Supporting Information

Target-Specific Prediction of Ligand Affinity with Structure-Based Interaction Fingerprints

*Florian Leidner, Nese Kurt Yilmaz, Celia A. Schiffer**

Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA

*Corresponding author: celia.schiffer@umassmed.edu

Figures

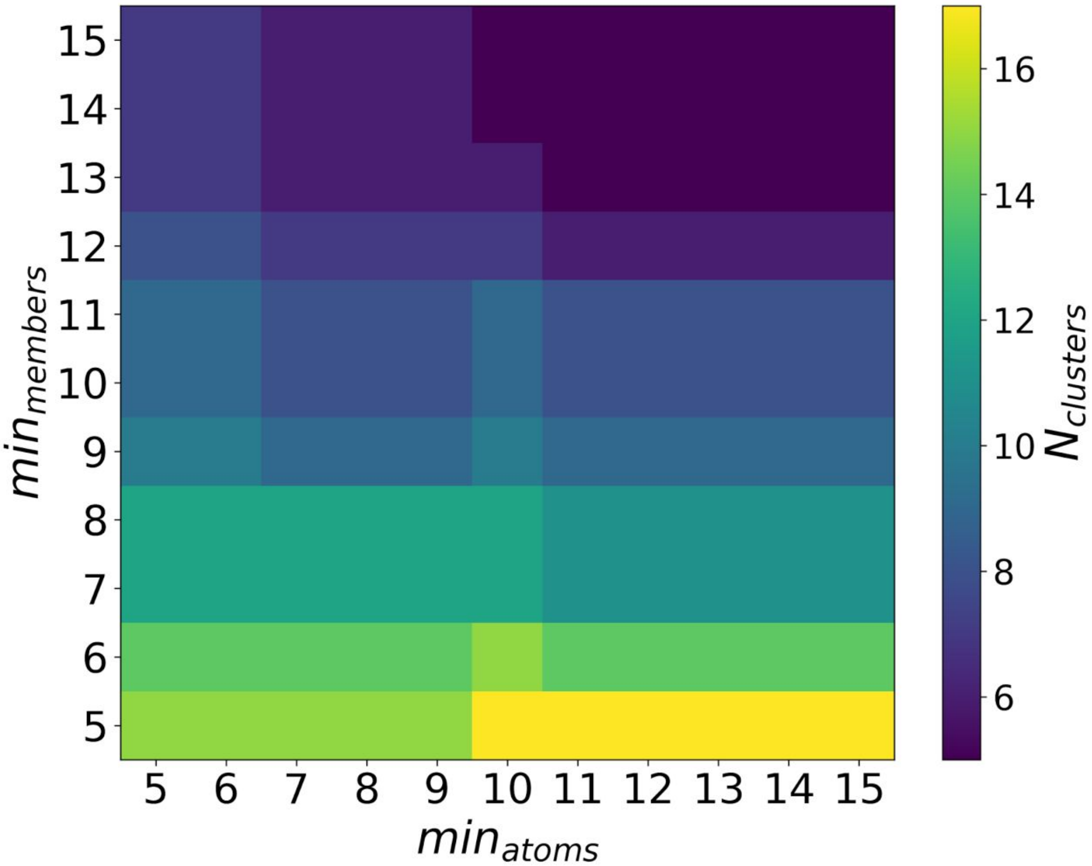

Figure S1 Effect of clustering parameters on number of clusters. Number of clusters shown as a function of atom threshold and minimum cluster size.

