PremPLI: Predicting the Effects of Missense Mutations on Protein-Ligand Interactions

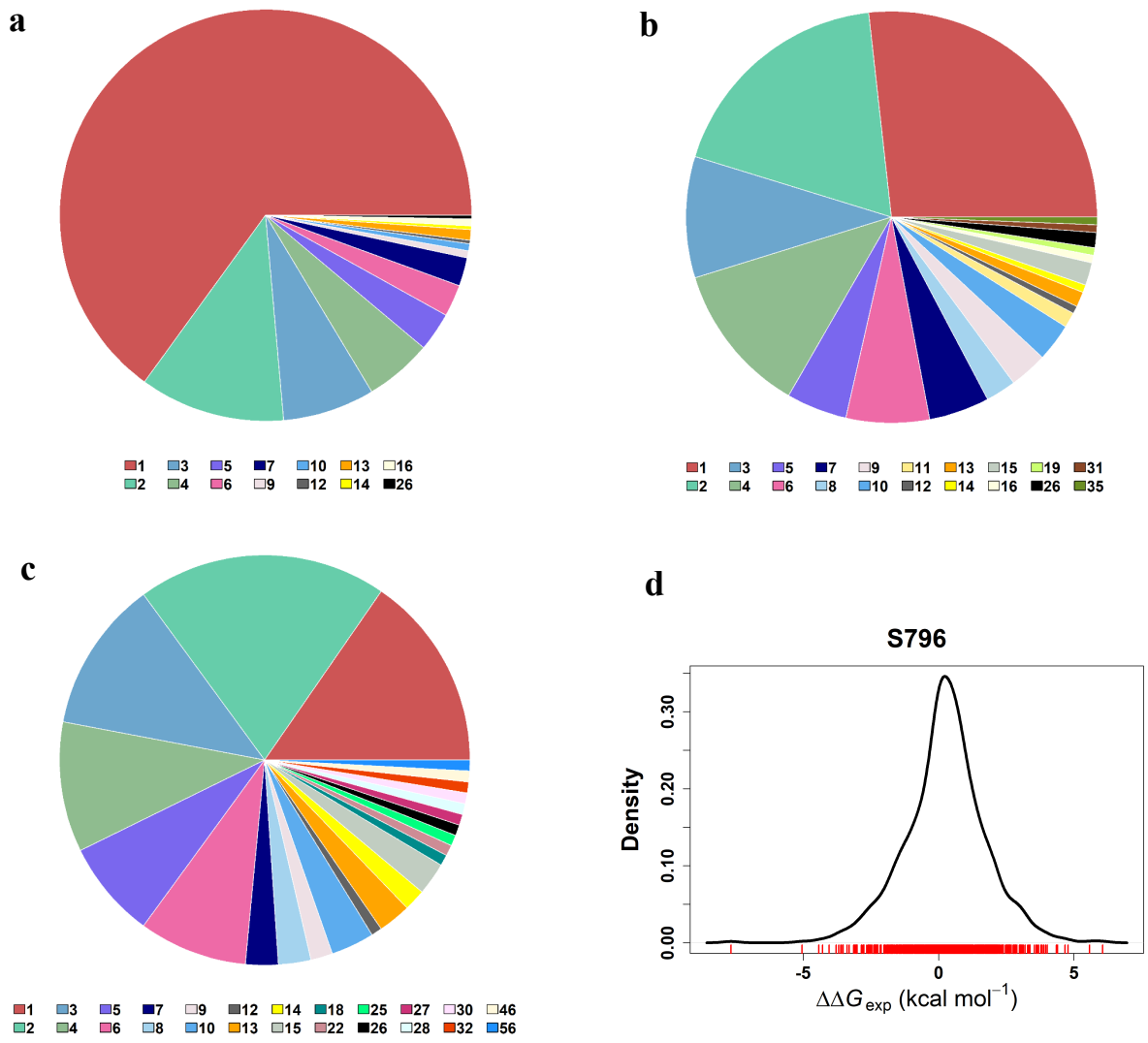


Figure S1. S796 dataset is composed of 796 single mutations from 360 different protein-ligand complex structures, in which the number of unique ligands and proteins is 168 and 117. Distribution of the number of protein-ligand complex structures (a), unique ligands (b) and proteins (c) across different number of mutations. Majority of complexes contain only one single mutation. (d) Distibution of experimental binding affinity changes ($\Delta\Delta G_{exp}$).

