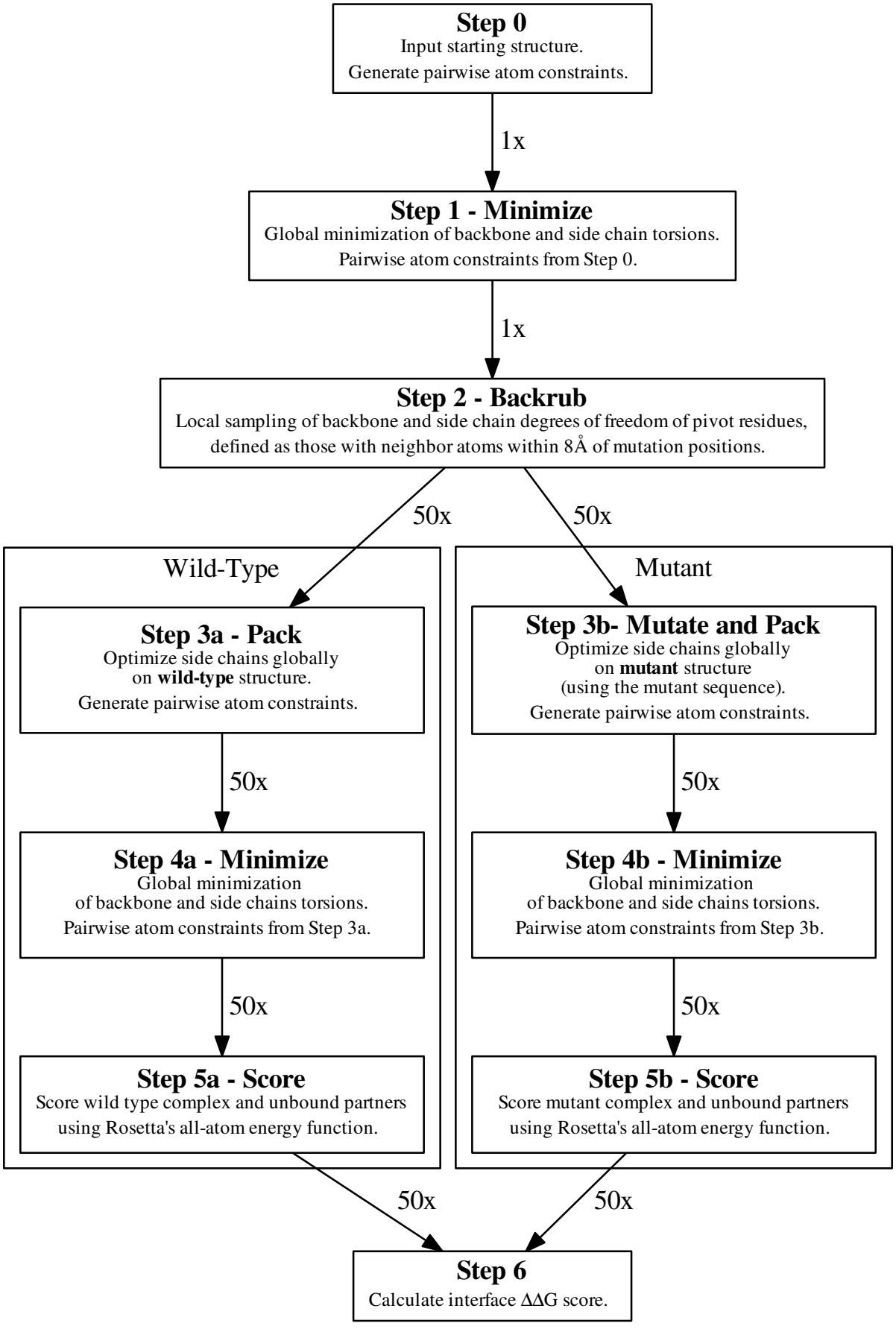
n	Name
1240	Complete dataset
748	Single mutation to alanine
273	Multiple mutations
130	Small-to-large mutation(s)
45	Multiple mutations, none to alanine

