Using Single Protein/Ligand Binding Models to Predict Active Ligands for Previously Unseen Proteins

Vikram Sundar
Department of Chemistry
University of Cambridge
Cambridge, UK
vs456@cam.ac.uk

Lucy Colwell
Department of Chemistry
University of Cambridge
Cambridge, UK
ljc37@cam.ac.uk

Abstract

Machine learning (ML) models trained to predict small molecule ligand binding to single proteins have achieved remarkable success, but cannot make predictions about protein targets other than the one they are trained on. Models that make predictions for multiple proteins and multiple ligands, known as drug-target interaction (DTI) models, aim to solve this problem but generally have lower performance. In this work, we improve the performance of DTI models by taking advantage of the accuracy of single protein/ligand binding models. Specifically, we first construct individual protein/ligand binding models for all train proteins with some experimental data, and then use each individual model to make predictions for all remaining ligands, against the corresponding protein target. Finally, we use the known and predicted ligand binding data for all targets in a DTI model to make predictions for the unseen test ligands and proteins. This approach significantly improves performance; most importantly, some of our models are able to achieve Areas Under the Receiver Operator Characteristic curve (AUCs) exceeding 0.9 on test datasets that contain only unseen proteins and unseen ligands.

1 Introduction

Identifying ligands that bind tightly to a given protein target is a crucial first step in drug discovery. Experimental methods such as high-throughput screening are time-consuming, and costly, while physics-based methods are computationally expensive and can be inaccurate [19, 32, 12, 2]. The emergence of large datasets enables data-driven approaches to be applied to this problem. In recent years, a variety of ML-based approaches have been developed to identify active ligands for a single protein target given training data from screening experiments [10, 3, 30]. These approaches report outstanding *in silico* success on benchmark datasets [34, 27, 5, 35, 14, 26, 28, 11]. However, they rely on the existence of experimental screening data that identifies active and inactive ligands for each protein target, which are costly and time-consuming to obtain.

Models that predict global drug-target interactions (DTI) aim to remove this bottleneck by predicting the interactions between multiple protein targets and multiple candidate ligands or drugs [1, 29, 25, 7]. Consider the matrix of interactions between a set of small molecule ligands, and a set of protein targets illustrated in Figure 1. DTI models use experimental data for some subset of interactions together with computational descriptors of the protein targets and the small molecule ligands to predict all entries of this matrix. The ultimate goal is to build models that accurately predict interactions in which no experimental data is available for either the protein target or the small molecule ligand [7, 1].

ML-based approaches for DTI prediction can be broadly classified into similarity-based methods, which build similarity matrices for the different proteins and ligands [7, 1] and feature-based methods, which use standard machine learning algorithms such as decision trees or neural networks on a given

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

feature set [7]. Recently, many deep-learning-based methods have become popular [36, 17, 16, 18, 20, 37, 38, 13, 33, 23, 8]. All methods predict the entire DTI matrix in one step, using all of the available information at the same time. However, performance analyses using standard benchmark datasets demonstrate that these methods often struggle to make accurate prediction for interactions involving proteins or ligands that are not present in the training data [7]. The difficulty of generalising to previously unseen ligands is particularly surprising, since single protein models successfully solve this problem without including data for related proteins [3]; this is likely because the DTI models have insufficient data and are not expressive enough to learn the true binding function between multiple proteins and multiple ligands. The most difficult and most interesting test case of generalising in both protein and ligand space simultaneously has rarely been tried in the literature [7]; when it is tested, models often perform poorly [7, 36, 33, 8] though in some cases this is partially accounted for by the difficulty of the dataset [33].

In this paper, we introduce a novel approach that uses single protein/ligand binding models to improve the accuracy of DTI models. Specifically, we use a two-step procedure: first we build single protein/ligand binding models and use them to generalise to unseen ligands for each known protein. Then we use this data to build DTI models using wide range of existing methods, and use these DTI models to generalise to unseen proteins. We show that this two-stage algorithm results in improvements in the performance of a wide variety of DTI models applied to a number of different datasets. Most importantly, our method makes it possible to tackle the most challenging case involving simultaneous generalisation to both unseen drugs and unseen targets with reasonable accuracy.

