Supporting Information

HAC-Net: A Hybrid Attention-Based Convolutional Neural Network for Highly Accurate Protein-Ligand Binding Affinity Prediction

Gregory W. Kyro[1], Rafael I. Brent[1], Victor S. Batista[1]*

[1]Department of Chemistry, Yale University, New Haven, Connecticut 06511-8499

* Email: victor.batista@yale.edu

List of Figures and Tables

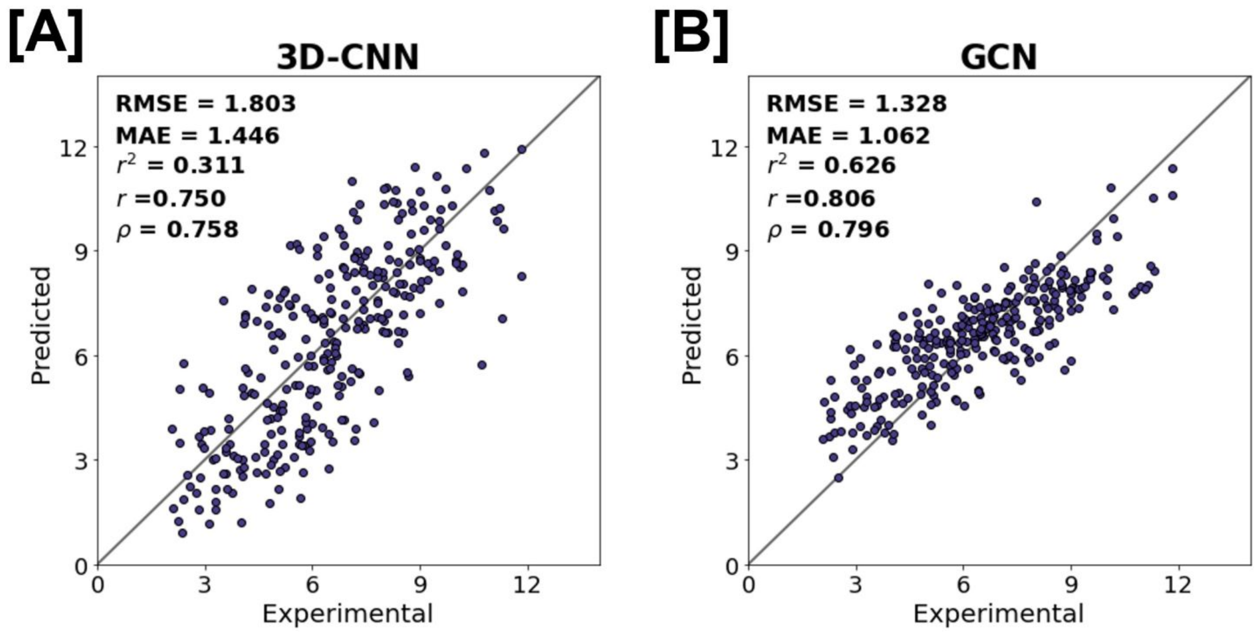

Figure S1. Correlation scatter plots depicting predictions of HAC-Net subcomponents on experimental pK_D values of protein-ligand complexes in the PDBbind v.2016 core set. (A) 3D-CNN and (B) GCN are shown. Root-mean-square error (RMSE), mean absolute error (MAE), r^2, Pearson r, and Spearman ρ are shown on plots.