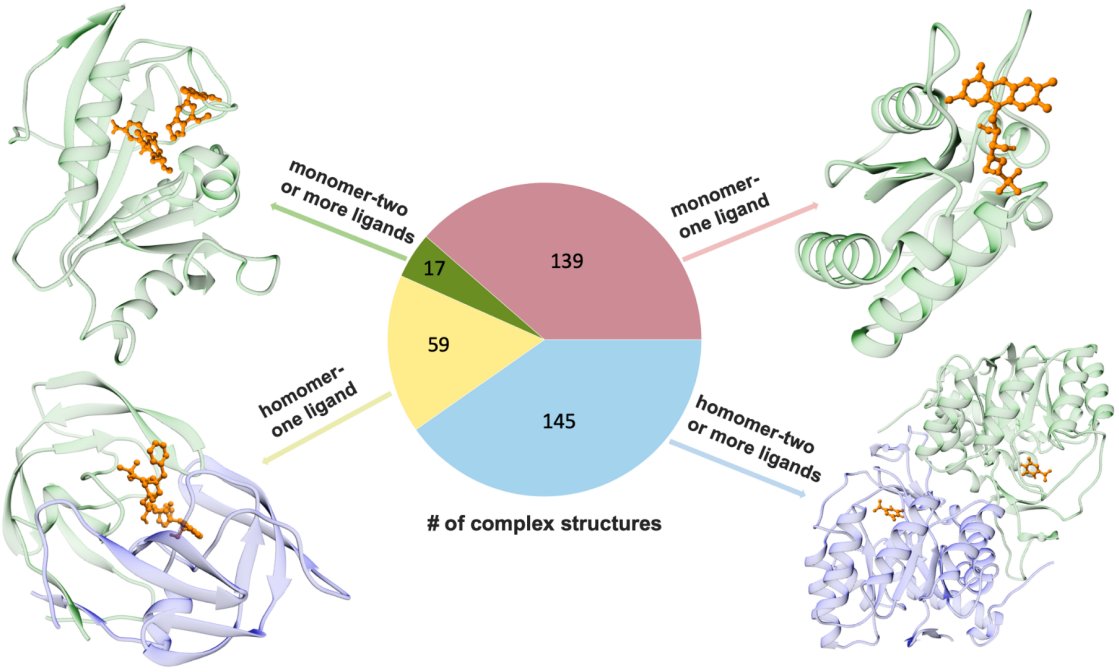


Figure S2. Four types of protein-ligand complex structures in S796 dataset. The number of complex structures for each type is shown in the pie chart.