Table 1: ZEMu dataset subset definition and composition

Mutation Category	Prediction Method	N	R	MAE	FC
Complete dataset	flex ddG	1240	**0.63**	**0.97**	**0.76**
	ddG monomer		0.51	1.57	0.64
	no backrub control		0.56	1.12	0.73
	ZEMu paper		0.61	1.08	0.71
Small-to-large mutation(s)	flex ddG	130	**0.64**	**0.80**	**0.71**
	ddG monomer		0.31	1.55	0.55
	no backrub control		0.41	1.11	0.62
	ZEMu paper		0.48	1.16	0.65
Mutation(s) to alanine	flex ddG	939	0.61	**0.97**	**0.78**
	ddG monomer		0.50	1.55	0.66
	no backrub control		0.58	1.06	0.75
	ZEMu paper		**0.62**	1.03	0.73
Single mutation to alanine	flex ddG	748	**0.50**	**0.76**	**0.77**
	ddG monomer		0.36	1.31	0.62
	no backrub control		0.44	0.90	0.74
	ZEMu paper		0.45	0.86	0.71
Multiple mutations	flex ddG	273	0.62	**1.62**	**0.78**
	ddG monomer		0.50	2.44	0.70
	no backrub control		0.58	1.73	0.73
	ZEMu paper		**0.64**	1.63	0.75
Multiple mutations, all to alanine	flex ddG	191	0.47	1.78	**0.84**
	ddG monomer		0.34	2.49	0.80
	no backrub control		0.50	**1.69**	0.81
	ZEMu paper		**0.55**	1.72	0.79
Multiple mutations, none to alanine	flex ddG	45	**0.67**	**1.34**	0.58
	ddG monomer		0.40	2.54	0.38
	no backrub control		0.44	1.66	0.58
	ZEMu paper		0.53	1.59	**0.60**
Antibodies	flex ddG	355	**0.60**	**0.93**	**0.75**
	ddG monomer		0.50	1.35	0.69
	no backrub control		0.49	1.06	0.72
	ZEMu paper		0.54	1.06	0.67

Table 2: Main results table. Backrub steps = 35000. R = Pearson's R. MAE = Mean Absolute Error. FC = Fraction Correct. N = number of mutations in the dataset or subset.


Step 0
Input starting structure.
Generate pairwise atom constraints.
1x
Step 1 - Minimize
Global minimization of backbone and side chain torsions.
Pairwise atom constraints from Step 0.
1x
Step 2 - Backrub
Local sampling of backbone and side chain degrees of freedom of pivot residues,
defined as those with neighbor atoms within 8Å of mutation positions.
50x
50x
Wild-Type
Mutant
Step 3a - Pack
Optimize side chains globally
on wild-type structure.
Generate pairwise atom constraints.
Step 3b- Mutate and Pack
Optimize side chains globally
on mutant structure
(using the mutant sequence).
Generate pairwise atom constraints.
50x
50x
Step 4a - Minimize
Global minimization
of backbone and side chains torsions.
Pairwise atom constraints from Step 3a.
Step 4b - Minimize
Global minimization
of backbone and side chains torsions.
Pairwise atom constraints from Step 3b.
50x
50x
Step 5a - Score
Score wild type complex and unbound partners
using Rosetta's all-atom energy function.
Step 5b - Score
Score mutant complex and unbound partners
using Rosetta's all-atom energy function.
50x
50x
Step 6
Calculate interface ΔΔG score.

