SUPPLEMENTARY DATA

mCSM-AB2: Guiding Rational Antibody Design Using Graph-Based Signatures

Yoochan Myung[1,2,3], Carlos H.M. Rodrigues[1,2,3], David B. Ascher[1,2,3,4,*], Douglas E.V. Pires[1,2,3,5,*]

[1]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria 3010, Australia

[2]ACRF Facility for Innovative Cancer Drug Discovery, Bio21 Institute, University of Melbourne, Melbourne, Victoria 3010, Australia

[3]Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute, Melbourne, Victoria 3004, Australia

[4]Department of Biochemistry, University of Cambridge, CB2 1GA, UK

[5]School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria 3010, Australia

Table S1. Mutations from same Ab-antigen complexes in mCSM-AB2 dataset.

PDB ID	# of forward mutations	
	AB-BIND	**Experimental blind-test**
1BJ1	19	10
1CZ8	19	1
1JRH	2	41
1N8Z	28	6
1VFB	38	8
1YY9	5	11
2JEL	43	1
2NYY	28	3
3BN9	34	1
3HFM	21	52
3NGB	11	30

Table S2. Predictive performance of individual feature classes in mCSM-AB2 regression model.

Class	Feature	Pearson's correlation (RMSE (Kcal/mol))		
		10-fold CV	**5-fold CV**	**Jackknife**
Structure-based	**Δ** Relative solvent accessibility	0.16 (10.92)	0.16 (11.36)	0.16 (11.05)
Energy-based terms	FoldX score	0.26 (6.61)	0.24 (7.41)	0.29 (5.89)
Structure-based	**Δ** Distance to interface[a]	0.26 (6.64)	0.25 (6.99)	0.27 (6.43)
Structure-based	Distance pattern	0.20 (8.94)	0.17 (10.42)	0.20 (9.06)
Sequence-based	PSSM score	0.42 (3.95)	0.42 (3.95)	0.41 (3.97)
Structure-based	**Δ** Pharmacophore count	0.50 (3.12)	0.49 (3.19)	0.50 (3.11)
Structure-based	**Δ** Arpeggio contacts[b]	0.60 (2.44)	0.59 (2.45)	0.60 (2.42)
Structure-based	Graph-based signatures	0.65 (2.14)	0.64 (2.19)	0.65 (2.09)
All	**mCSM-AB2**	**0.73 (1.68)**	**0.72 (1.72)**	**0.73 (1.70)**

[a] The closest distance between a mutation site and its binding partner.

[b] Interatomic interactions.

Table S3. Performance comparison among 8 different algorithms.

Algorithm	**Pearson**	**RMSE (Kcal/mol)**	**Pearson (90%)**	**RMSE (90%) (Kcal/mol)**
Extra Trees	**0.73**	**1.68**	**0.84**	**1.14**
Random forest	0.70	1.85	0.83	1.28
XGBOOST	0.68	1.93	0.78	1.32
Gradient boost	0.68	1.96	0.77	1.31
Adaboost	0.61	2.37	0.73	1.88
Support vector machine	0.13	13.92	0.11	2.71
K-nearest neighbors	0.02	105.47	0.18	60.11
Gaussian regression	0.02	91.18	0.01	3.01

Table S4. Performance comparison between mCSM-AB2 models, before and after including features derived from modelled mutant structures.

Method	Pearson's Correlation (RMSE (Kcal/mol))		
	Training	Blind Test	
	AB-BIND	Experimental Structures	Homology Models
Without mutant features	0.75 (RMSE = 1.58)	0.62 (RMSE = 1.92)	0.72 (RMSE = 1.81)
With mutant features	0.76 (RMSE = 1.51)	0.64 (RMSE = 1.85)	0.77 (RMSE = 1.66)

Table S5. RMSD between wild-type crystal and modelled structures.