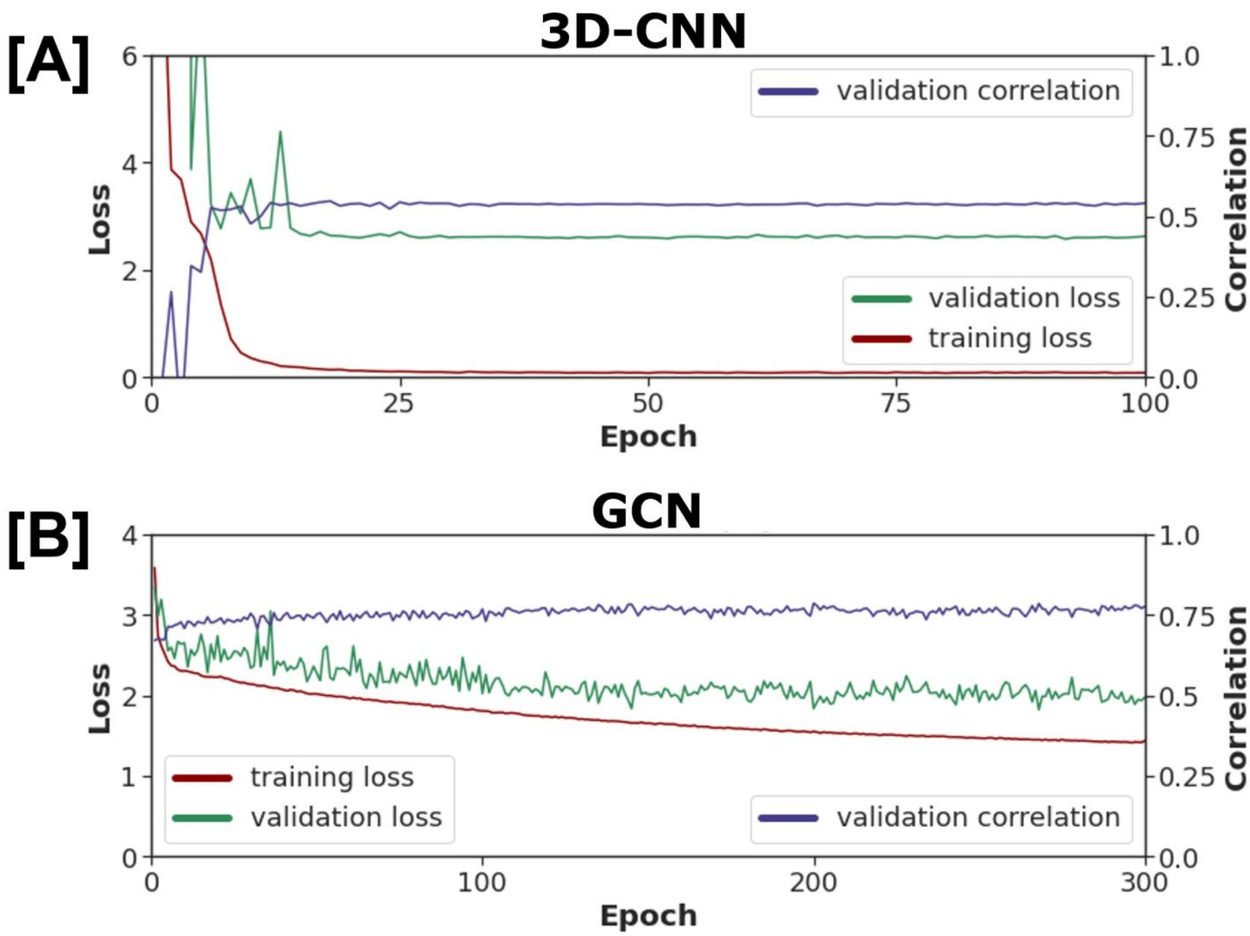

Figure S2. Representative learning curves for testing on the PDBbind v.2016 core set. Validation and training loss (left y-axis) and average correlation ($(\text{Spearman}\,\rho + \text{Pearson}\,\text{r})/2$) on the validation set (right y-axis) are shown as a function of epoch for the (A) 3D-CNN feature extraction and (B) one of the GCNs.

Table S3. Performance of HAC-Net on the Comparative Assessment of Scoring Functions (CASF)-2016 ranking, docking, and screening tests for protein ligand complexes in the CASF-2016 test set.

	Ranking			Docking			Screening					
Model	Spearman ρ	PI	Kendall τ	SR Top 1	SR Top 2	SR Top 3	SR 1% F/R	SR 5% F/R	SR 10% F/R	Mean EF 1%	Mean EF 5%	Mean EF 10%
HAC-Net	0.705	0.731	0.611	0.368	0.572	0.702	0.088/0.042	0.211/0.109	0.386/0.168	2.24	1.91	1.71

[a]We assess ranking power with mean Spearman ρ, predictive index (PI) and Kendall τ across all 57 proteins, and docking power with success rate (SR), where a complex is marked as a success if the root-mean-square deviation (RMSD) of the top 1, 2 and 3 identified ligands is below a preset cutoff of 2.0 Å. To assess screening power, we calculate the SR of identifying the highest-affinity binder among the 1%, 5%, and 10% top-ranked ligands for each target protein in the test set (F: forward) and the SR of identifying the highest-affinity binder among the 1%, 5%, and 10% top-ranked proteins for each target ligand (R: reverse). Additionally, we utilize the mean enhancement factor (EF) among all proteins in the test set. This entire procedure is outlined by Su et al. (*J. Chem. Inf. Model.* 2019, 59, 2, 895–913)

