Evaluating Voxelized Representations with XGBoost Models

With Some Footnotes to Plato

Gabriel Ong[1]

[1]Karanicolas Lab
Program in Molecular Therapeutics - Fox Chase Cancer Center

Group Meeting, 18th May 2021

Table of Contents

Table of Contents

Some Old Wisdom from Cambridge

A Quote

The safest general characterization of the European philosophical tradition is that it consists of a series of footnotes to Plato.

–A.N. Whitehead

A Revalation!

Nb. Permission was sought to use this screenshot.

On the Correctness of Names

Cratylus 383a3-b1

Herm.: Cratylus says, Socrates, that there is a correctness of name (*onomata*; names/nouns) for each thing, one that belongs to it by nature. A thing's name isn't what people agree to call it – some bit of their native language that applies to it – but there is a natural correctness of names, which is the same for everyone, Greek or foreigner.

Question: Are the representations (names) for kinase-inhibitor complexes correct?

Translation from Cooper, ed. *Complete Works of Plato*; Greek from Burnet, ed. *Opera I*

Cratylus was Right

Cratylus 386d10-e3

Soc.: ... then it is clear that things have some fixed being or essence of their own. They are not in relation to us and are not made to fluctuate by how they appear to us. They are by themselves, in relation to their own being or essence, which is theirs by nature.
Herm.: I agree, Socrates.

- In previous work with Kiruba, we have shown that SMILES strings (names that we "agree" to call molecules) make bad representations for machine learning.
- We need representations that represent 'natural essences' – a representation that captures the ground truth of a kinase-inhibitor reaction.

Table of Contents

The Importance of Kinases

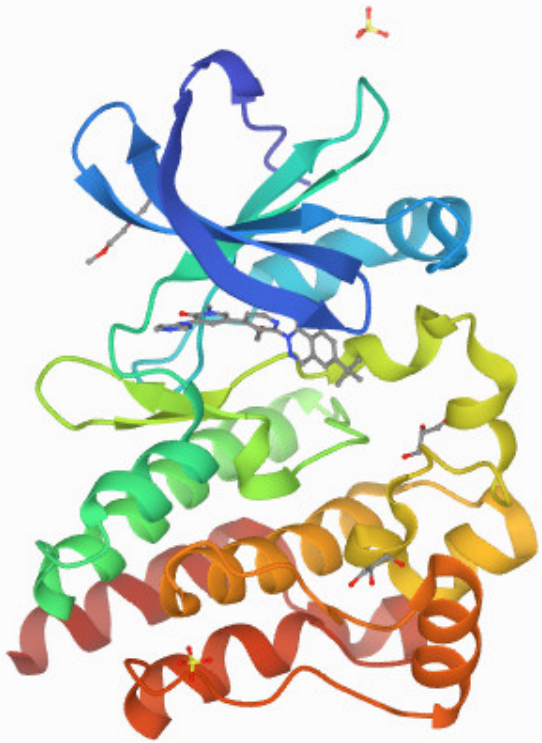
Kinases play an important role in several human diseases:

- Cancers
- Inflammatory diseases
- Autoimmune disorders

Kinases already have a strong record as pharmaceuticals with over 48 drugs gaining FDA approval.