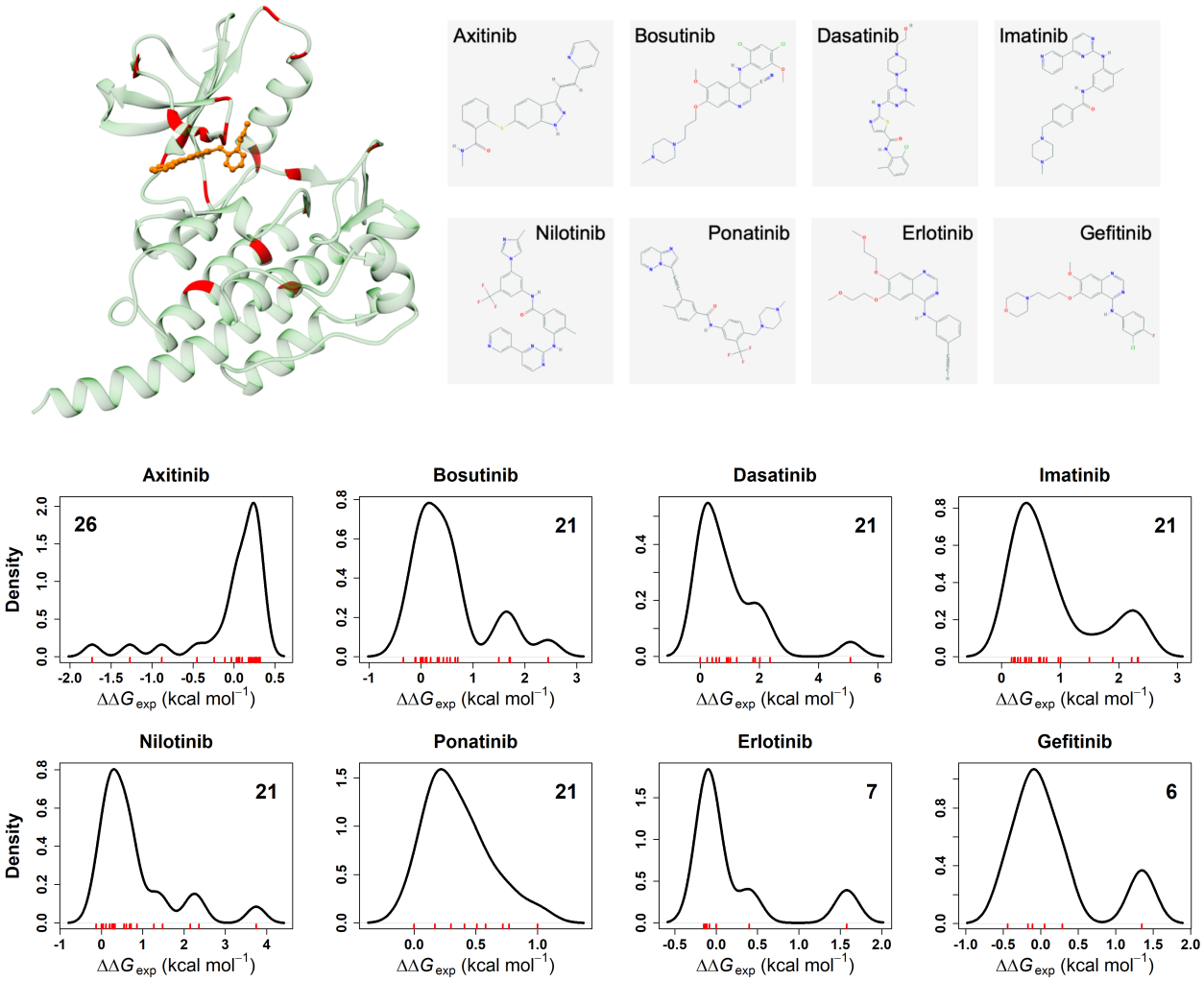
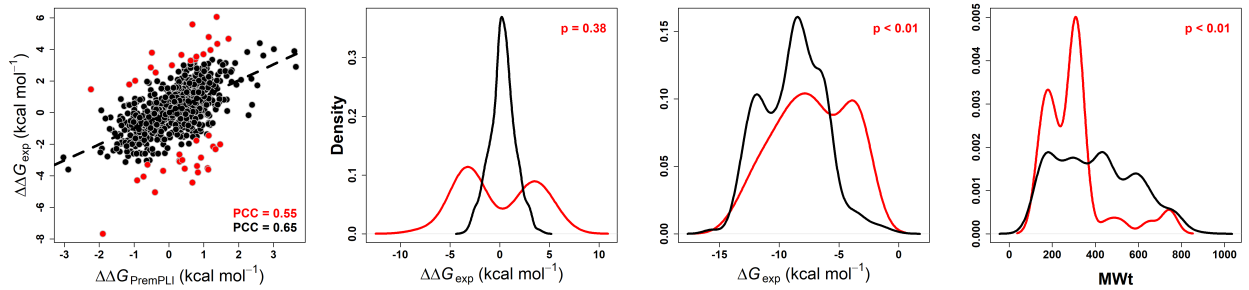
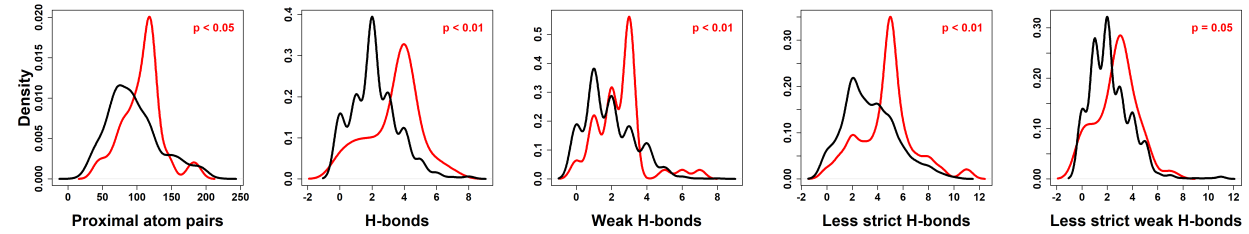