Wild-type	**Mutant model**	**RMSD (Å) on single position**	**RMSD (Å) on all atoms**	**RMSD (Å) on Cα**
Tyr32 of 1KIQ chain B	Ala32Tyr of 1KIPchain B	0.87	0.17	0.14
Tyr101 of 1KIP chain B	Phe101Tyr of 1KIQ chain B	0.87	0.17	0.14
Ser50 of 1KIP chain A	Tyr50Ser of 1KIR chain A	0.35	0.15	0.13
Phe33 of 1XGU chain B	Ala33Phe of 1XGP chain B	0.53	0.19	0.17
Val33 of 1XGQ chain B	Ala33Val of 1XGP chain B	1.01	0.16	0.13
Leu33 of 1XGT chain B	Ala33Leu of 1XGP chain B	0.44	0.15	0.13
Ile33 of 1XGR chain B	Ala33Ile of 1XGP chain B	0.27	0.14	0.12
Average		0.62	0.16	0.13
Standard deviation		0.29	0.02	0.02

Table S6. Performance of mCSM-AB2 stratified by wild-type and mutant amino acids residue types, considering forward and hypothetical reverse mutations.

10-fold cross validation results using the complete set of mutations (1810 mutations)			
Class (Wild-type to Mutant)	**Pearson's correlation** RMSE (Kcal/mol) (Number of mutations)		
	Forward & hypothetical reverse mutations	**Forward mutation**	**Reverse mutation**
any to any	**0.73** **RMSE = 1.68** **(1810)**	**0.62** **RMSE = 1.70** **(905)**	**0.59** **RMSE = 1.66** **(905)**
non-ALA to ALA	0.73 RMSE = 1.58 (569)	0.72 RMSE = 1.58 (554)	0.83 RMSE = 1.38 (15)
ALA to non-ALA	0.67 RMSE = 1.57 (569)	0.91 RMSE = 1.29 (15)	0.66 RMSE = 1.58 (554)
non-ALA to non-ALA	0.55 RMSE = 1.85 (672)	0.46 RMSE = 1.90 (336)	0.48 RMSE = 1.81 (336)

Table S7. mCSM-AB2 performance on leave-one-complex-out cross-validation. Mutations were grouped by complex in a total of 60 groups. Each group was used as test set and the remaining used as training set. mCSM-AB2 on leave-one-complex-out cross-validation achieved a correlation of ρ = 0.70.

Group #	PDB ID	RMSE (Kcal/mol)	# of mutation in complex	# of outliers in training	# of outliers in 10-fold CV
1	1AHW	1.61	18	2	2
2	1AK4	1.01	32	0	0
3	1AO7	1.8	10	1	2
4	1BJ1	1.75	58	3	3
5	1CZ8	2.42	40	7	7
6	1DQJ	2.36	42	7	10
7	1DVF	1.96	50	9	7
8	1FC2	1.94	18	3	4
9	1FCC	1.06	14	0	0
10	1FFW	0.65	18	0	0
11	1JRH	2.08	86	11	13
12	1JTG	1.88	10	0	0
13	1KIP	2.05	2	0	0
14	1KIQ	0.45	2	0	0
15	1KIR	0.7	2	0	0
16	1KTZ	1.09	44	0	2
17	1MHP	2.42	100	22	11

18	1MLC	2.4	22	4	4
19	1N8Z	1.33	68	2	2
20	1NCA	1.66	8	0	0
21	1NMB	1.82	16	3	3
22	1VFB	2.09	92	15	13
23	1XGP	1.63	8	0	0
24	1XGQ	1.4	8	0	0
25	1XGR	1.99	8	0	0
26	1XGT	1.78	8	0	0
27	1XGU	4.5	8	7	6
28	1YQV	1.57	2	0	0
29	1YY9	1.45	32	2	4
30	2B2X	0.47	6	0	0
31	2BDN	1.29	24	0	3
32	2JEL	1.7	88	8	10
33	2NYY	1.85	62	8	4
34	2NZ9	2.41	38	6	3
35	2VIR	3.64	4	3	2
36	2VIS	3.96	2	2	1
37	3BDY	1.67	68	8	5
38	3BE1	1.38	68	4	4
39	3BN9	1.5	70	4	4

40	3G6D	4.03	4	2	2
41	3HFM	2.19	146	17	31
42	3K2M	1.09	14	0	1
43	3L5X	2.58	2	0	1
44	3LZF	0.37	4	0	0
45	3N85	2.09	18	2	3
46	3NGB	1.28	82	0	0
47	3NPS	1.17	54	2	0
48	3SE8	1.56	56	2	2
49	3SE9	1.41	50	3	1
50	3W2D	2.45	8	2	1
51	4GXU	2.53	4	1	1
52	4I77	2.58	24	6	5
53	4JPK	1.72	16	1	1
54	4KRL	2.19	4	0	0
55	4KRO	0.9	4	0	0
56	4KRP	1.81	2	0	0
57	4NM8	1.82	18	2	3
58	4U6H	1.78	4	0	0
59	4ZS6	1.45	4	0	0
60	5C6T	1.33	36	0	0