Building Voxelized Representations

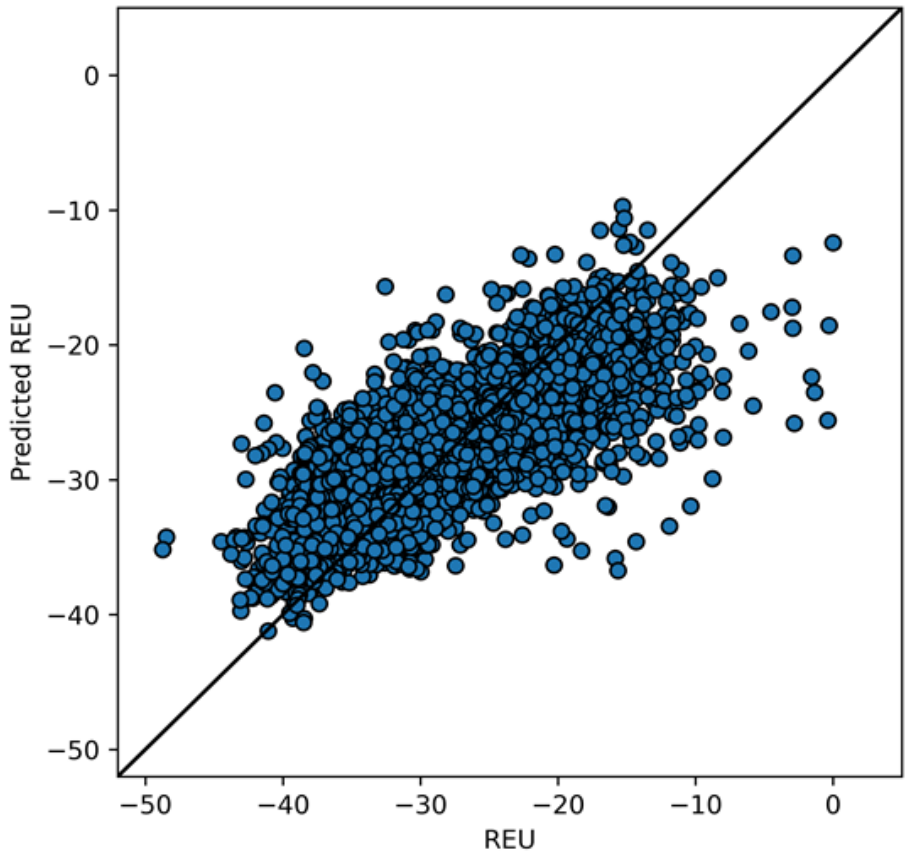
- Take the SMILES string of a given inhibitor and generate conformers in Omega.
- Align conformers against the PDB structure with structurally similar inhibitors.
- Keep the top 10 models by Rosetta Energy.



Some Promising Results

- Our 3D CNN can reasonably predict Rosetta Energies from our 3D voxelized input.

Pearson R: 0.76
MSE: 20.79



Bridging the Gap

Our model seems to perform reasonably well predicting Rosetta Energies from the voxelized 3D representations.

- Can we somehow use ML techniques to discriminate between voxelized representations?
- Is there a "nice" relationship between Rosetta Energies and binding affinity?

