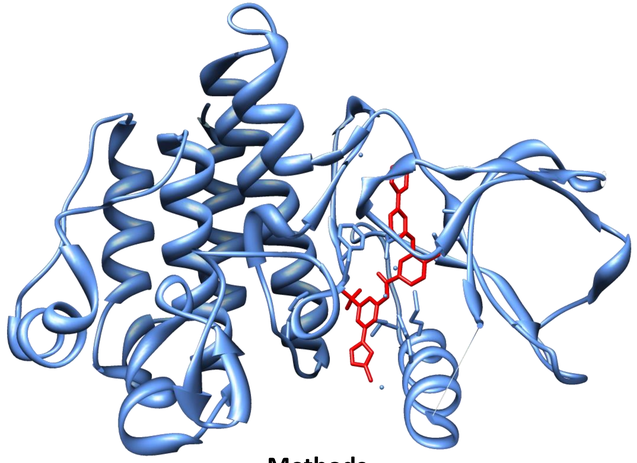
XGBoost Models to Evaluate 3D Ligand Representations

Wern Juin Gabriel Ong[1 2], Grigorii V. Andrianov[1], and John Karanicolas[1] | Karanicolas Lab -- Fox Chase Cancer Center
Fox Chase Cancer Center[1] and Bowdoin College[2]

Motivations

Accurate 3D representations are key to training effective deep learning models.

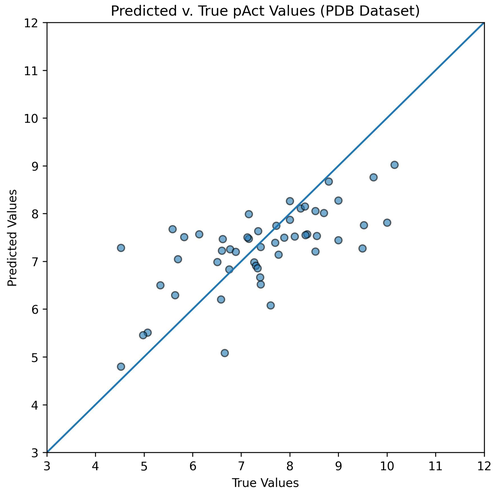
- Kinases are highly druggable biological targets broadly implicated in biological disorders.
- Previous work in the Karanicolas Lab has demonstrated the need for structure-based ligand representations.
- However, it is unclear if these 3D structure-based representations are accurate representations of the ground truth.

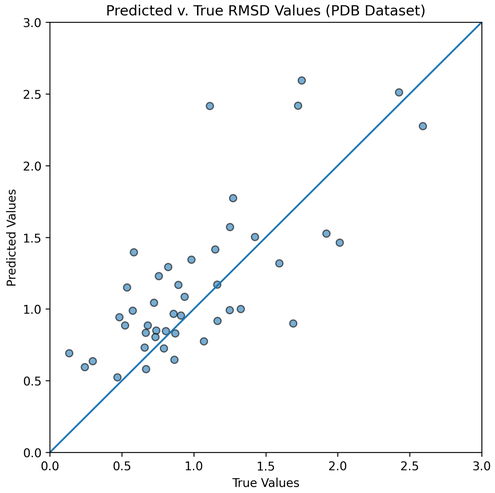


Methods

- We start with a curated subset of kinase-inhibitor complexes in the PDB with recorded binding affinity that are then standardized to pAct scores.
- Train XGBoost pipeline to predict binding affinity from energy features.
- We hypothesize that our pipeline will accurately predict pAct for models close to ground truth – we can then use this XGBoost pipeline to select between models when a PDB structure is unavailable.
- Features used:
  - Rosetta energy components
  - RDKit features
  - OpenEye Omega features

Results





Pipeline	Pearson R	MSE
Predicting RMSD on PDB Dataset	0.250	3.231
Predicting pAct on PDB Dataset	0.671	1.041

Discussion

- Our pipeline demonstrates competence on a range of tasks and is able to recapitulate both RMSD (especially at the lower ranges) and pAct.
- The ability to recapitulate pAct from energy features is a promising sign that our pipeline can discriminate between kinase-inhibitor models with poor RMSD.
- This is further evidenced by the fact that there is a performance decrease when we use the Christmann-Franck database that does not rely directly upon PDB structures (Pearson R: 0.354).

Future Directions

- Using similarity scores – such as Maximum Common Substructure – to cluster inhibitors so as to ensure minimal information leakage.
- Further filtering the data to ensure only models with a sufficiently small RMSD are used as training data
- Training on different sets of features – Rosetta energy components only or RDKit features only.

Acknowledgements

Many thanks to Grigorii, John, and Kiruba for their suggestions and assistance throughout this project.

References

Christmann-Franck, Serge, et al. "Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design?" *Journal of Chemical Information and Modeling*, vol. 204, 2 Aug. 2016, pp. 1654-1675, doi:.org/10.1021/acs.jcim.6b00122.
Francoeur, Paul G., et al. "Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design." *Journal of Chemical Information and Modeling*, vol. 60, no. 9, 31 Aug. 2020, doi:.org/10.1021/acs.jcim.0c00411.

Hassan-Harrirou, Hussein, et al. "RosENet: Improving Binding Affinity Prediction by Leveraging Molecular Mechanics Energies with an Ensemble of 3D Convolutional Neural Networks." *Journal of Chemical Information and Modeling*, vol. 60, no. 6, 11 May 2020, pp. 2791-2802, doi:.org/10.1021/acs.jcim.0c00075.

Heinzinger, Michael, et al. "Modeling aspects of the language of life through transfer-learning protein sequences." *BMC Bioinformatics*, vol. 723, 17 Dec. 2019, doi:.org/10.1186/s12859-019-3220-8.

Jiménez, Jose, et al. "KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks." *Journal of Chemical Information and Modeling*, vol. 58, no. 2, 8 Jan. 2018, pp. 287-296, doi:.org/10.1021/acs.jcim.7b00650.
Laufkötter, Oliver, et al. "Identifying representative kinases for inhibitor evaluation via systematic analysis of compound-based target relationships." *European Journal of Medicinal Chemistry*, vol. 204, 15 Oct. 2020, doi:.org/10.1016/j.ejmech.2020.112641.

Rifaioglu, Ahmet S., et al. "MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery." *Bioinformatics*, 7 Dec. 2020, doi:.org/10.1093/bioinformatics/btaa858.