Using pre-training to improve binding affinity predictions

Jesse Murray

What is pre-training?

- Typically, the initial phase of training a neural network involves random weight and bias assignment.
- Pre-training, however, involves borrowing these weights and biases from a previously trained network, which already holds valuable pattern understanding.
- This method mirrors the use of informative Bayesian priors over representations, which can help reduce chances of overfitting by providing a constructive bias.

Challenges in predicting binding affinity

- **Data Scarcity:** Models predicting binding affinity face a data limitation issue, with only 19,443 structures available in PDBbind for learning.
- **Data Quality:** The structures in PDBbind are of moderate quality, with a median resolution of 2Å.
- **Tendency for Overfitting:** These models often overfit the training data, leading to substantial variability and poor generalization.
- **Performance Concerns:** High scores on CASF-2016 might be attributed to possible data-leakage between CASF-2016 and the remainder of PDBbind.

Advantages of Pre-training in Addressing Challenges

- **Larger Datasets**: Pre-training enables the use of significantly larger datasets than those limited to protein-ligand complexes with known binding affinity values.
- **Improved Data Quality**: Pre-training datasets can provide superior quality data, including improved resolution and more precise energetics values.
- **Mitigating Overfitting**: By introducing a physics-based bias via pre-training, we can effectively reduce the model's variance.
- **Enhanced Generalization**: Given that the principles of physics are consistent across novel protein-ligand contexts, a model with a physics-bias should offer better generalization capabilities.

Energetics as a pre-training regime

- A pre-trained model that correlates structure with energetics could potentially offer valuable features for mapping structure to binding affinity.
- The connection between energetics and p*K* is shown on the right.
- The energetics of intermolecular interactions could be especially relevant to the energetics of the interaction between protein residues (P) and the ligand (L).

$$\mathrm{p}K = \log[PL] - \log[P] - \log[L]$$

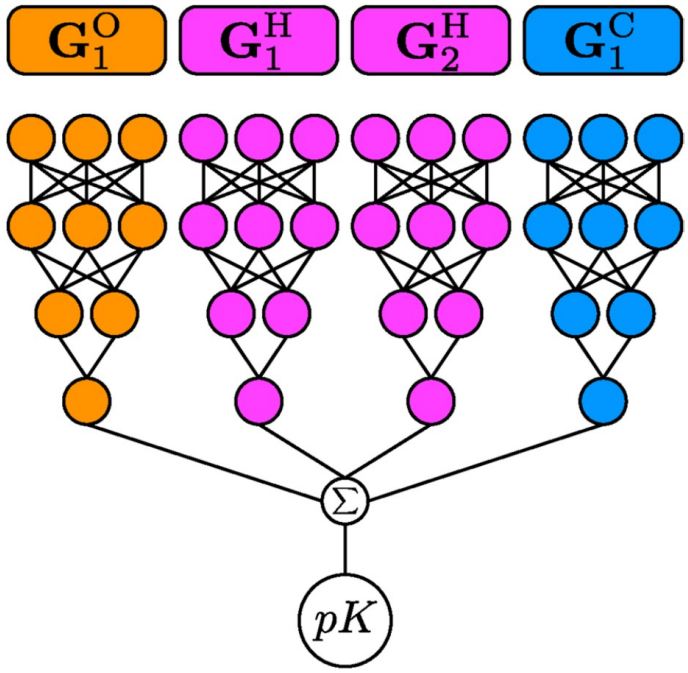
$$\mathrm{p}K = -\frac{1}{\ln(10)}\frac{\Delta G_0^{\mathrm{bind}}}{RT}$$

$$\mathrm{p}K \propto \Delta\mathrm{H} = E_{PL} - E_P - E_L$$

Preliminary experiment to test this idea

Rocco Meli proposed **AEScore** to predict binding affinity

- Each atom in the complex encodes its surrounding ligand and protein atoms in an atomic environment vector (AEV)
- The AEV is a function of the atom-types, distances, and angles of the surrounding atoms.
- The AEV for an atom is fed into the feedforward NN, specific to the atom's element.
- The atomic outputs are summed to get pK.
- Achieved 0.80 on CASF-2016