Table of Contents

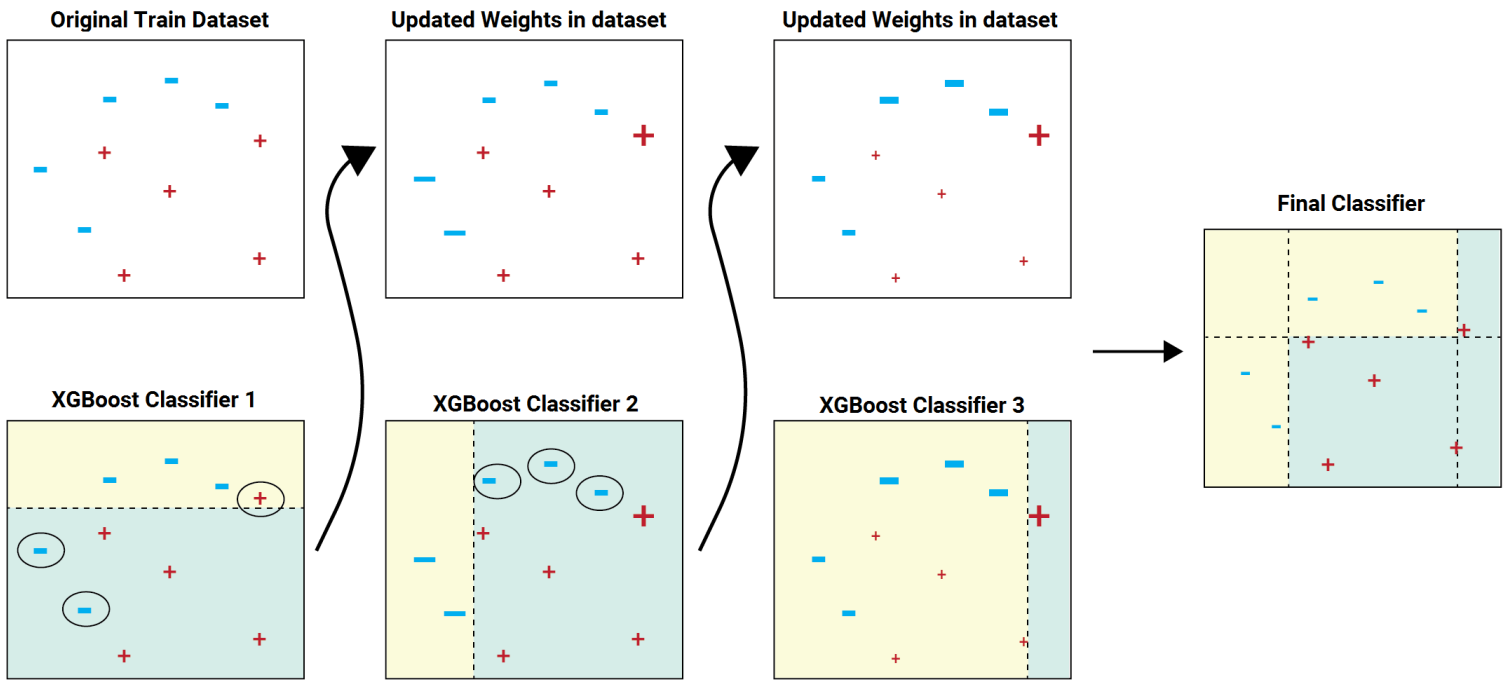
The Preliminaries

Hypothesis

Suppose we have a pipeline that can accurately predict binding affinity from Rosetta component energies. Should we have a model where RE components do not match binding affinity, we can reasonably assume that that model is not representative of the ground truth.

Step 1: Build models that accruately predict binding affinity from RE components.

What are XGBoost Models?



- A series of consecutive decision trees that successively correct each other.

Methods (Dataset Preparation)

1. Take SMILES string of a given inhibitor from Christmann-Franck dataset and generate conformers in Omega.
2. Align conformers against PDB structure with structurlaly similar inhibitors.
3. Keep top 10 models by Rosetta Energy (Total/Interaction)
4. Calculate component energies for each model.

Methods (Model Training)

Use Yusuf's hyperparameters as outlined in vScreenML. Training done in Google Colab.

- 1945 Estimators
- 7 Maximum Depth

All 10 models used in training for Components $\rightarrow$ RE; only model with lowest energy used for training Components $\rightarrow$ pAct.

A Proof of Concept (Total Component RE $\rightarrow$ RE)

- Trivial task: total RE is weighted sum of components.
- Some models are very poorly aligned and have high RE.