2 Methods

2.1 Dataset Construction and Problem Set-up

The datasets analyzed update the gold-standard datasets developed by Yamanishi et al. [39] to reflect the vast quantity of protein/ligand binding data that is now available. For these protein targets we found active ligands from ChEMBL 24.1 [4, 9] by filtering for compounds with an IC_{50}, K_i, K_d, or EC_{50} of less than $1\,\mu M$; to prevent duplication of proteins we only used proteins from *Homo sapiens*. Targets with fewer than 20 active ligands in ChEMBL were eliminated. We found between 22 and 3182 active ligands for 91 GPCR targets, and between 26 and 1864 active ligands for 21 nuclear receptor (NR) targets.

Inactive ligands were acquired from PubChem indexed by UniProt Protein ID [21, 15] and for targets with a DUD-E decoy set, randomly sampled inactives from the DUD-E set were included [22]. To ensure a reasonable balance of actives to inactives, we randomly selected 500 decoy ligands per run from ChEMBL [9] that were not in any previous set of ligands, either active or inactive. Unlike previous work [7], we did not assume when training or testing our models that unknown interactions between proteins and non-decoy ligands were inactive.

We next split our dataset into training and test sets. To explicitly test for generalisation in both protein and ligand space, we used the incomplete training submatrix of protein/ligand interactions illustrated in Figure 1a. Specifically, we first split the protein targets randomly into 80% train and 20% test. For each protein target, we randomly selected 20% of the experimentally validated ligands (active and inactive) for the test set. We further included 20% of decoys in the test set. All remaining ligands were placed in the training set, ensuring that all protein targets in the training set have some active and inactive ligands in the training set. The models were not provided with any information about interactions involving any of the test ligands or proteins. This allows us to evaluate performance on the subproblem of greatest interest in which both the protein and ligand are not seen in the training data.

Some models required hyperparameter tuning, so we created a validation submatrix within the training submatrix using the same methodology as for the train/test split. Hyperparameters were tuned separately for every repetition using the AUC on the validation set. We performed 20 replicates for all models; error bars reported are to 1 SEM.

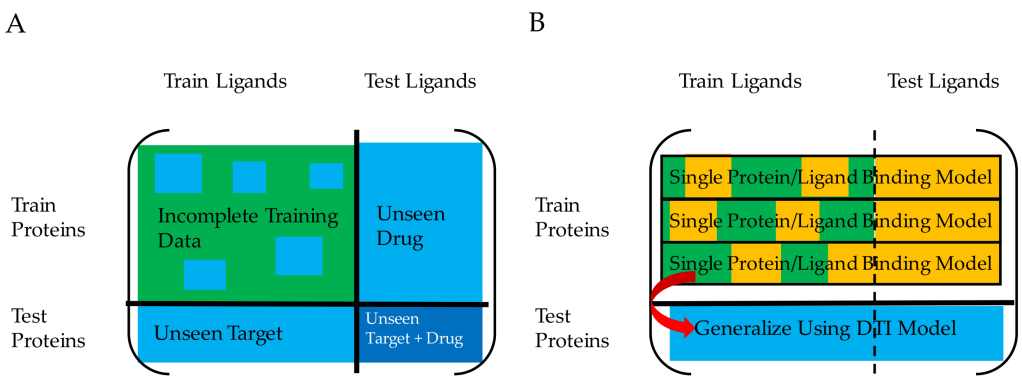

Figure 1: (a) Our split into training and test sets for the DTI problem. This paper focuses on the cases in which neither the protein target nor the small molecule ligand were seen during training, which is most difficult as it requires generalisation in both protein and ligand space. (b) A depiction of our proposed method, wherein we use single protein/ligand binding models to predict the interactions between known proteins and all ligands. We then use a standard DTI model to generalise to the test proteins.