Figure S3. 3D structure of human Abl kinase with axitinib bound (PDB ID: 4WA9, mutation sites are shown in red), names and chemical structures of eight tyrosine kinase inhibitors (TKIs), and the distibution of experimental binding affinity changes ($\Delta\Delta G_{exp}$) for each inhibitor. The number of mutations for each type of inhibitor is shown on the density figure. Co-crystal structures of Abl bound to erlotinib or gefitinib were not available, so the docking models obtained from Hauser et al.[1] were used.

a. **b.**



c.

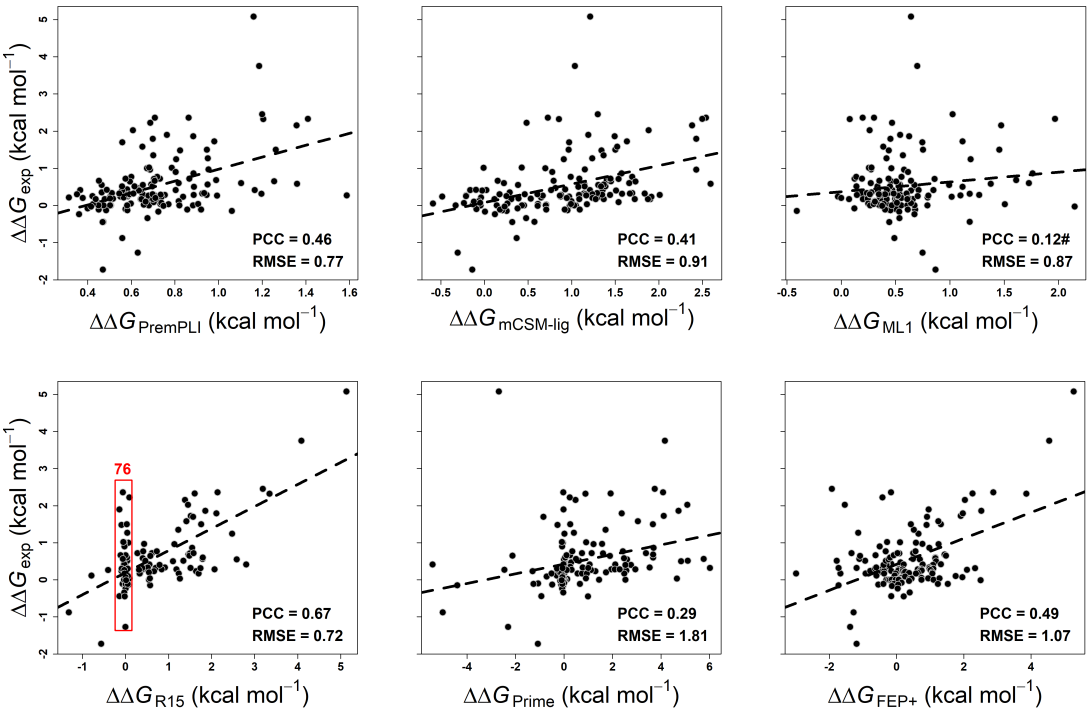


d.