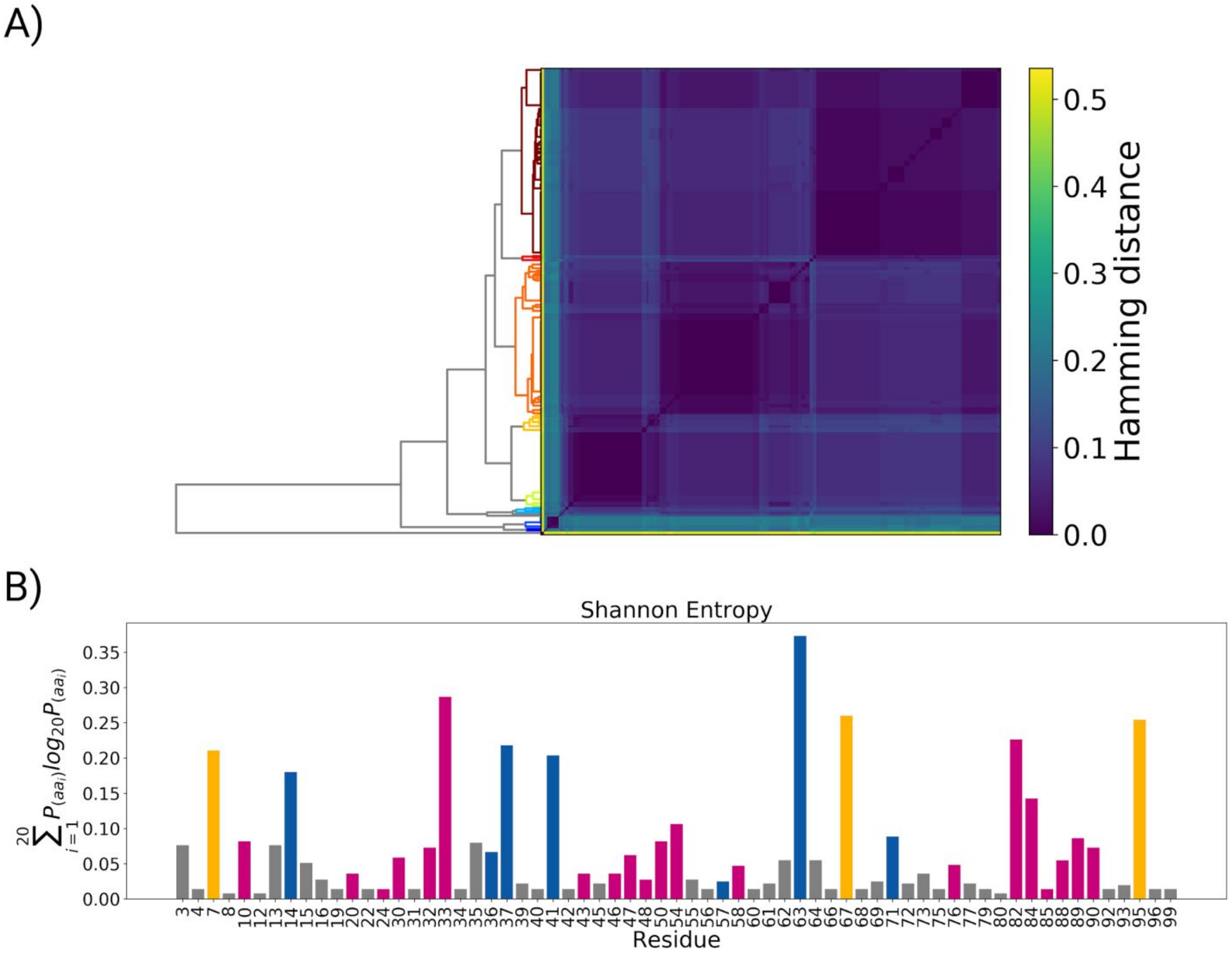

Figure S2 Sequence diversity of the HIV protease dataset. **A)** Hierarchical clustering of HIV protease sequences by hamming distance. **B)** Per residue sequence variation measured by Shannon entropy. Yellow bars indicate common engineered mutations, blue bars indicate natural polymorphism, purple bars indicate drug resistance associated mutations.

