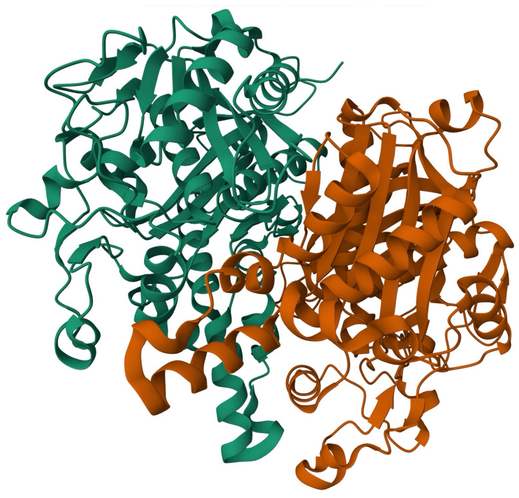
Predictive Modeling of Small Molecule Binding Affinity through a NequIP-based Equivariant Neural Network

Vladislav Cherdantsev

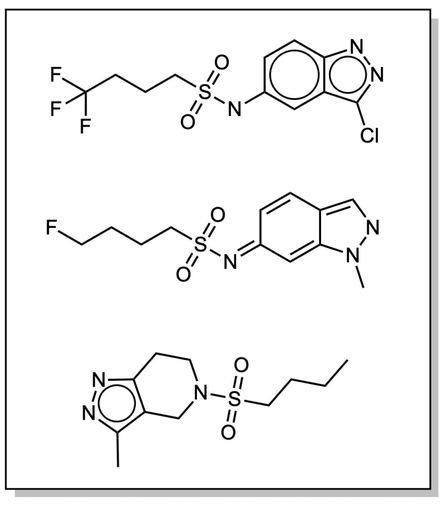
Background and Motivation



Accurately predicting protein-ligand interactions is paramount in drug design. In recent years, the advent of deep learning models has significantly contributed to addressing this challenge. Many of these models leverage 3D structures of protein-ligand interactions as input data, enabling more precise predictions of binding affinity.

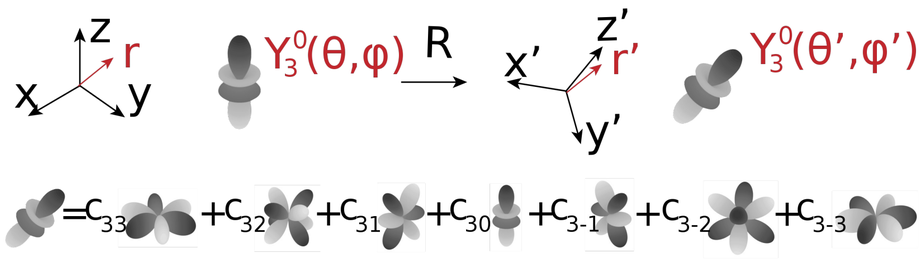
Antitubercular drug discovery is one example of an area that might benefit from these methods.

TB drug discovery faces limited success, producing only two FDA-approved drugs in the last fifty years. One potential target, β-ketoacyl synthase KasA, has a characterized bound inhibitor, while another hit molecule with promising efficacy acts on multiple targets within mycobacterial cell wall biosynthesis. Our research introduces an equivariant neural network to predict antitubercular hit properties from in vitro assays, including minimum inhibitory concentration (MIC).



Methods

Although our aim is to predict an energy value that remains invariant to rotations and translations, we may still benefit from the enhanced expressivity offered by symmetry-aware models. One such successful equivariant model for energy prediction is NequIP, which utilizes E(3)-equivariant convolutions to handle interactions of geometric tensors.



Although NequIP embeds features solely based on atomic numbers, we sought to investigate the impact on model accuracy by substituting atomic numbers with **more descriptive atom types** and incorporating **partial charges** into node features. The atomic numbers are embedded into l=0 features, which are refined through a sequence of interaction blocks, generating scalar and higher-order tensor features:

$$f_i' = \frac{1}{\sqrt{z}} \sum_{j \in \mathcal{N}(i)} f_j \otimes (R(r_{ij}))Y_m^{(l)}(\hat{r_{ij}})$$

with the following filter:

$$F_m^{(l)}(\vec{r_{ij}}) = R(r_{ij})Y_m^{(l)}(\hat{r_{ij}})$$

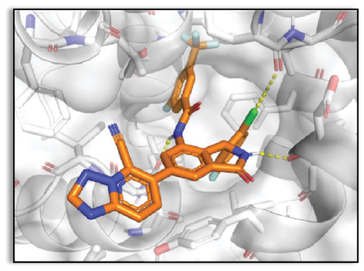
$$R(r_{ij}) = W_n\sigma(\ldots\sigma(W_2\sigma(W_1B(r_{ij}))))$$

Embedding: One-Hot Atom Encoding, Radial Basis Edge Embedding, Spherical Harmonics Edge Embedding.

