Exploring $\Delta\Delta G$ prediction with Siamese Networks

Andrew McNutt and David Koes
Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh,PA
Carnegie Mellon - University of Pittsburgh Joint Program in Computational Biology

Abstract

During lead optimization, lead molecules are refined for potency via slight modifications of their chemical structure. Relative binding free energy (RBFE) methods allow comparisons of molecular potency during this optimization. We utilize a Siamese Convolutional Neural Network (CNN) to directly estimate the RBFE with higher throughput than simulation based methods. Our models show improved performance over a previously published Siamese RBFE predictor. We observe decreased performance on out-of-domain RBFE predictions.

Background

Lead optimization:

- Early phase of the drug discovery process
- Simultaneously optimize a lead molecule for many pharmaceutical properties.
- Modifications made to the chemical scaffold of the lead molecule and tested for their effect on the properties of interest
- Congeneric series - collection of molecules with slight modifications on same scaffold structure

Relative Binding Freee Energy (RBFE) Methods:

- Typical methods:
  - Molecular dynamics with alchemical perturbations (i.e. FEP)
  - Thorough sampling of the endpoints of the transformation(i.e. MMGBSA, MMBPSA)
- Methods suffer from:
  - lack of applicability to large changes between ligands
  - high computational cost of prediction

Machine Learning (ML) Binding Affinity Predictors:

- low error and high throughput for absolute binding affinity predictions
- Jimenez-Luna et al.[1] utilize a Siamese Convolutional Neural Network (CNN) architecture to predict RBFE

Siamese Network

- Two arms with shared weights take in two inputs
- Usually for determining distances between inputs

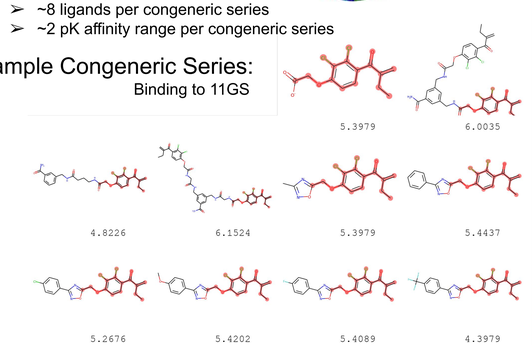
1. Jiménez-Luna, José, et al. "DeltaDelta neural networks for lead optimization of small molecule potency." *Chemical science* 10.47 (2019): 10911-10918.

Data

BindingDB 3D Strucure Series

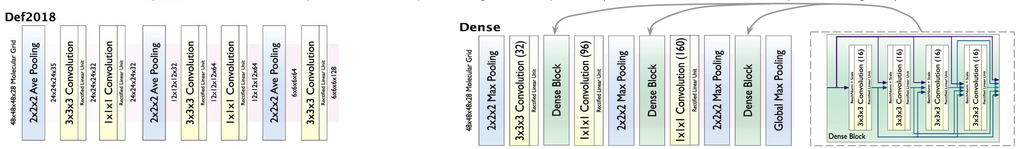
- 943 unique receptor structures
- 1082 congeneric series
- ~8 ligands per congeneric series
- ~2 pK affinity range per congeneric series