AEScore architecture is almost identical to ANI-2X architecture

- ANI2x has a higher angular resolution in the AEV
  - (AEV length is 1008 versus 270)
- ANI2x has slightly different layer sizes (256-192-160-1 for H), (160-128-96-1 for S/F/CL), etc.
  - 256-128-64-1 for AEScore

ANI-2X is similar to AEScore, but pretrained

- Only molecules containing H, C, N, O, S, F, and Cl

Trained on:

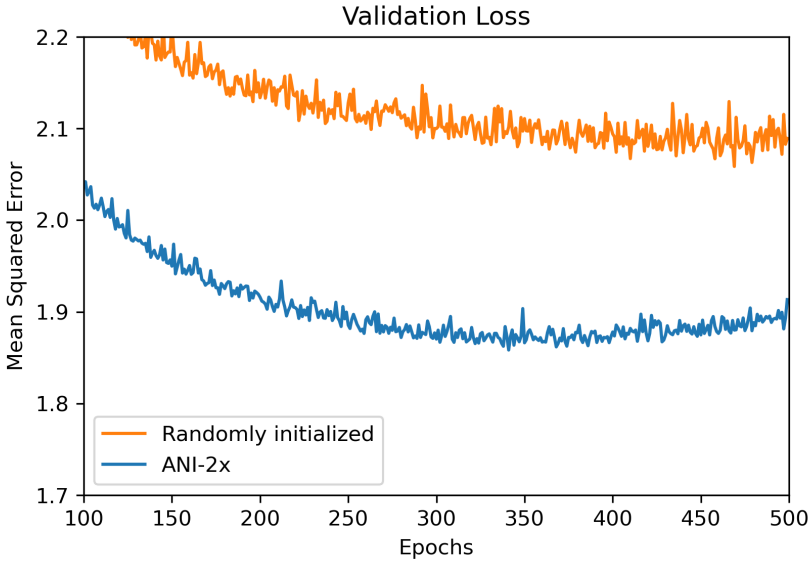
- GDB-11 database
- CheMBL database
- S66x8 (dimers)
- ANI-1x and ANI-ccx database
- 8.9 million molecular conformations
- Normal mode sampling of s66x8 dataset (eight locations along dissociation path of C, H, N, and O-containing dimers with HB, pi-stacking, and LDFs). 195,000 conformations of the 66 dimers.
- MD simulations of bulk water

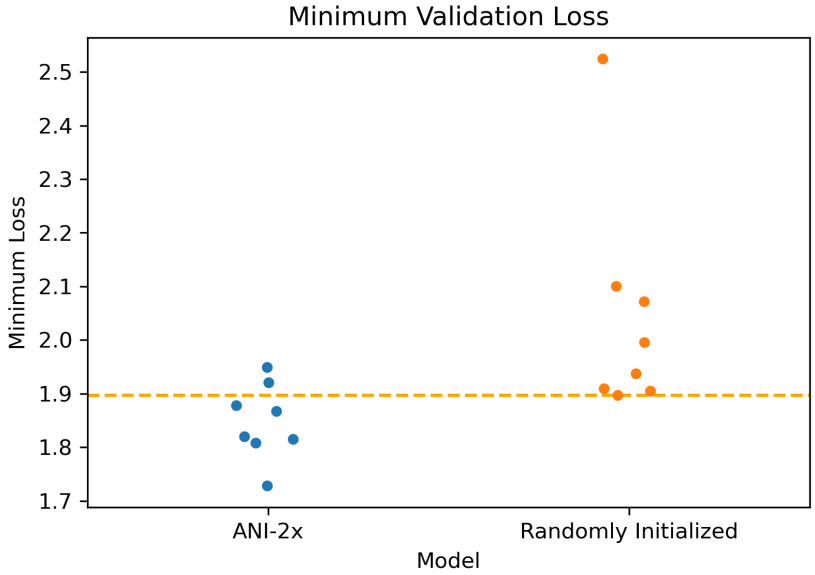
Experimental design to evaluate the impact of pre-training

1. Train and evaluate the ANI-2x model using PDBbind data, utilizing ANI-2x parameters for initial settings.
2. To gauge the impact of pre-training, establish an identical copy of ANI-2x as a control but randomly initialize the parameters before training. The only distinction between the control and test conditions will thus be the initial parameters.
3. Any performance disparity between the pre-trained ANI-2x and the randomly initialized model can be attributed to the effects of pre-training. We have the advantage of having eight ANI-2x models for this purpose.