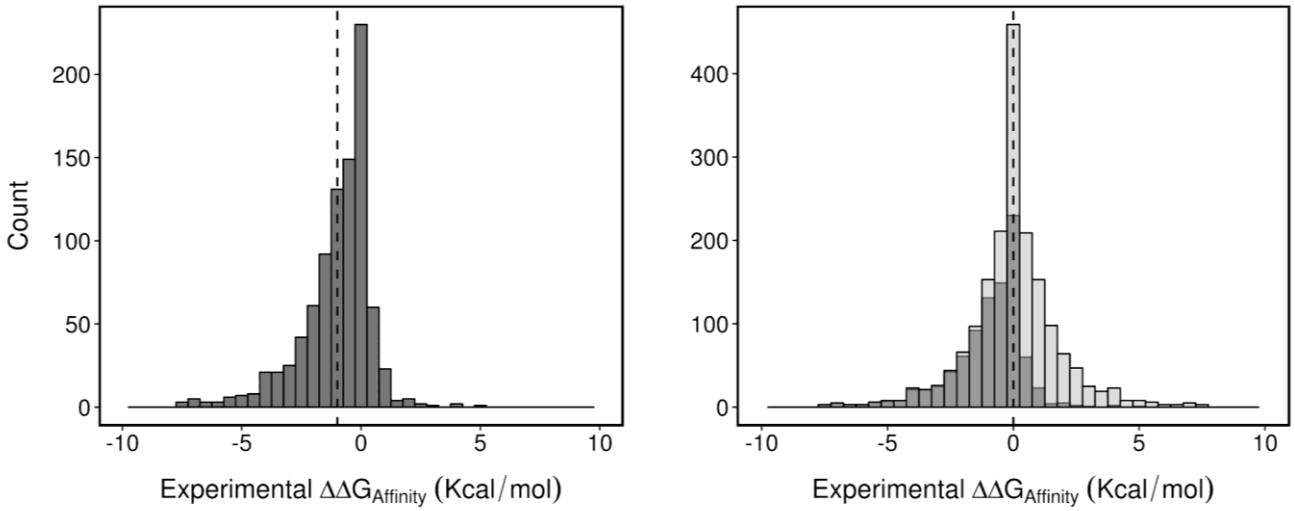

Figure S1. Distribution of experimental $\Delta\Delta G_{Affinity}$ for the mCSM-AB2 data set. The distribution of the experimental $\Delta\Delta G_{Affinity}$ for the 905 missense mutations is shown on the left, with an average of -1.00 Kcal/mol (mutation leading to a reduced affinity- dashed line). The graph on the right depicts the distribution of $\Delta\Delta G_{Affinity}$ for the data set after 905 hypothetical reverse mutations are included, showing a balanced distribution (average of 0 Kcal/mol).