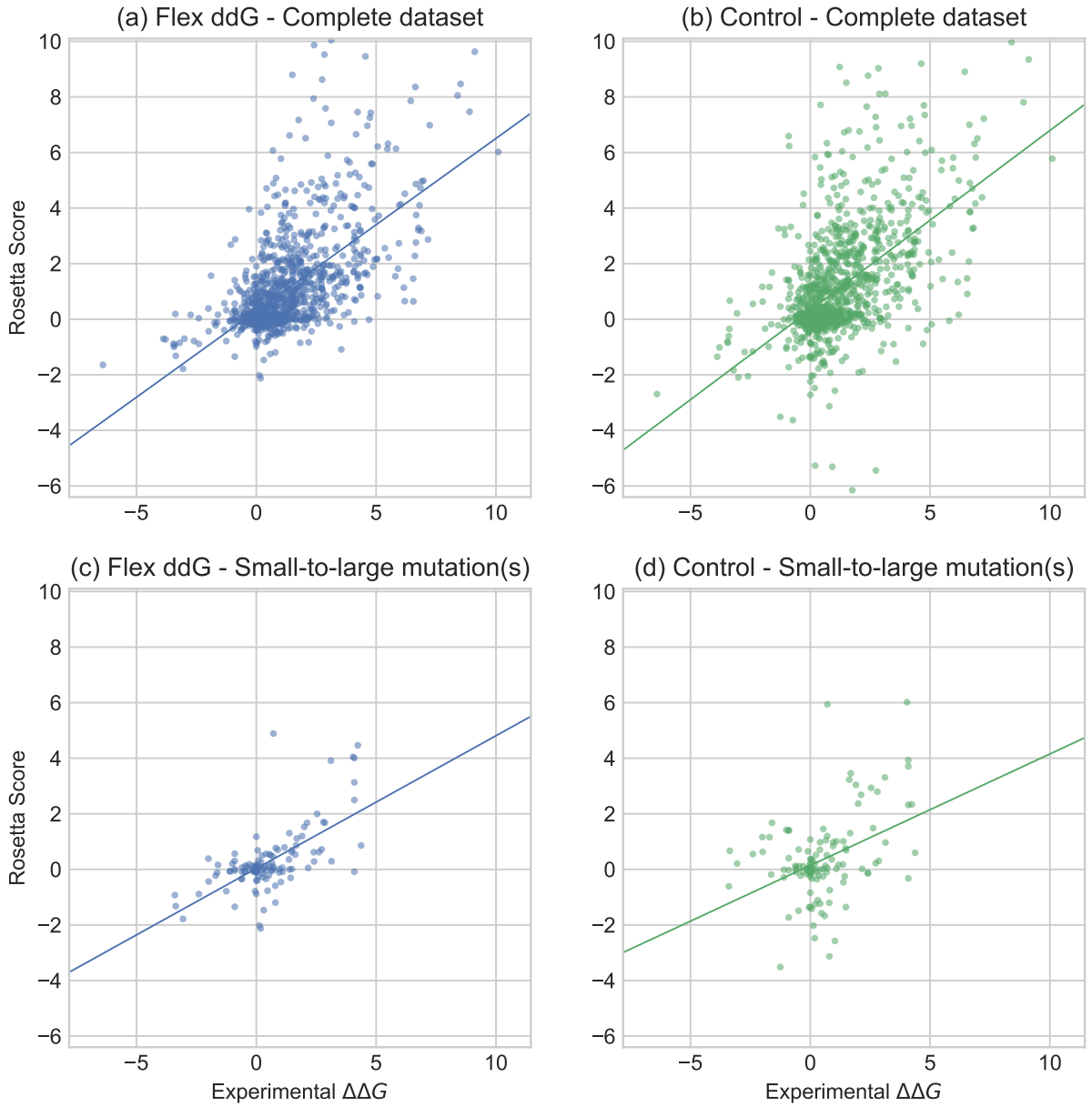


Figure 2: Experimentally determined $\Delta\Delta G$ values (y-axis) vs. Rosetta predictions. (a) Flex ddg method (35000 backrub steps); Complete dataset mutation set (n=1240). (b) No backrub control; Complete dataset mutation set (n=1240). (c) Flex ddg method (35000 backrub steps); Small-to-large mutation(s) mutation set (n=130). (d) No backrub control; Small-to-large mutation(s) mutation set (n=130).