Definition	5% outlier
$(\Delta\Delta G_{PremPLI} > 0$ & $\Delta\Delta G_{exp} > 1.36)/\Delta\Delta G_{exp} > 1.36$	0.67
$(\Delta\Delta G_{PremPLI} < 0$ & $\Delta\Delta G_{exp} < -1.36)/\Delta\Delta G_{exp} < -1.36$	0.27

Figure S4. Outlier analysis of PremPLI. Leave-one-complex-out validation (CV3) results were used for these analyses. Black: all mutations except for 5% outliers; red: 5% outliers. (a) Pearson correlation coefficient between experimental and calculated changes in binding affinity. (b) Distibution of experimental binding affinity change ($\Delta\Delta G_{exp}$), experimental binding affinity (ΔG_{exp}) between wild-type protein and ligand, and molecular weight (MWt) of ligand. p-value (p) is shown indicating whether the difference between 5% outliers and the remaining mutations is statistically significant (t-test). (c) Distibution of the number of proximal atom pairs, hydrogen bonds (H-bonds), weak hydrogen bonds, less strict hydrogen bonds, and less strict weak hydrogen bonds formed between the mutated site and other atoms in the complex, repectively, which were calculated by Arpeggio[2]. (d) True positive rate.

a. S144



b. S99

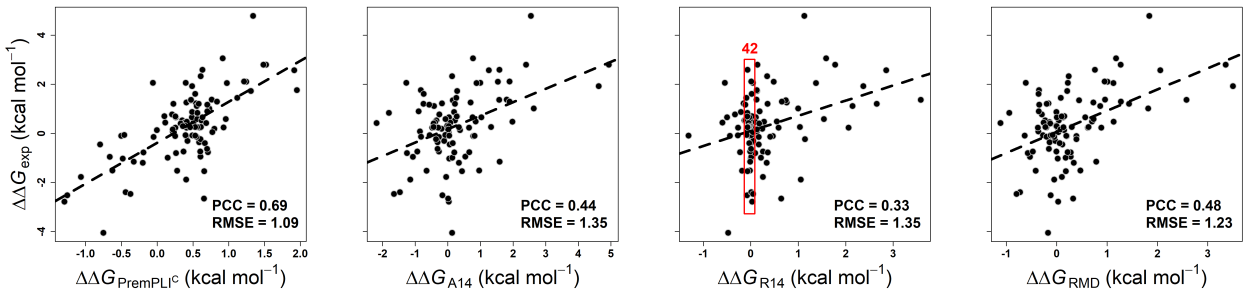


Figure S5. Pearson correlation coefficients (PCC) between experimental and calculated binding affinity changes for different methods tested on S144 (a) and S99 (b) datasets. All correlation coefficients are statistically significantly different from zero (p-value < 0.01, t-test) except $^{\#}$p-value = 0.14 for ML1. The prediction values of Rosetta for half of mutations in S144 (R15, 76 mutations in red box) and S99 (R14, 42 mutations in red box) are around zero.