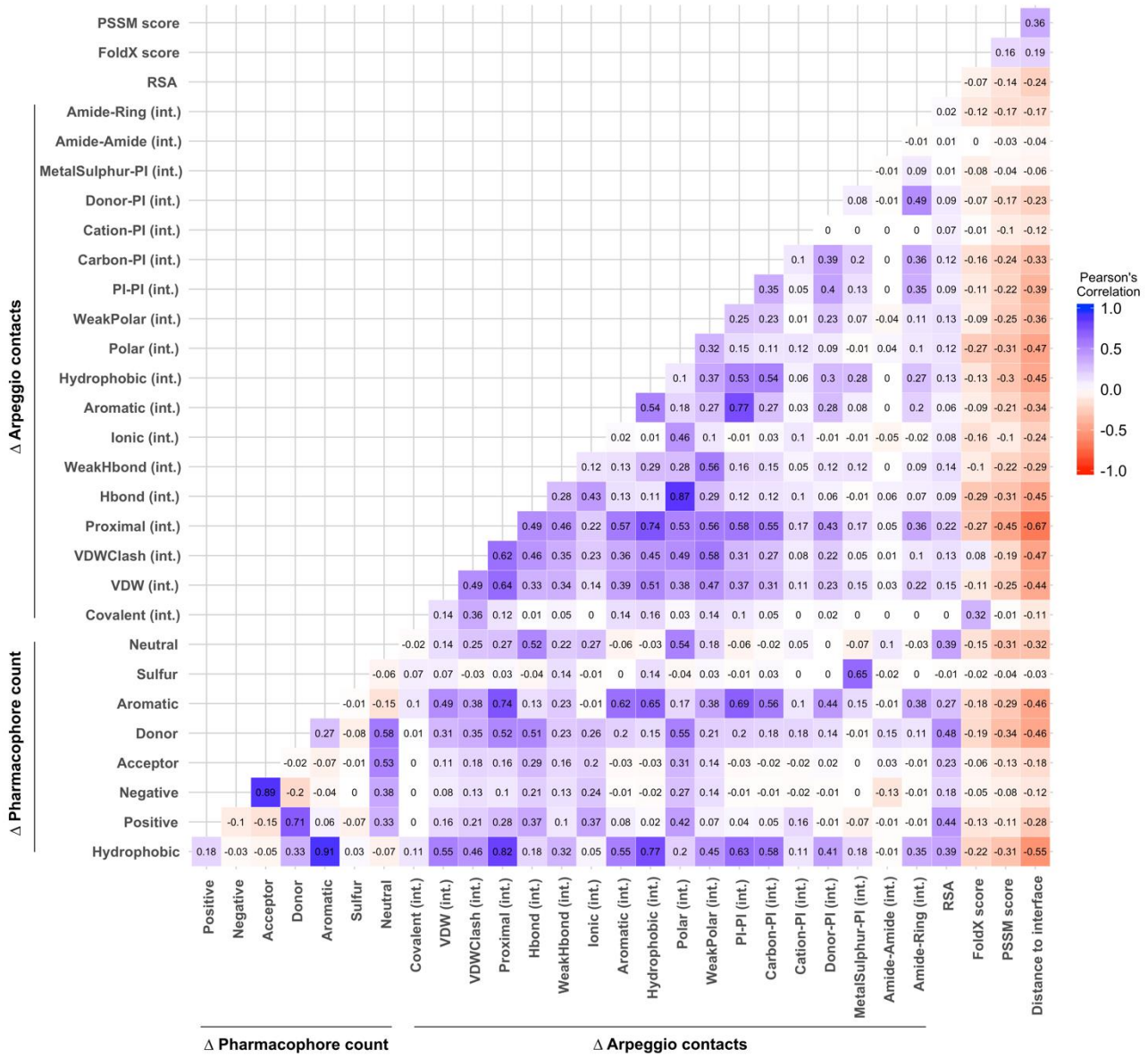

Figure S2. Pairwise associations among features used for mCSM-AB2. The heatmap is coloured by Pearson's correlation coefficient showing how features used in mCSM-AB2 are related to each other highlighting that, in general, there is little correlation between features, meaning there are no redundant feature pairs.

