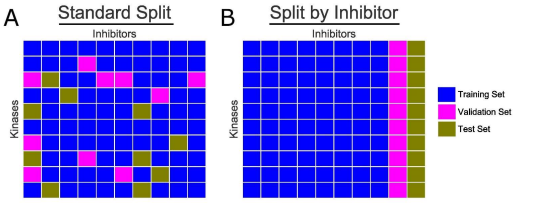
Exploring Poor Generalizations of Kinase-Drug Affinity Predictions in Existing Neural Network Models

Wern Juin Gabriel Ong[1,2], Palani Kirubakaran[1], John Karanicolas[1] | Karanicolas Lab -- Fox Chase Cancer Center

Fox Chase Cancer Center[1] and Bowdoin College[2]

Abstract

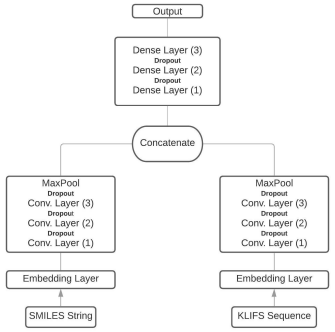
Neural networks have been widely used to predict drug-target interactions (DTIs) since the resurgence of neural network use in the 2000s. A model that can accurately make DTI predictions has the potential to screen large chemical libraries and significantly streamline the drug discovery process. However, despite previous studies reporting models that accurately predict DTI with protein sequences and SMILES strings, it is still unclear how these models would perform in a prospective evaluation due to the possibility of information leakage. In our study, we first design a Convolutional Neural Network-based model comparable to those previously reported when trained on randomly split cross validation data. We then evaluated the model on datasets processed in two additional ways: one where a set of ligands were withheld from the training set and another where the SMILES strings were replaced by a random string of characters. While the model struggles to make accurate predictions on the dataset with ligands withheld (Pearson's Correlation of 0.433), the model trained on random strings (0.825) performed almost identically to the model trained on SMILES strings (0.829). We demonstrate that these trends hold across the two kinase datasets we used for testing. This demonstrates the prevalence of information leakage in existing models and suggests the use of richer feature extraction for training CNN-based models that would generalize well and make accurate predictions in prospective experiments.



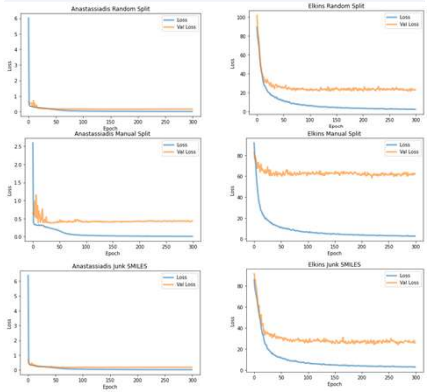
Methodology

- Evaluated three different types of dataset processing on both the Anastassiadis and Elkins dataset.
  - Anastassiadis transformed to lg(IC50), Elkins used as is
- Datasets are split as follows:
  - Training and test set split randomly
  - Selected ligands withheld from training set and randomly split into test and independent validation set
  - SMILES String replaced by a random string of characters

Model Training



Parameters	Range
CNN	
Number of filters	32; 64; 96
Filter length (SMILES)	4; 6; 8
Filter length (KLIFS)	4; 8; 12
Dense Network	
Hidden neurons	1024; 1024, 512
Training Settings	
Epochs	300
Batch size	256
Dropout	0.1
Optimizer	Adam
Learning rate	0.001



Results

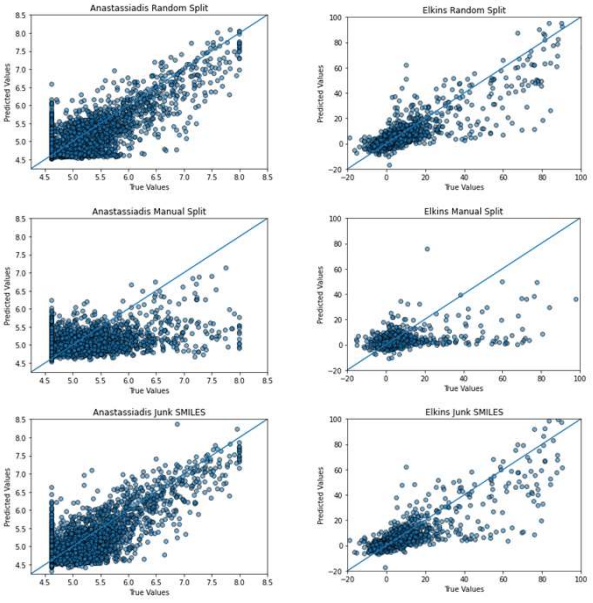
Existing Models

Model	Dataset	Protein	Ligand	CI (Std. Dev.)	MSE
Pahikkala *et. al.* 2014	Davis	*S-W	ᵗPubchem	0.782 (0.0008)	0.379
He *et. al.* 2017	Davis	S-W	Pubchem	0.872 (0.002)	0.282
Öztürk *et. al.* 2018	Davis	CNN	CNN	0.878 (0.004)	0.261
Pahikkala *et. al.* 2014	KIBA	S-W	Pubchem	0.782 (0.0009)	0.441
He *et. al.* 2017	KIBA	S-W	Pubchem	0.836 (0.001)	0.222
Öztürk *et. al.* 2018	KIBA	CNN	CNN	0.863 (0.002)	0.194

Experimental Results

Splitting Type	Dataset	Protein	Ligand	Pearson's (*P*-val.)	MSE
Random	Anastassiadis	CNN	CNN	0.802 (0.0)	0.161
Manual	Anastassiadis	CNN	CNN	0.408 (3.71E-145)	0.297
Random SMILES	Anastassiadis	CNN	CNN	0.801 (0.0)	0.162
Random	Elkins	CNN	CNN	0.855 (0.0)	22.9
Manual	Elkins	CNN	CNN	0.457 (1.16E-180)	48.7
Manual SMILES	Elkins	CNN	CNN	0.849 (0.0)	23.7

* Proteins are parametrized by the Smith-Waterman algorithm
ᵗ Ligands are parametrized by the Pubchem Sim similarity scores



Conclusions

- Our randomly split models are split similarly to existing models and perform well due to information leakage in the training set.
- Our manually split models perform significantly worse indicating the presence of information leakage when compared to the randomly split model.
- The fact that the model with random strings of characters in place of SMILES strings indicates that the model learns poorly from SMILES strings.
- Our experiments, overall, indicate a need for extraction of richer features from a molecule when using as a training input – that 3D representations are needed to meaningfully represent molecules

Acknowledgements

Many thanks to Kiruba for his mentorship during this process, John for welcoming me into this lab, and the group.

References

Anastassiadis, Theonie, et al. "Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity." *Nature Biotechnology*, vol. 29, 30 Oct. 2011, pp. 1039- 1045, doi:.org/10.1038/nbt.2017.

Elkins, Jonathan M., et al. "Comprehensive characterization of the Published Kinase Inhibitor Set." *Nature Biotechnology*, vol. 34, 1 Jan. 2016, pp. 95-103, doi:.org/10.1038/nbt.3374.

Öztürk, Hakime, et al. "DeepDTA: deep drug-target binding affinity prediction." *Bioinformatics*, vol. 34, no. 17, 8 Sept. 2018, pp. i821-i829, doi:.org/10.1093/bioinformatics/bty593.