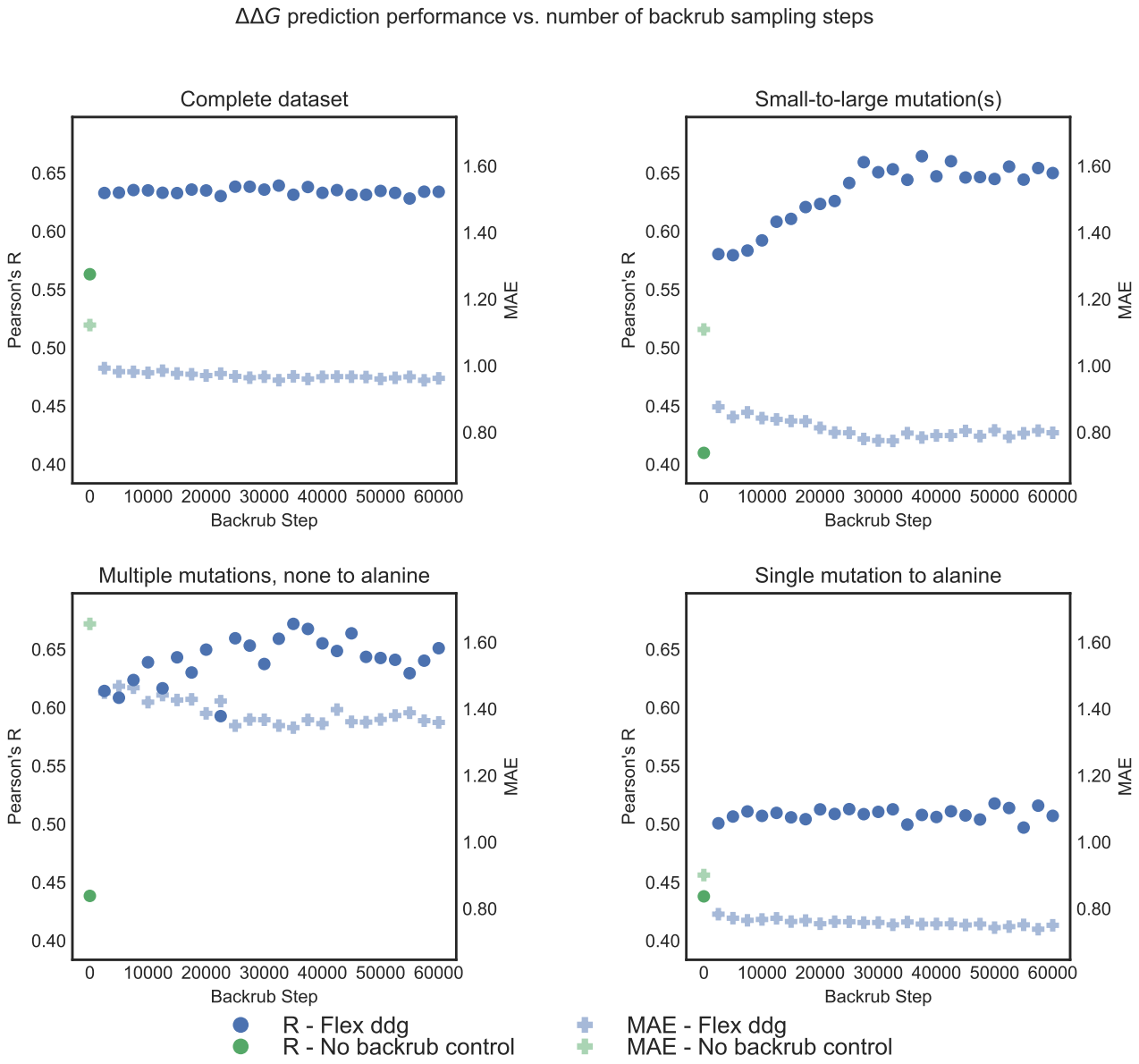


Figure 3: Correlation (Pearson's R) and MAE (Mean Absolute Error) vs. number of backrub steps, on the complete ZEMu set, and subsets. (a) Complete dataset (n=1240) (b) Small-to-large mutation(s) (n=130) (c) Multiple mutations, none to alanine (n=45) (d) Single mutation to alanine (n=748)

Mutation Category	Prediction Method	N	R	MAE	FC
Complete dataset	flex ddG	1240	**0.64**	0.98	**0.76**
	flex ddG (1.6 kT)		**0.64**	**0.93**	0.75
Small-to-large mutation(s)	flex ddG	130	0.59	0.84	0.71
	flex ddG (1.6 kT)		**0.64**	**0.81**	**0.72**
Single mutation to alanine	flex ddG	748	**0.51**	0.77	**0.76**
	flex ddG (1.6 kT)		**0.51**	**0.72**	0.75
Multiple mutations	flex ddG	273	**0.63**	1.64	**0.79**
	flex ddG (1.6 kT)		**0.63**	**1.52**	0.75
Multiple mutations, none to alanine	flex ddG	45	**0.64**	1.42	**0.60**
	flex ddG (1.6 kT)		0.62	**1.38**	0.58

Table 3: Flex ddG performance comparison, when backrub is run with a sampling temperature (kT) of 1.2 or 1.6. Backrub steps = 10000. R = Pearson's R. MAE = Mean Absolute Error. FC = Fraction Correct. N = number of mutations in the dataset or subset.