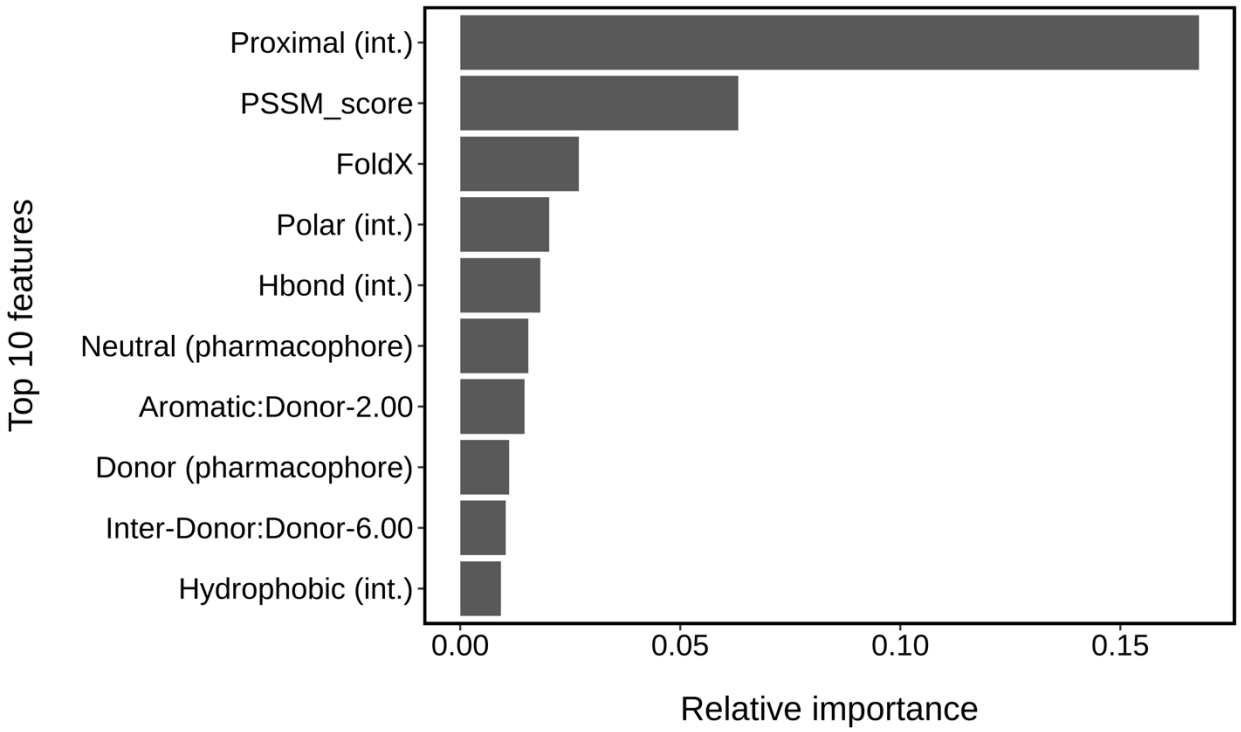

Figure S3. Relative importance of features. The relative importance based on Extra Trees algorithm shows the importance of individual features on the mCSM-AB2 model. The features with "(int.)" are changes of interatomic interaction between wild-type and mutant. The "Aromatic:Donor" and "Inter-Donor:Donor" are distance patterns which represents the geometry of surrounding environment of mutation site in wild-type at atomic level.