Interaction Block: Encodes interactions between neighboring atoms using a convolution function, combined with linear atom-wise self-interaction layers and a ResNet-style update, followed by processing through an equivariant SiLU-based gate nonlinearity.

Output Block: The final convolution's l=0 features undergo processing via a couple of self-interaction linear layers.

Data and Featurization



GAFF Parametrization + Preprocessing

1. 3D Coordinates
2. Partial Charges
3. Atom Types Assigned by the Force Field

- Binding poses were derived from a combination of experimental data and energy minimization techniques.
- Atom types were assigned according to the General Amber Force Field (GAFF) parametrization and are distinct from the atomic numbers.
- Partial charges were determined using the AM1-BCC charge scheme.

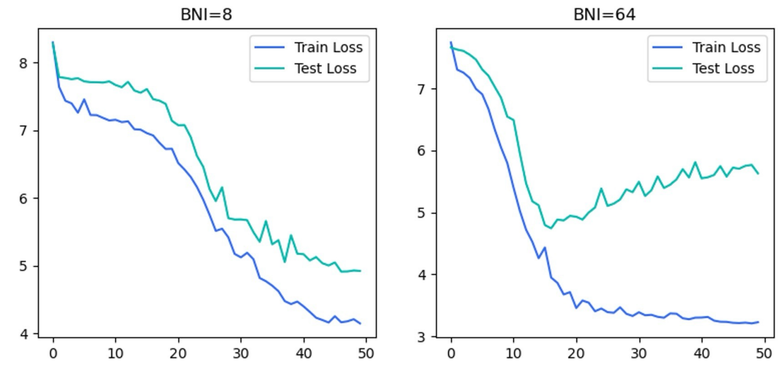
Atom type	Description
c	sp^2 carbon in C═O, C═S
c2	sp^2 carbon, aliphatic
ca	sp^2 carbon, aromatic
n1	sp^1 nitrogen
n3	sp^3 nitrogen with 3 subst.

Results and Conclusions

HP Name	HP Value	R^2 score	HP Name	HP Value	R^2 score	HP Name	HP Value	R^2 score
LMAX	0	0.63	INVARIANT_LAYERS	1	0.64	RESNET	True	0.59
	1	0.65		2	0.65		False	0.64
	2	0.67		3	0.62	USE_SC	True	0.63
	3	0.63	INVARIANT_NEURONS	16	0.60		False	0.49
	4	0.65		32	0.64	BATCH_SIZE	2	0.56
NUM_BASIS	6	0.62		64	0.64		5	0.65
	8	0.63		128	0.63		10	0.65
	10	0.58	NUM_CONV_LAYERS	1	0.62	LEARNING_RATE	1e-2	0.46
BNI	8	0.71		2	0.64		5e-3	0.52
	16	0.67		3	0.66		1e-3	0.66
	32	0.60		4	0.70		1e-4	0.52
	64	0.57		5	0.69		1e-5	0.52

In the first experiment, the hyperparameters that achieve the highest accuracy for NequIP with atomic numbers as input were determined. R2-scores served as a metric of the model's accuracy.

The hyperparameter that exerted the most pronounced impact on the model's accuracy was BNI (base number of irreps) , which represents the multiplicity of irreps of each type in hidden layers.



The need for a smaller number of irreps in training the data could be attributed to the dataset's relatively small size (approximately 300 molecules) and the structural similarities among the ligands. Notably, when larger values of BNI were used, overfitting was observed, as evidenced by the train/validation curves.

Both modifications yielded models with higher accuracies. This improvement could be attributed to the fact that the new atom types inherently capture more chemically relevant information, especially considering that GAFF is commonly used to parametrize small molecules.

Featurization	R^2-score, average
Atomic Numbers	0.70
GAFF Atom Types	0.76
GAFF Atom Types + Charges	0.79

The inclusion of partial atomic charges additionally improved the model and stabilized the training process.