a. PCC

b. RMSE

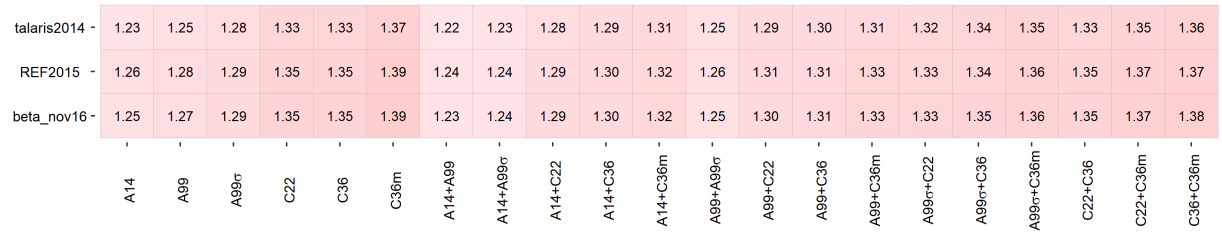
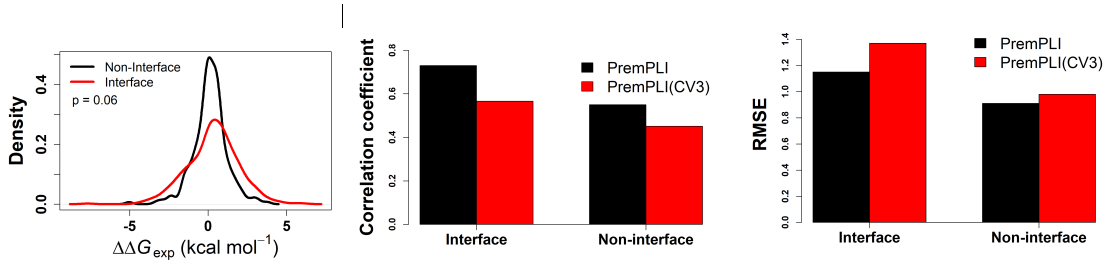
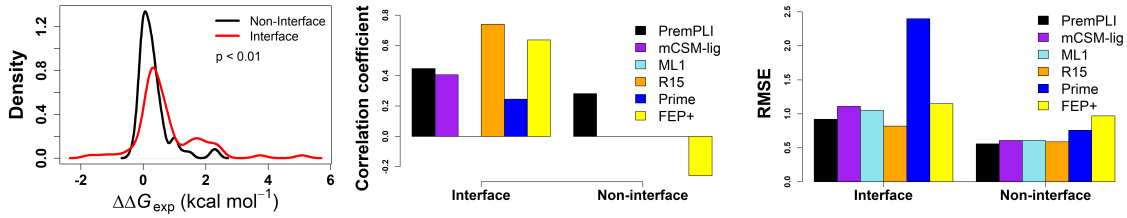


Figure S6. Performance of combined Rosetta using three different scoring functions and MD approaches under six different force fields tested on S99 dataset. The PCC and RMSE were calculated using the predicted values taken from Aldeghi et al.[3] and the average values from Rosetta and MD approaches were used.

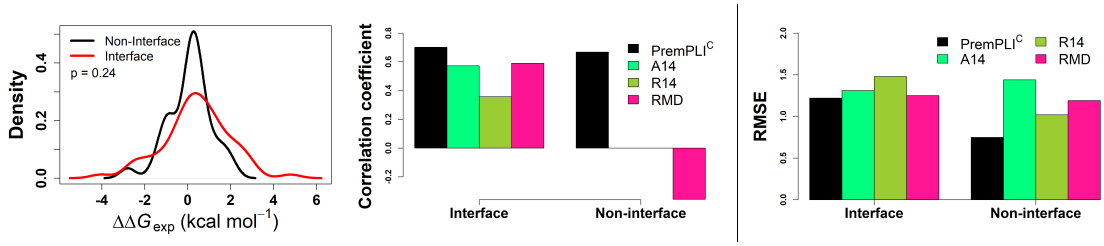
a. S796



b. S144



c. S99



d.

Category	S796	S144	S99
Interface	545	75	67
Non-interface	251	69	32

Figure S7. Distribution of experimental binding affinity changes, Pearson correlation coefficients, and root-mean-square errors for interface and non-interface mutations from S796 (a), S144 (b) and S99 (c) datasets, respectively. Only statistically significant PCC values were shown in the figure. (d) The number of interface and non-interface mutations. Interface residues were defined if any heavy atoms are within 5Å distance from any heavy atoms of ligand.