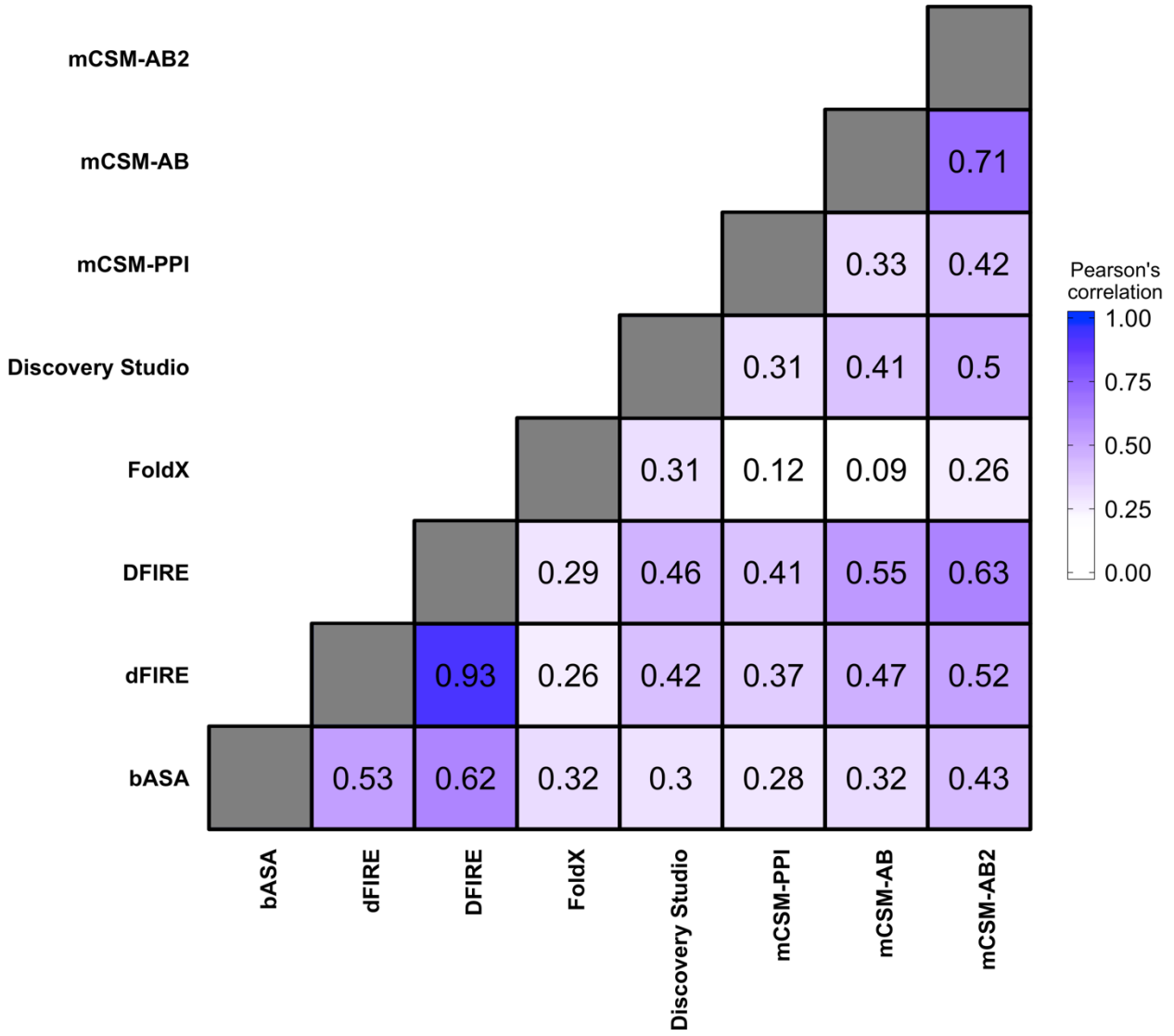

Figure S4. Pairwise correlation analysis of individual tools. The heatmap shows correlation between experimental $\Delta\Delta G_{Affinity}$ and predicted $\Delta\Delta G_{Affinity}$ of each method on the 754 non-redundant mutation data (blind test).

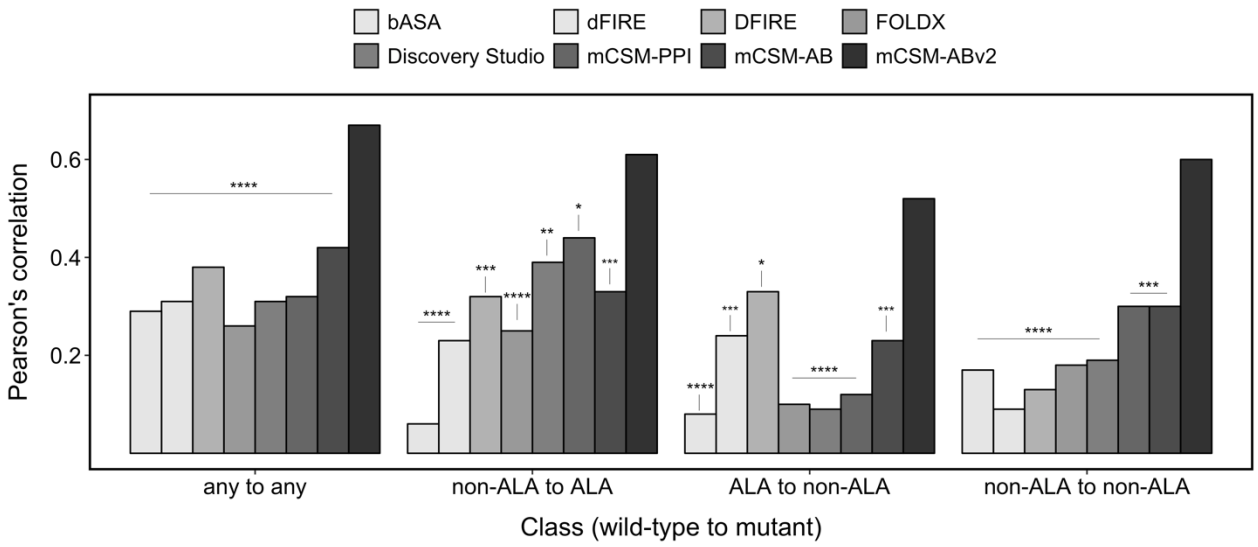

Figure S5. Performance comparison of individual tools considering different residue types. The individual methods were evaluated on the experimental blind test, and AB-BIND was used for building mCSM-AB and mCSM-AB2 predictive models. The *any to any* set (754 mutations) consists of non-ALA to ALA (216 mutations), ALA to non-ALA (216 mutations) and non-ALA to non-ALA (322 mutations) classes. Statistical significance of individual method was compared to mCSM-AB2 using Fisher's r-to-z transformation (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, and **** $p \leq 0.0001$).