2.2 Models

Our proposal is to add an initial step for each training protein that builds a single protein/ligand binding model with the available data and uses these models to predict the interactions between the corresponding protein and all other ligands in the dataset (see Figure 1b) . Here we use logistic regression with 2048 bit ECFP6 fingerprints implemented using scikit-learn to build the initial target specific models. [31, 24] We use regularization constant $C = 1$, which is known to perform well on the single protein/ligand binding problem. We then use a standard DTI model to generalise to the test proteins using the predicted and experimentally validated ligands for each training protein. We tested weighted nearest neighbor (w-NN) as a baseline DTI model, and four high-performing DTI models from the literature, namely random forest with one-hot features (RF one-hot), regularized least squares (RLS-WNN), collaborative matrix factorization (CMF), and weighted graph-regularized matrix factorization (WGRMF). [7] In all cases, we used normalized pairwise E-value scores from HMMER to describe similarity between protein targets, [6] and Tanimoto similarities calculated using ECFP6 fingerprints for the ligands.

3 Results

3.1 Performance

Figure 2 presents the AUCs achieved by the five DTI algorithms evaluated using test sets in which neither the protein targets nor the small molecule ligands were seen during training. Similar results are obtained if a random forest is used for the single protein ligand binding model in place of logistic regression (data not shown), and also for further datasets containing ion channels, kinase or other enzyme protein targets (data not shown). Across all results in Figure 2, incorporation of the single protein ligand binding models results in improvement that is statistically significant at a $p < 0.01$ level. Further analysis indicates that our methods achieve similar performance even for targets in the test dataset only distantly related to those in the train dataset (data not shown). Our results suggest that adding an initial step in which single protein ligand binding models are constructed has the potential to unblock a crucial bottleneck in the ability for models build using existing experimental data to generalise to new protein targets.

3.2 Proposed Explanation

The key advantage provided by our approach is that the DTI models are built using significantly more data than is available from experiments for each protein target. Although some of this additional data

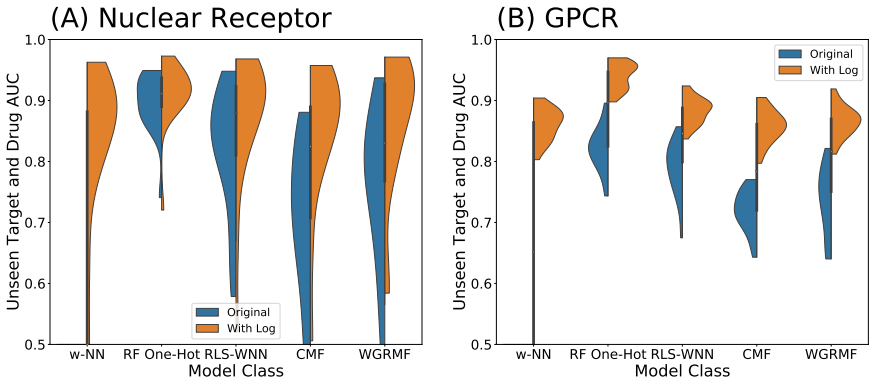

Figure 2: Logistic Regression single protein-ligand binding models improve the ability of DTI models to make accurate predictions for held-out test data in which neither the protein target or small molecule ligand were seen during training. (A) Results using five different DTI models for 21 Nuclear Receptor protein targets and (b) 91 GPCR protein targets. The baseline DTI model is weighted nearest neighbor (w-NN), while the other DTI models are random forest with one-hot features (RF one-hot), regularized least squares (RLS-WNN), collaborative matrix factorization (CMF), and weighted graph-regularized matrix factorization (WGRMF). We see consistent improvement in all models regardless of dataset size or sparsity and type of DTI model used.

consists of predicted interactions, including these predictions adds useful information and results in improved DTI models because the single protein-ligand binding models are highly accurate. To further examine the information that is added by our approach for the GPCR dataset, Figure 3 shows the interaction probabilities predicted by (A) the logistic regression and (B) the random forest single protein ligand binding models, aggregated across the 73 GPCR protein targets included in the training dataset. While many interactions are predicted to be inactive, a small but significant proportion of the interactions are predicted to be active; in contrast, standard approaches either ignore this data or assume that any unknown interactions are inactive [7].