Pearson R: 0.999
MSE: 3575

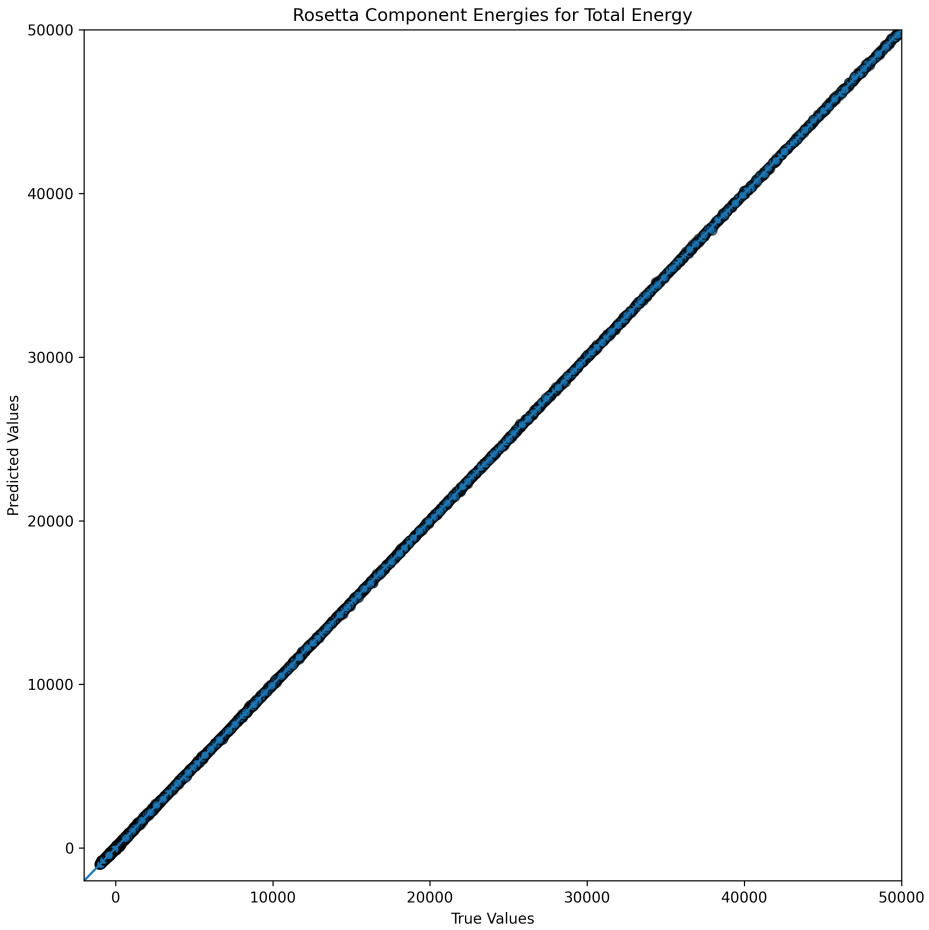
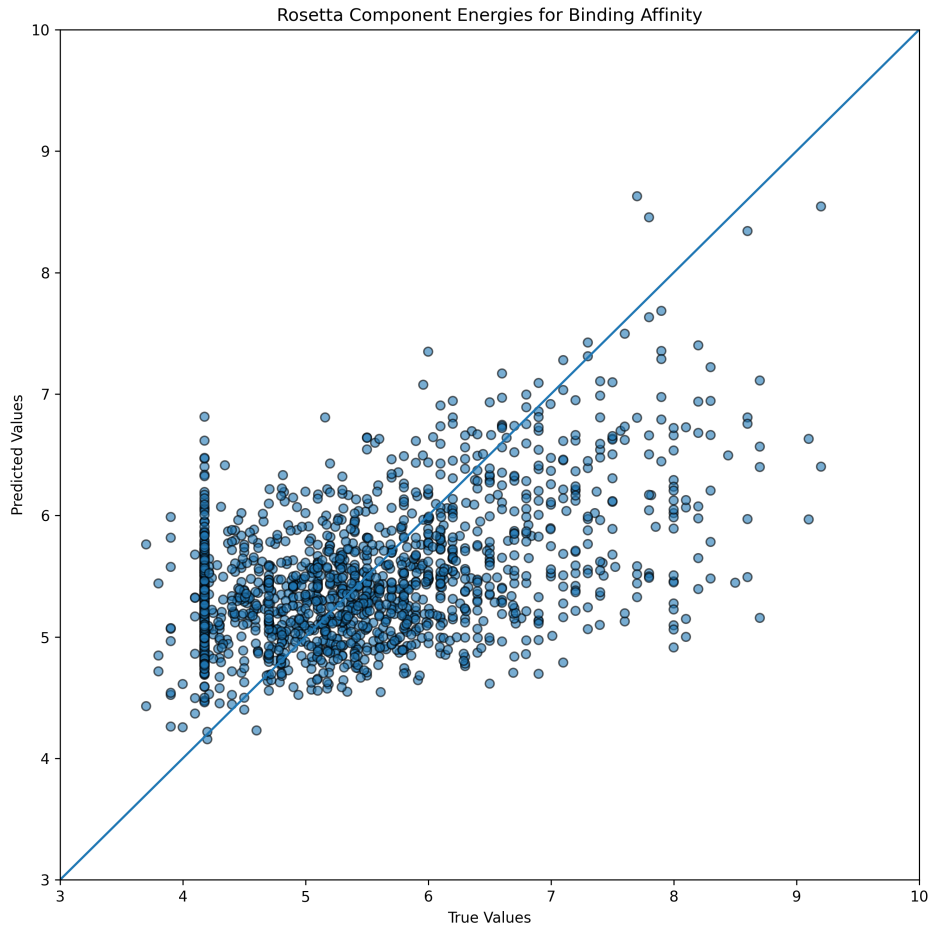


