott and Arthur J Olson. 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.

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

[47] Andrew R Leach, Brian K Shoichet, and Catherine E Peishoff. Prediction of protein- ligand interactions. docking and scoring: successes and gaps. *Journal of medicinal chemistry*, 49(20):5851–5855, 2006.

[48] José Jiménez, Miha Skalic, Gerard Martinez-Rosell, and Gianni De Fabritiis. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling*, 58(2):287–296, 2018.

[49] Mohammad A Rezaei, Yanjun Li, Dapeng Wu, Xiaolin Li, and Chenglong Li. Deep learning in drug design: protein-ligand binding affinity prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):407–417, 2020.

[50] Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*, 2015.

[51] Evan N Feinberg, Debnil Sur, Zhenqin Wu, Brooke E Husic, Huanghao Mai, Yang Li, Saisai Sun, Jianyi Yang, Bharath Ramsundar, and Vijay S Pande. Potentialnet for molecular property prediction. *ACS central science*, 4(11):1520–1530, 2018.

[52] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[53] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[54] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[55] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[56] Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35:12964–12978, 2022.

[57] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.

[58] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE, 2016.

[59] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[60] Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7322–7329, 2019.

[61] Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*, 2023.

[62] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

A Appendix

A.1 Related Works

A.1.1 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 [47]. Furthermore, PLD is highly sensitive to the conformational state of structures, which can often lead to inaccurate predictions [23].

A.1.2 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 PLIs and their binding affinity from 3D information [48, 49, 50, 51]. These structure-based machine learning methods use a variety of methods to incorporate 3D information into their models for predicting binding affinity. K_{DEEP} [48] and DeepAtom [49] 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 [50] also uses a form of voxel representation. Instead of representing the entire protein, it represents only the binding site of the target protein. PotentialNet [51] 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) [21, 33, 34, 35]. DeepDTA [35] performs training on label encoded SMILES and protein sequences after using 1D CNNs to obtain the representations. DeepAffinity [21] 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 [33] 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 [34] 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.

A.1.3 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 [52]. Existing methods in machine learning commonly assumed that both training and testing data are independent and identically distributed (iid) [53]. 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 [12, 54, 55, 56, 57]. PortalCG [12] 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.

A.1.4 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 [41, 58, 59] and natural language processing [60, 61, 62]. However, the application of uncertainty quantification to the field of biology is relatively limited. Zeng and Gilford [31] 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. [29] 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. [30] and notice similar or better results when compared to the GP.

A.2 Model Architecture & Hyperparameters

Table A.1, and Table A.2 depicts the model architecture, configuration, and hyperparameters.

Table A.1: Model and training hyperparameters

Module	Hyperparameter	Value
Binding Affinity Prediction Module	Learning Rate	0.0001
	Batch Size	256
	Epochs	50
	Loss	NLPD-MSE Loss
	Optimizer	Adam
Uncertainty Quantification Module	Learning Rate	0.0001
	Batch Size	32
	Epochs	50
	Loss	Exact Marginal Log Likelihood
	Optimizer	'Adam

Table A.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
Uncertainty Quantification Module	Gaussian Process Regression	Kernels	Radial basis function
		Likelihood	Gaussian Likelihood
		Mean	Linear