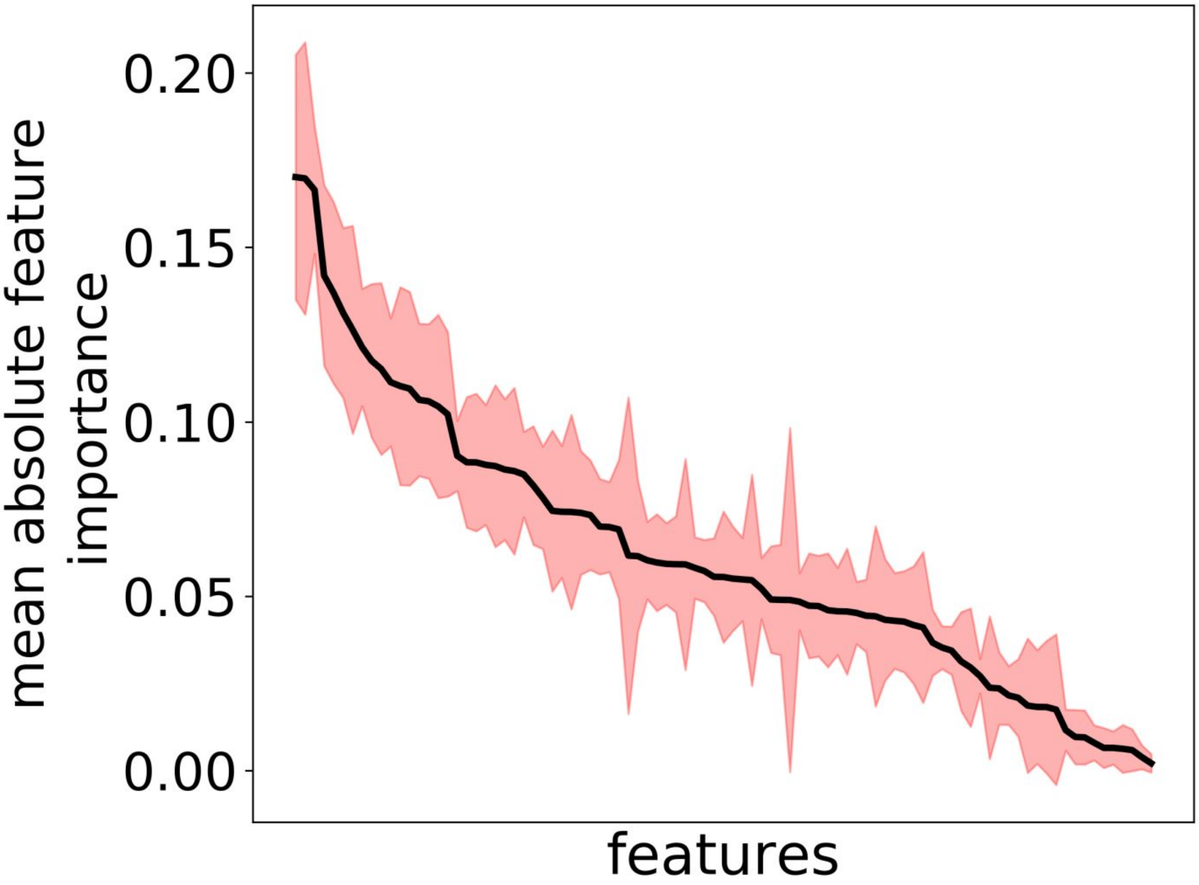

Figure S3 Global feature importance. Rank ordered mean absolute shapley values averaged across 10 gradient boosting models. Red area indicates average feature importance ± standard deviation.

Tables

Table S1. List of hyperparameters for machine learning models used.

Algorithm	Parameter	Optimized Value
Elastic Net	l1_ratio	0.5
	alpha	0.00304
Support Vector Machine Regression	kernel	rbf
	C	29.71
	epsilon	0.27
	gamma	scale[1]
Random Forest	n_estimators	500
	max_features	0.5
	min_impurity_decrease	0.0018
	max_depth	10
Gradient Boosting Machine	learning _rate	0.063
	max_depth	2
	l2_leaf_reg	9
	bagging_temperature	5
	random_strength	73
	iterations	10000
	eval_metric	RMSE

[1] 1/(N*σ^2) where N is the number of features and σ^2 is the feature variance