Table of Contents

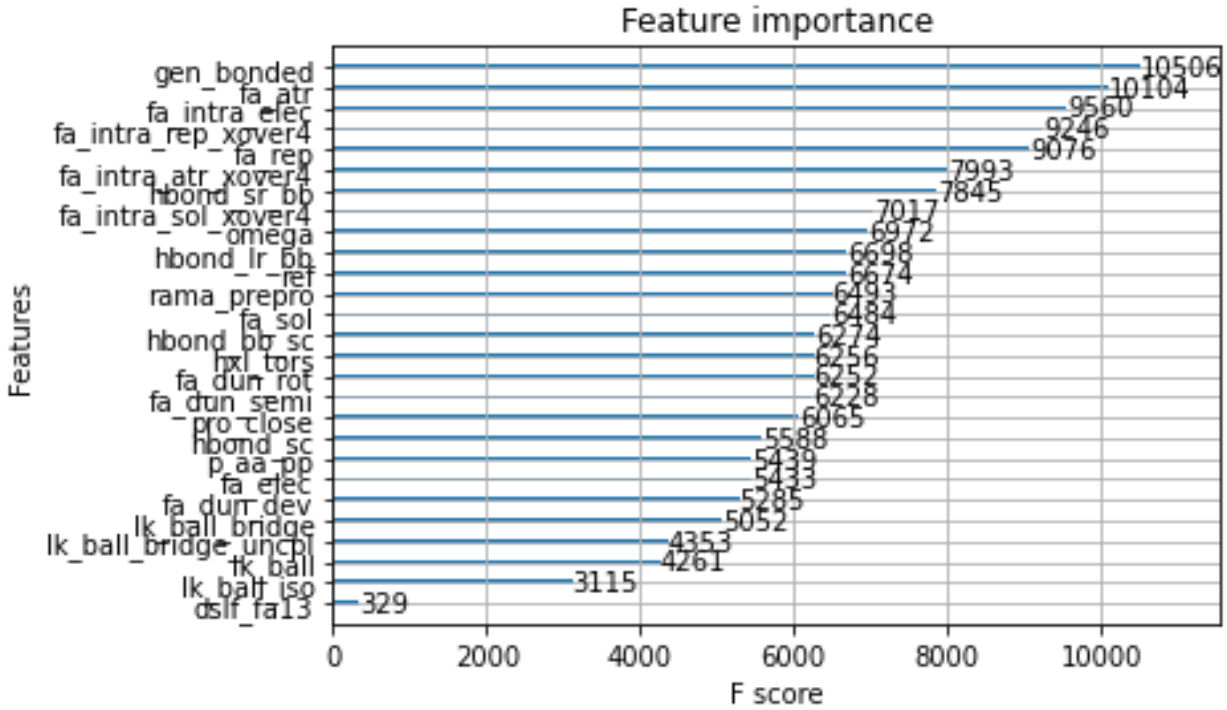
Results (Total Component RE $\rightarrow$ pAct)

- Significant performance dropoff from Component $\rightarrow$ Total.
- Adjusting hyperparameters to increase model capacity (more estimators, greater depth) does not alter performance.

Pearson R: 0.357
MSE: 0.948



Results (Total Component RE → pAct)