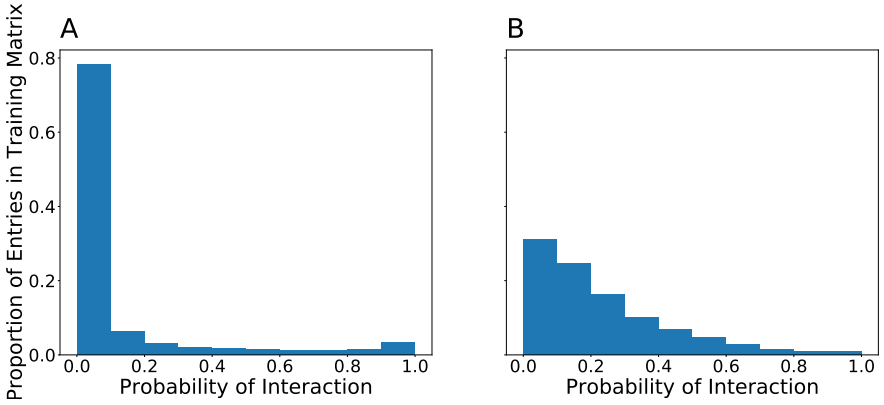

Figure 3: Histogram showing the predicted interaction probabilities added to the training matrix by single protein binding models built using (A) Logistic Regression and (B) Random Forests for the 91 GPCR protein targets in the GPCR dataset.

4 Conclusions

In this work, we demonstrate that incorporating predictions made by robust single protein/ligand binding models into DTI algorithms allows generalisation simultaneously in both protein and ligand space to predict interactions between unseen drugs and unseen targets. We observe this effect consistently regardless of dataset size across multiple DTI algorithms. Our best-performing DTI

method was random forest with the one-hot feature set of amino acid counts and ECFP6 ligand fingerprints. Since this is a very simple model, we suspect that more complicated models like deep neural networks will be able to learn more information about the dataset and perform better. Our work also suggests feature-based methods perform better than similarity-based methods and should be preferred for future research into DTI models.

References

[1] Ruolan Chen, Xiangrong Liu, Shuting Jin, Jiawei Lin, and Juan Liu. Machine learning for drug-target interaction prediction. *Molecules*, 23(9):1–15, 2018.

[2] Yu-Chian Chen. Beware of docking! *Trends in Pharmacological Sciences*, 36(2):78–95, feb 2015.

[3] Lucy J. Colwell. Statistical and machine learning approaches to predicting protein–ligand interactions. *Current Opinion in Structural Biology*, 49:123–128, 2018.

[4] Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P. Overington. ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, 43(W1):W612–W620, 2015.

[5] David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P. Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems*, 28:2224–2232, 2015.

[6] Sean Eddy. HMMER 3.1b2.

[7] Ali Ezzat, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics*, pages 1–21, 2018.

[8] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable Drug Target Prediction Using Deep Neural Representation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3371–3377, California, jul 2018. International Joint Conferences on Artificial Intelligence Organization.

[9] Anna Gaulton, Anne Hersey, Micha L. Nowotka, A. Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrian-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, Maŕia Paula Magarinos, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, 2017.

[10] Erik Gawehn, Jan A. Hiss, and Gisbert Schneider. Deep Learning in Drug Discovery. *Molecular Informatics*, 35(1):3–14, jan 2016.

[11] Joseph Gomes, Bharath Ramsundar, Evan N. Feinberg, and Vijay S. Pande. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *arXiv Preprint*, (arXiv:1703.10603), mar 2017.

[12] Sam Grinter and Xiaoqin Zou. Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design. *Molecules*, 19(7):10150–10176, jul 2014.

[13] Pengwei Hu, Yu-an Huang, Zhuhong You, Shaochun Li, Keith C. C. Chan, Henry Leung, and Lun Hu. Learning from Deep Representations of Multiple Networks for Predicting Drug–Target Interactions. pages 151–161. Springer, Cham, aug 2019.

[14] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, aug 2016.

[15] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. PubChem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016.

[16] Hanbi Lee and Wankyu Kim. Comparison of Target Features for Predicting Drug-Target Interactions by Deep Neural Network Based on Large-Scale Drug-Induced Transcriptome Data. *Pharmaceutics*, 11(8):377, aug 2019.