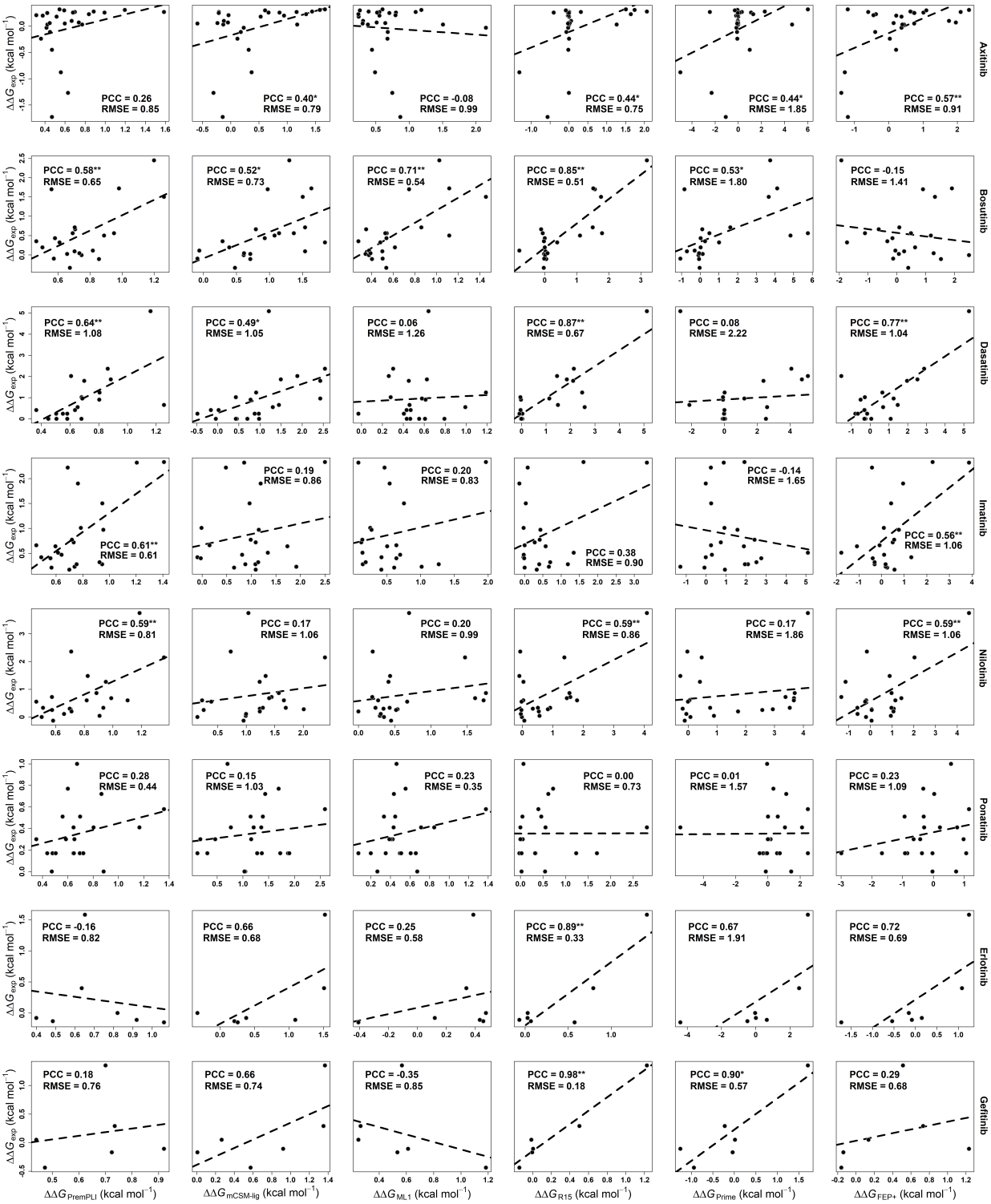


Figure S8. Performance for six different methods tested on eight Abl-inhibitor complexes. * and ** indicate statistically significant difference from zero in terms of PCC with p-value < 0.05 and p-value < 0.01, respectively.