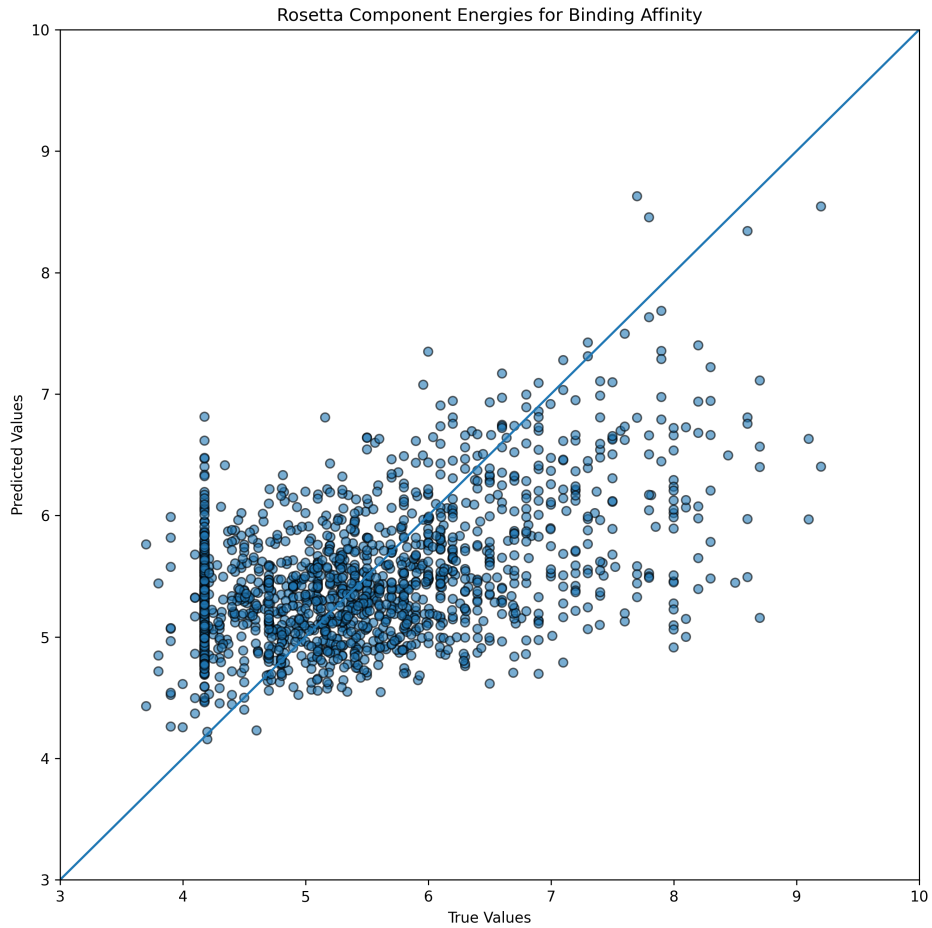


Higher F Score = More Important Feature

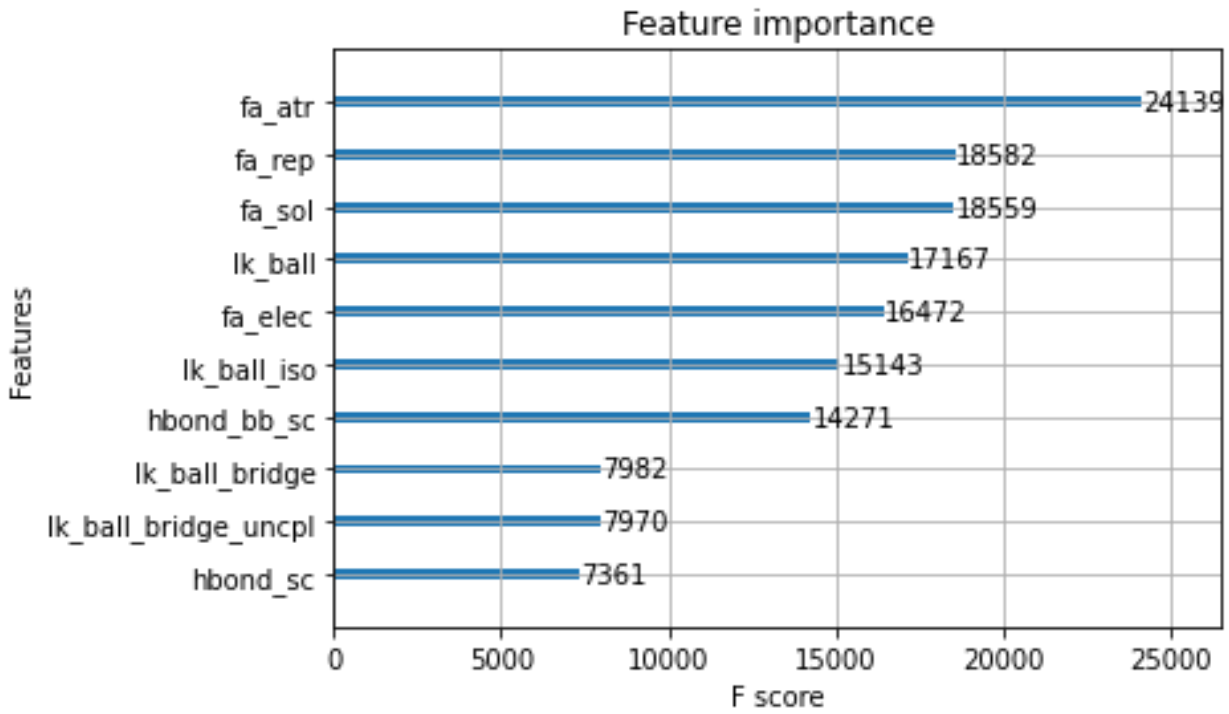
Results (Interaction Component RE → pAct)