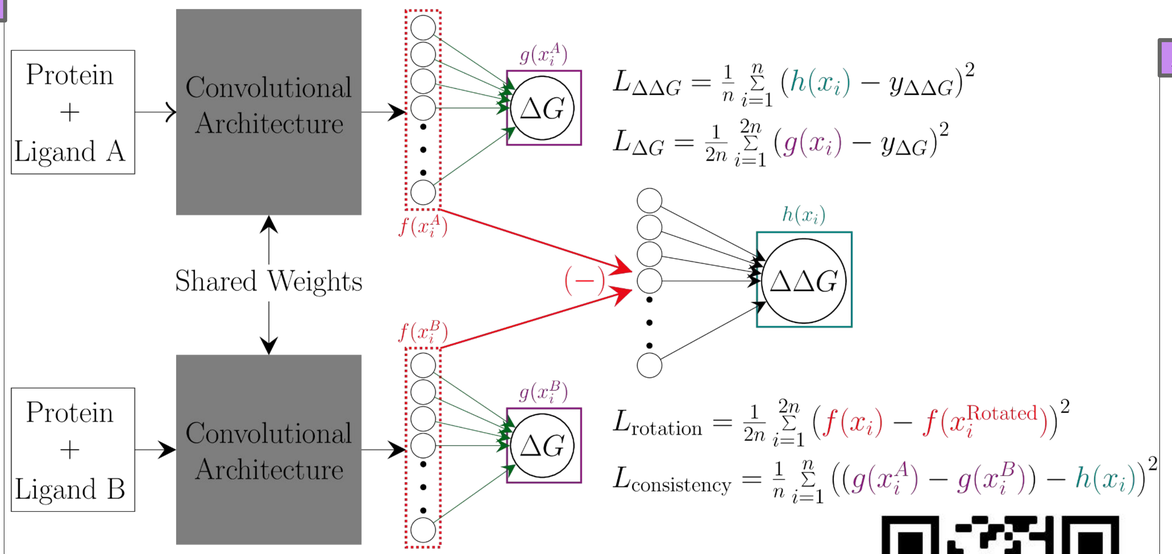
Example Congeneric Series:
Binding to 11GS



Convolutional Architectures

Inputs: 3D voxelized representation of protein-ligand complexes (14 channels for both receptor and ligand)





$$L_{\Delta\Delta G} = \frac{1}{n}\sum_{i=1}^{n}\left(h(x_i) - y_{\Delta\Delta G}\right)^2$$

$$L_{\Delta G} = \frac{1}{2n}\sum_{i=1}^{2n}\left(g(x_i) - y_{\Delta G}\right)^2$$

$$L_{\text{rotation}} = \frac{1}{2n}\sum_{i=1}^{2n}\left(f(x_i) - f(x_i^{\text{Rotated}})\right)^2$$

$$L_{\text{consistency}} = \frac{1}{n}\sum_{i=1}^{n}\left(\left(g(x_i^A) - g(x_i^B)\right) - h(x_i)\right)^2$$

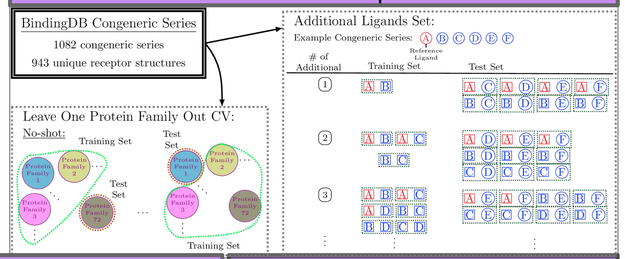
Multi-task loss increases performance

Ablation	Pearson's R	RMSE (pK)	MAE (pK)
Standard	$\mathbf{0.553(\pm 0.0233)}$	$\mathbf{1.11(\pm 0.0309)}$	$\mathbf{0.82(\pm 0.0187)}$
No $L_{\Delta\Delta G}$	$\mathbf{0.551(\pm 0.0202)}$	$\mathbf{1.12(\pm 0.0248)}$	$\mathbf{0.829(\pm 0.0179)}$
No $L_{\Delta G}$	$0.459(\pm 0.0238)$	$1.27(\pm 0.0289)$	$0.945(\pm 0.0182)$
No L_{rotation}	$\mathbf{0.556(\pm 0.0188)}$	$\mathbf{1.11(\pm 0.0233)}$	$\mathbf{0.819(\pm 0.0162)}$
No $L_{\text{consistency}}$	$\mathbf{0.536(\pm 0.021)}$	$1.14(\pm 0.0356)$	$0.842(\pm 0.0186)$
No $L_{\Delta\Delta G}$, $L_{\text{consistency}}$	$-0.0576(\pm 0.136)$	$1.24(\pm 0.0143)$	$0.908(\pm 0.0144)$
No $L_{\Delta G}$, $L_{\text{consistency}}$	$0.456(\pm 0.0231)$	$1.28(\pm 0.0319)$	$0.95(\pm 0.0233)$
Concatenation	$\mathbf{0.554(\pm 0.0134)}$	$\mathbf{1.11(\pm 0.0223)}$	$\mathbf{0.821(\pm 0.0174)}$
No Siamese Network	$0.5(\pm 0.0347)$	$1.15(\pm 0.0362)$	$0.854(\pm 0.021)$