[17] Ingoo Lee, Jongsoo Keum, and Hojung Nam. DeepConv-DTI: Prediction of Drug-Target Interactions via Deep Learning with Convolution on Protein Sequences. 2018.

[18] Munhwan Lee, Hyeyeon Kim, Hyunwhan Joe, and Hong-Gee Kim. Multi-channel PINN: investigating scalable and transferable neural networks for drug discovery. *Journal of Cheminformatics*, 11(1):46, dec 2019.

[19] Andrew B MacConnell, Alexander K Price, and Brian M Paegel. An Integrated Microfluidic Processor for DNA-Encoded Combinatorial Library Functional Screening. *ACS combinatorial science*, 19(3):181–192, 2017.

[20] Yosef Masoudi-Sobhanzadeh, Yadollah Omidi, Massoud Amanlou, and Ali Masoudi-Nejad. Trader as a new optimization algorithm predicts drug-target interactions efficiently. *Scientific Reports*, 9(1):9348, dec 2019.

[21] Lewis H. Mervin, Krishna C. Bulusu, Leen Kalash, Avid M. Afzal, Fredrik Svensson, Mike A. Firth, Ian Barrett, Ola Engkvist, and Andreas Bender. Orthologue chemical space and its influence on target prediction. *Bioinformatics*, 34(1):72–79, 2018.

[22] Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, jul 2012.

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

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[25] Tianyi Qiu, Jingxuan Qiu, Jun Feng, Dingfeng Wu, Yiyan Yang, Kailin Tang, Zhiwei Cao, and Ruixin Zhu. The recent progress in proteochemometric modelling: Focusing on target descriptors, cross-term descriptors and application scope. *Briefings in Bioinformatics*, 18(1):125–136, 2017.

[26] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 57(4):942–957, apr 2017.

[27] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively Multitask Networks for Drug Discovery. *arXiv Preprint*, (arXiv:1502.02072), 2015.

[28] Bharath Ramsundar, Bowen Liu, Zhenqin Wu, Andreas Verras, Matthew Tudor, Robert P. Sheridan, and Vijay Pande. Is Multitask Deep Learning Practical for Pharma? *Journal of Chemical Information and Modeling*, 57(8):2068–2076, aug 2017.

[29] Ahmet Sureyya Rifaioglu, Heval Atas, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Dogan. Recent Applications of Deep Learning and Machine Intelligence on In Silico Drug Discovery: Methods, Tools, and Databases. *Briefings in Bioinformatics*, pages 1–35, 2018.

[30] Peter Ripphausen, Britta Nisius, and Jürgen Bajorath. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today*, 16(9-10):372–376, may 2011.

[31] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, may 2010.

[32] Gisbert Schneider. Automating drug discovery. *Nature Reviews Drug Discovery*, 17(2):97–113, dec 2017.

[33] Wen Torng and Russ B. Altman. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *bioRxiv*, page 473074, nov 2018.

[34] Thomas Unterthiner, Andreas Mayr, Günter Klambauer, Marvin Steljaert, Jorg Wenger, Hugo Ceulemans, and Sepp Hochreiter. Deep Learning as an Opportunity in Virtual Screening. *Proceedings of the Deep Learning Workshop at NIPS*, pages 1–9, 2014.

[35] 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, oct 2015.

[36] Fangping Wan and Jianyang (Michael) Zeng. Deep learning with feature embedding for compound-protein interaction prediction. *bioRxiv*, page 086033, nov 2016.

[37] Lei Wang, Zhu-Hong You, Xing Chen, Shi-Xiong Xia, Feng Liu, Xin Yan, Yong Zhou, and Ke-Jian Song. A Computational-Based Method for Predicting Drug–Target Interactions by Using Stacked Autoencoder Deep Neural Network. *Journal of Computational Biology*, 25(3):361–373, mar 2018.

[38] Lingwei Xie, Song He, Xinyu Song, Xiaochen Bo, and Zhongnan Zhang. Deep learning-based transcriptome data classification for drug-target interaction prediction. *BMC Genomics*, 19(S7):667, sep 2018.

[39] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):232–240, 2008.