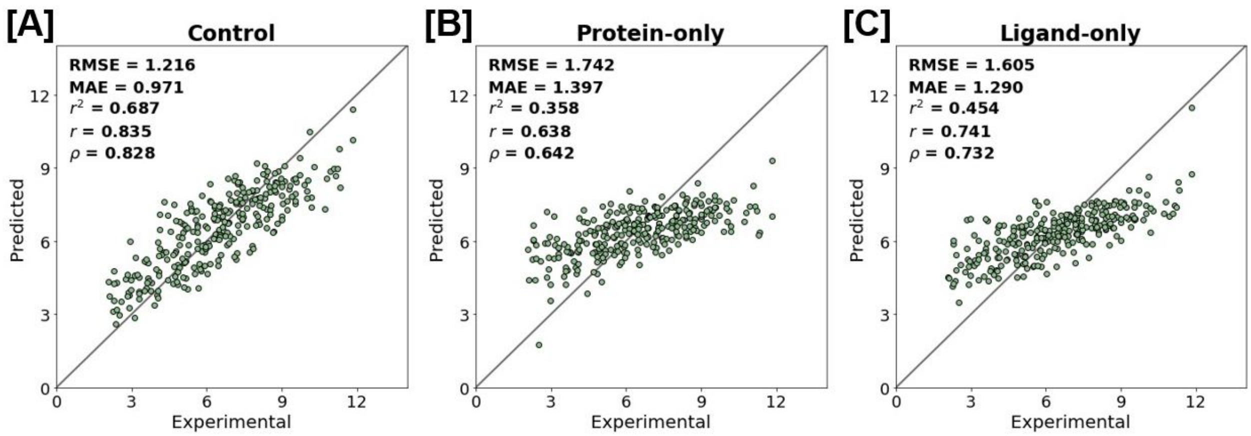

Figure S4. Correlation scatter plots depicting the performance of HAC-Net on the protein-ligand complexes of the PDBbind v.2016 core set compared to protein-only and ligand-only trainings and tests. Root-mean-square error (RMSE), mean absolute error (MAE), r^2, Pearson r, and Spearman ρ are shown. Predictions of experimental pK_D values are shown on the (A) protein-ligand complex data (control), (B) protein-only data, and (C) ligand-only data.

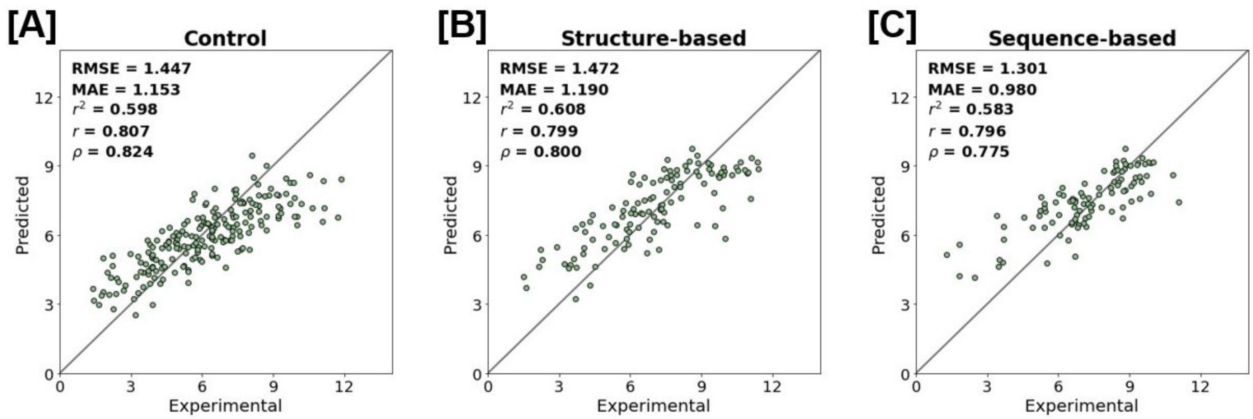

Figure S5. Correlation scatter plots depicting generalizability of HAC-Net across protein structure and sequence. Root-mean-square error (RMSE), mean absolute error (MAE), r^2, Pearson r, and Spearman ρ are shown for predictions of experimental pK_D values for complexes in the A) PDBbind v.2007 core set (Control), (B) test set based on protein structure-dissimilarity (Structure-based), and (C) test set based on protein sequence-dissimilarity (Sequence-based).

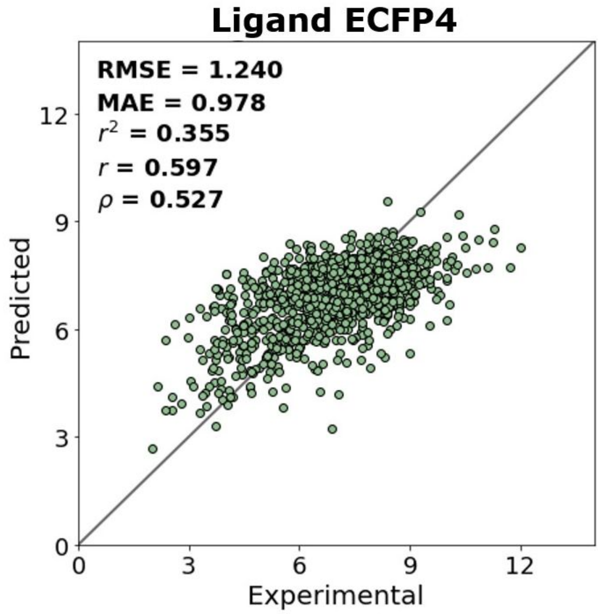

Figure S6. Correlation scatter plot depicting generalizability of HAC-Net based on ligand extended-connectivity fingerprints across four bonds (ECFP4s). Root-mean-square error (RMSE), mean absolute error (MAE), r^2, Pearson r, and Spearman ρ are shown for predictions of experimental pK_D values.