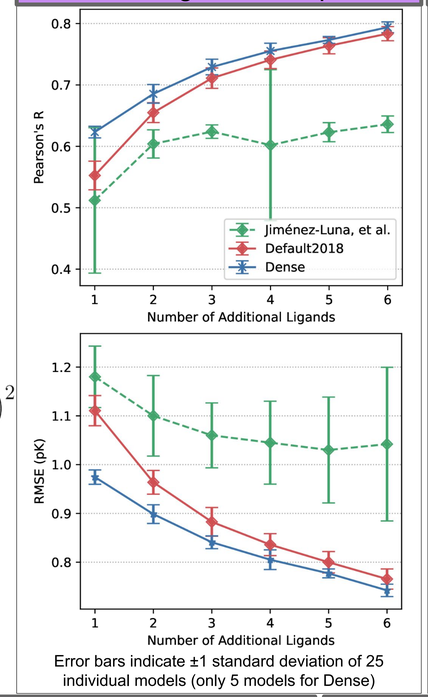
Bold indicates that results are not significantly different from Standard ($p > 0.05$)

This work is supported by R35GM140753 from the National Institute of General Medical Sciences and CHE-2102474 from the National Science Foundation

Evaluations of $\Delta\Delta G$ prediction

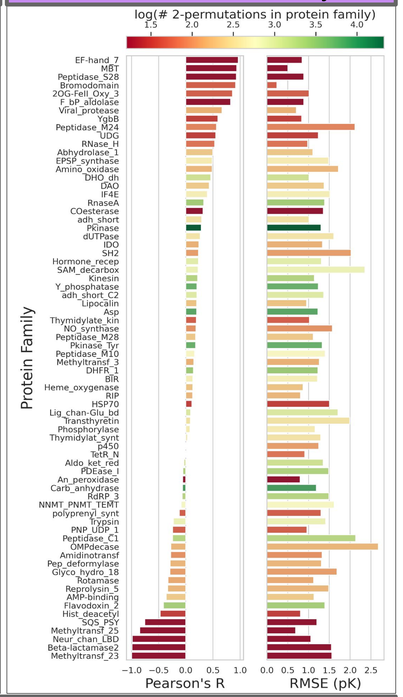


Additional Ligands Comparison



Error bars indicate ±1 standard deviation of 25 individual models (only 5 models for Dense)

Leave One Protein Family Out



Conclusions

- Our models show higher correlation and lower error than previous state-of-the-art models
- Multi-task loss increases the predictive performance of the model
  - ΔG prediction performance enhances $\Delta\Delta G$ prediction performance
  - $\mathcal{L}_{\text{consistency}}$ decreases error of $\Delta\Delta G$ prediction
- Increased regularization of latent space improves $\Delta\Delta G$ prediction
- Latent space subtraction not necessary for performance
- $\Delta\Delta G$ prediction does not generalize well to new protein families

Future Work

- Embed more symmetries of $\Delta\Delta G$ prediction problem into network
  - Equivariant network architectures
  - Cycle consistency
- Build larger, high-quality dataset of congeneric series
  - Redock BindingDB dataset
- Semi-supervised learning for increased generalization performance