We only include PDBbind complexes with H, C, N, O, S, F, and Cl in the binding pocket for this experiment.

Performance Comparison: Pre-trained vs Randomly Initialized Models





- The pre-trained model demonstrates marginally better performance.
- The pre-trained model scores 0.84 PCC on CASF-2016, compared to 0.79 for the randomly initialized counterpart.

Intriguing characteristics of ANI-2x initializations vs random initializations

1. Pre-trained models reach convergence faster and overfit faster, indicating that there is a useful prior representation in the network.
2. These models exhibit a more stable validation loss between parameter updates on Stochastic Gradient Descent (SGD).
3. These models also present a slightly lower correlations between predictions on CASF-2016, indicating more feature variability in the convergence.
4. In the ligand-only comparison (where only the structure of the ligand is used to predict affinity), the pretrained model significantly outperforms random initializations: 0.6 vs 0.1 PCC on CASF-2016

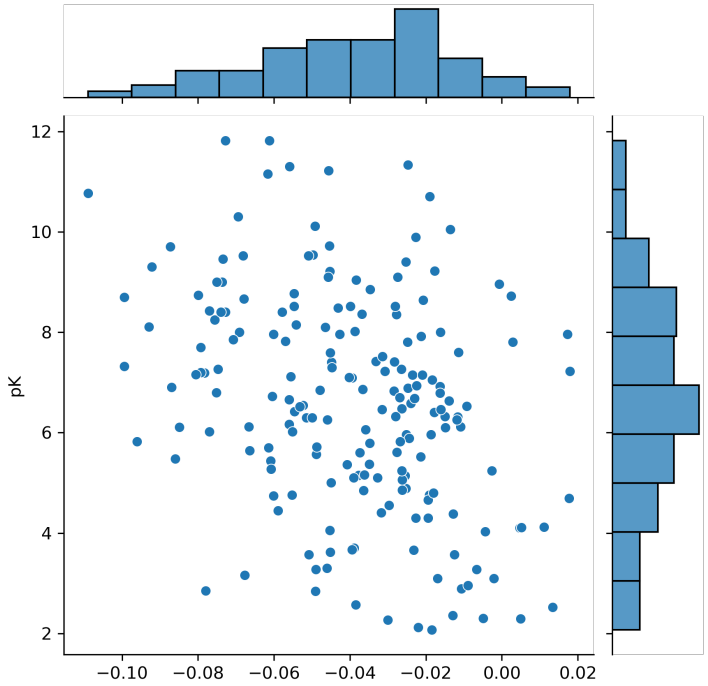
Future directions with pre-training

- Experiment with using the last layer of ANI-2x as input to a neural network (freezing the ANI-2x layers) and then fine-tuning the whole of ANI-2x to obtain a deeper model with less catastrophic interference.
  - (Update: did not work.)
- Use the TorchMD-Net model trained on the SPICE dataset *(Eastman et al., 2023, Nature)*
- Experiment with showing the model the bound and unbound complexes to better capture the *difference* in energy between the two.

Usefulness of pre-trained model before finetuning

- In order to evaluate the relevance of pre-training with respect to interaction energetics, we employ the following approach with the ANI-2x model:

1. Predict the energy of the ligand when it's within the protein and subtract from it the energy of the ligand in a vacuum.
2. The resulting energy difference (represented on the x-axis), signifies the bound and unbound states, exhibits an expected correlation with CASF-2016 affinity values (-0.33 PCC).



Questions?