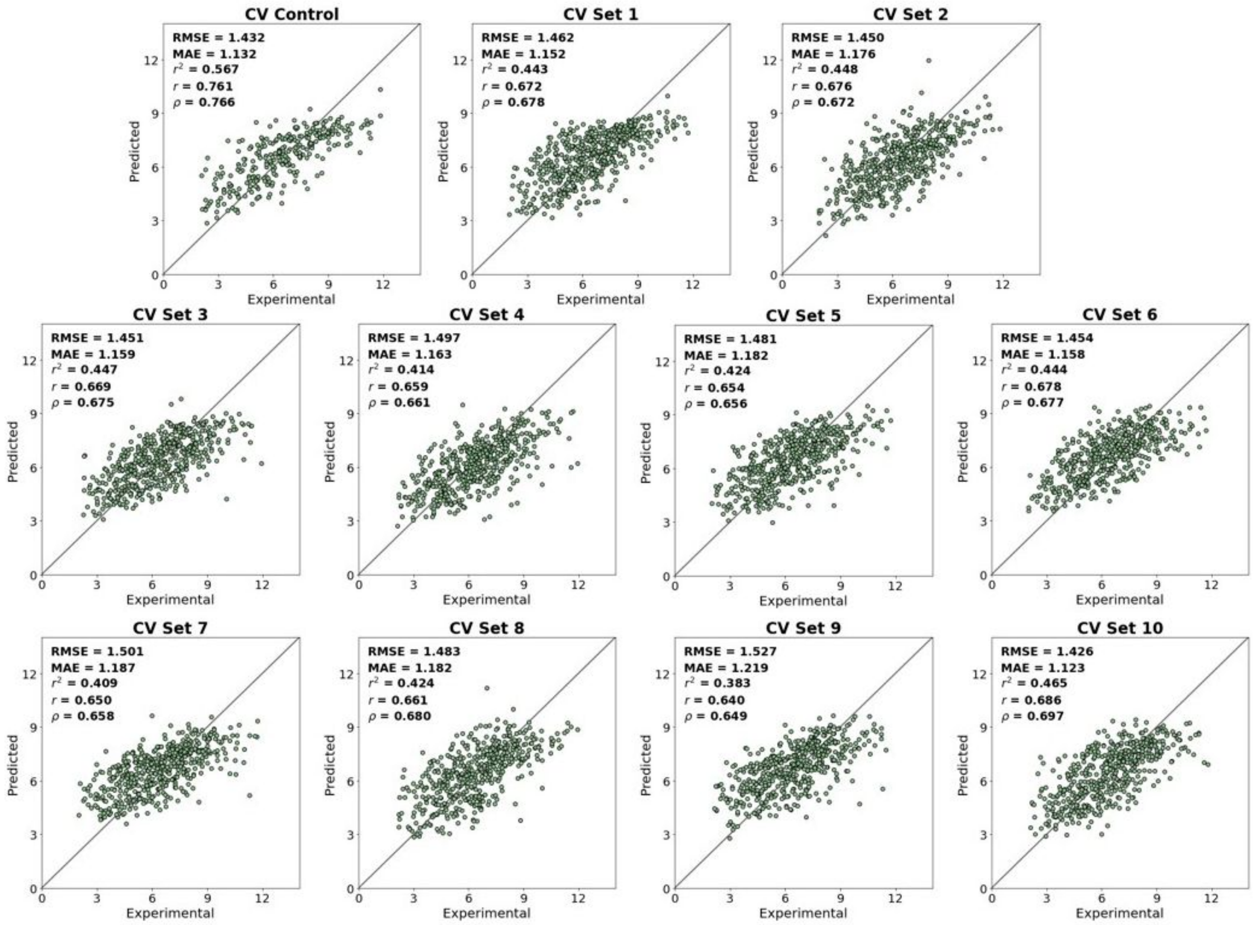

Figure S7. Correlation scatter plots depicting performance of HAC-Net on 10-fold cross-validation based on Tanimoto coefficient (T_c) cutoff applied to ligand SMILES strings. Root-mean-square error (RMSE), mean absolute error (MAE), r^2, Pearson r, and Spearman ρ are shown for predictions of experimental pK_D values on the PDBbind v.2016 core set (CV Control), as well as the ten cross-validation test sets.