- Performance similar to that of Total Components → pAct.
- Interaction energy eliminates features that are solely dependent on the protein.

Pearson R: 0.350
MSE: 1.055

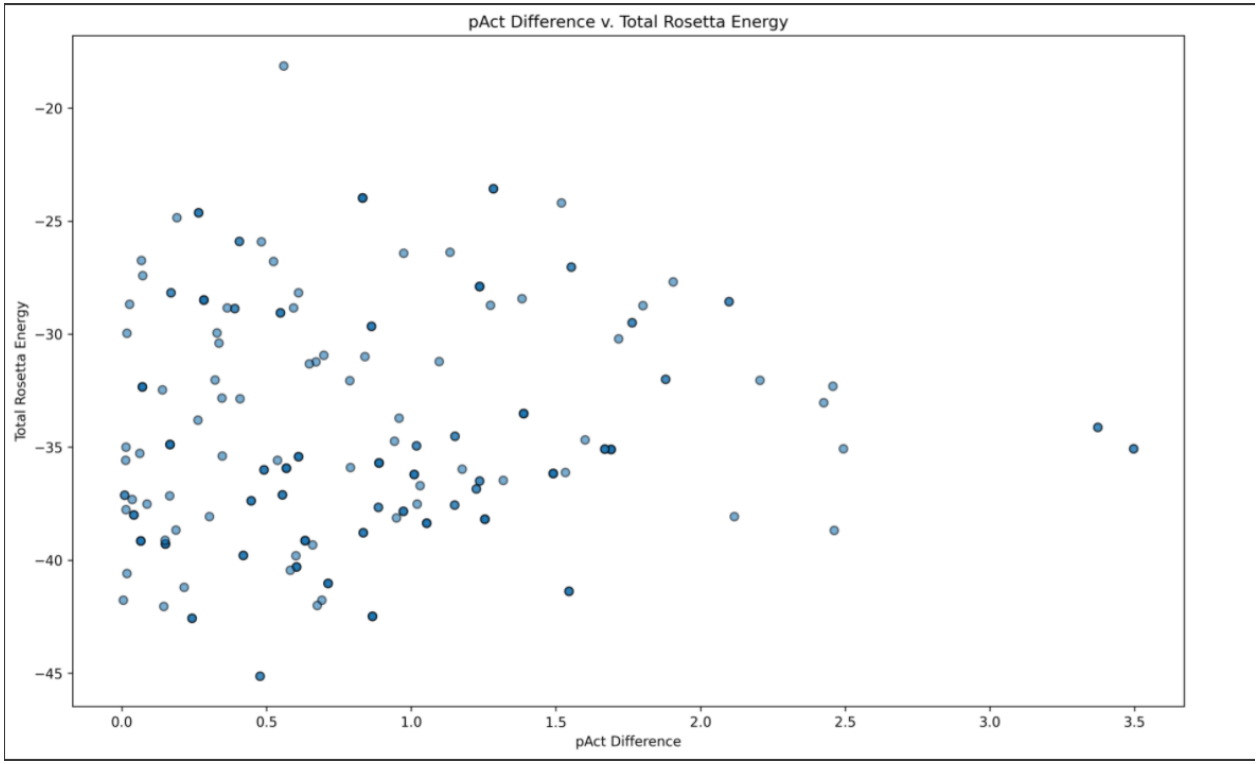


Results (Interaction Component RE → pAct



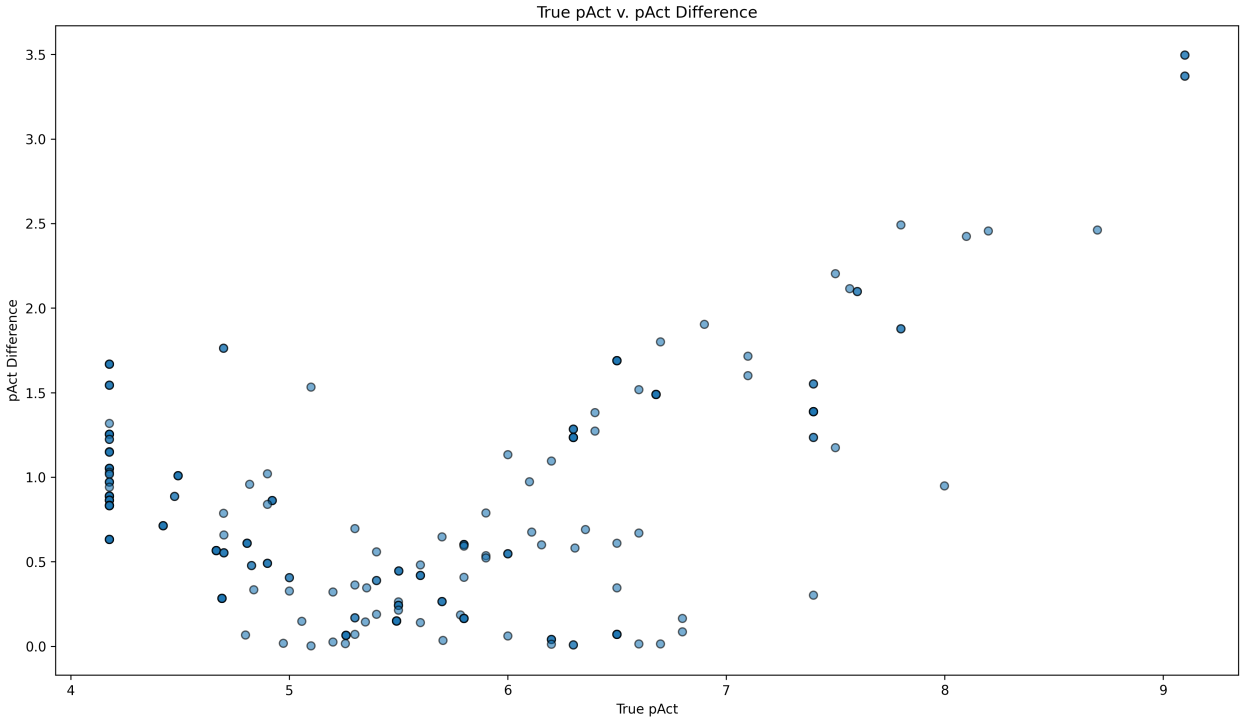
Since more features are identically 0, there are different weights for features when compared to Total Component RE → pAct.

When does our pipeline predict pAct poorly?



There is a weak correlation between Interaction Energy and pAct difference. As RE increases, the more likely it is that our model will make an incorrect prediction. Why?

When does our pipeline predict pAct poorly?



It seems our model is making predictions around the mean (5.5).
What does this tell us about the data?

Table of Contents

Next Steps

Kiruba has a set of models with RMSD measured against PDB structures.

- Develop metrics (other than RE) that can discriminate the lowest RMSD structure from a set.
- Retrospectively apply it to our representations for the 3D CNN binding affinity project.

vScreenML incorporates additional features such as interatomic interactions and ligand properties.

- Will incorporating these features improve our model performance?

Thanks and Acknowledgements

Thanks to:

- John Karanicolas
- Grigorii Andrianov
- Jake Khowsathit
- Chris Parry
- Mariam Fouad

- Daniel Yeggoni
- Sven Miller
- Lei Ke