Table S1. Experimental datasets used for training and testing different methods. S763 is the training set of mCSM-lig and S675 is the overlapped data between S796 and S763.

Dataset	# of mutations	# of complexes	# of ligands	# of proteins
S796	796	360	168	117
S859	859	360	168	117
S144	144	8	8	1
S99	99	42	22	14
S763	763	254	177	125
S675	675	240	166	116

Table S2. The importance of each feature in random forest scoring function of PremPLI. IncNodePurity is used for describing the importance which is the total decrease in node impurities from splitting on the variable, averaged over all trees.

Feature	Importance
$\boldsymbol{\Delta CS}$	371
$\boldsymbol{MW}$t	164
$\boldsymbol{Prox^{wt}}$	148
$\boldsymbol{PSSM}$	142
$\boldsymbol{P_{RKDE}}$	137
$\boldsymbol{P_{Q}}$	134
$\boldsymbol{N_{NO}}$	132
$\boldsymbol{H}$	125
$\boldsymbol{M_{AA1}}$	113
$\boldsymbol{Prox^{mut}}$	112
$\boldsymbol{M_{AA2}}$	96

Table S3. The performance for PremPLI trained and tested on S859 and S796 dataset and performing four types of cross-validation, respectively.

Method	S859			S796		
	PCC	RMSE	Slope	PCC	RMSE	Slope
PremPLI	0.74	1.06	1.12	0.70	1.08	1.14
PremPLI (CV1)	0.73	1.07	1.14	0.69	1.10	1.17
PremPLI (CV2)	0.74	1.06	1.14	0.70	1.08	1.16
PremPLI (CV3)	0.51	1.34	0.96	0.55	1.26	1.01
PremPLI (CV4)	0.50	1.35	0.92	0.53	1.27	0.98

Pearson correlation coefficient between experimental and predicted $\Delta\Delta G$ values. RMSE (kcal mol^{-1}): root-mean-square error. Slope: the slope of the regression line between experimental and predicted $\Delta\Delta G$ values. All correlation coefficients are statistically significantly different from zero (p-value < 0.01, t-test).

Table S4. The performance using Random Forest (RF), Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost) algorithms to build PremPLI model, respectively. Leave-one-complex-out results on S796 are shown.

Algorithm	PCC	RMSE	Slope
RF	0.55	1.26	1.01
SVM	0.43*	1.36	0.89
XGBoost	0.56	1.25	0.85

*p-value < 0.01 compared to Random Forest (Hittner2003 test).

Table S5. Computational cost for different approaches. The data for ML1, FEP+, and R15 were taken from Aldeghi et al.[4].

Abbreviation	Method	Force filed or scoring function	Approximate cost per $\Delta\Delta G$ calculation	
			Hardware	Compute hours
PremPLI	Machine Learning	n/a	1 CPU core	0.17
ML1	Machine Learning	n/a	1 CPU core	0.02
FEP+	Molecular Dynamics	OPLS3	1 GPU	72
R15	Rosetta	REF15	1 CPU core	32

References

(1) Hauser, K.; Negron, C.; Albanese, S. K.; Ray, S.; Steinbrecher, T.; Abel, R.; Chodera, J. D.; Wang, L. Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Communications biology* **2018,** *1*, 70.

(2) Jubb, H. C.; Higueruelo, A. P.; Ochoa-Montano, B.; Pitt, W. R.; Ascher, D. B.; Blundell, T. L. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of molecular biology* **2017,** *429* (3), 365.

(3) Aldeghi, M.; Gapsys, V.; de Groot, B. L. Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation. *ACS central science* **2018,** *4* (12), 1708.

(4) Aldeghi, M.; Gapsys, V.; de Groot, B. L. Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches. *ACS central science* **2019,** *5* (8), 1468.