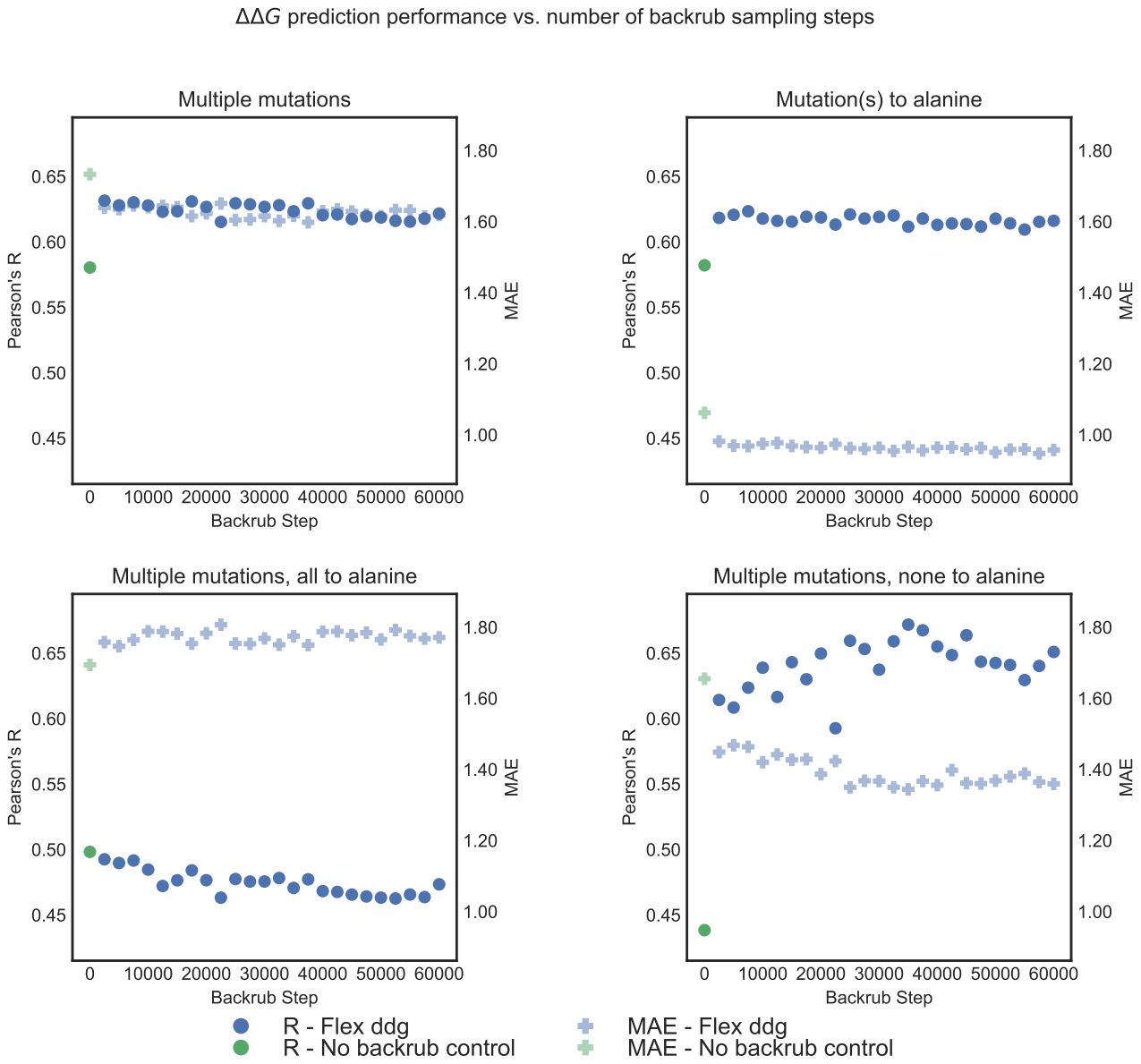


Figure 4: Correlation (Pearson's R) and MAE (Mean Absolute Error) vs. number of backrub steps, on the complete ZEMu set, and subsets. (a) Multiple mutations (n=273) (b) Mutation(s) to alanine (n=939) (c) Multiple mutations, all to alanine (n=191) (d) Multiple mutations, none to alanine (n=45)

Mutation Category	Prediction Method	N	R	MAE	FC
Multiple mutations	flex ddG	273	0.62	**1.62**	**0.78**
	no backrub control		0.58	1.73	0.73
	ZEMu paper		**0.64**	1.63	0.75
Multiple mutations, all to alanine	flex ddG	191	0.47	1.78	**0.84**
	no backrub control		0.50	**1.69**	0.81
	ZEMu paper		**0.55**	1.72	0.79
Multiple mutations, none to alanine	flex ddG	45	**0.67**	**1.34**	0.58
	no backrub control		0.44	1.66	0.58
	ZEMu paper		0.53	1.59	**0.60**
Mutation(s) to alanine	flex ddG	939	0.61	**0.97**	**0.78**
	no backrub control		0.58	1.06	0.75
	ZEMu paper		**0.62**	1.03	0.73

Table 4: Multiple mutations results (backrub steps = 35000). R = Pearson's R. MAE = Mean Absolute Error. FC = Fraction Correct. N = number of mutations in the dataset or subset.

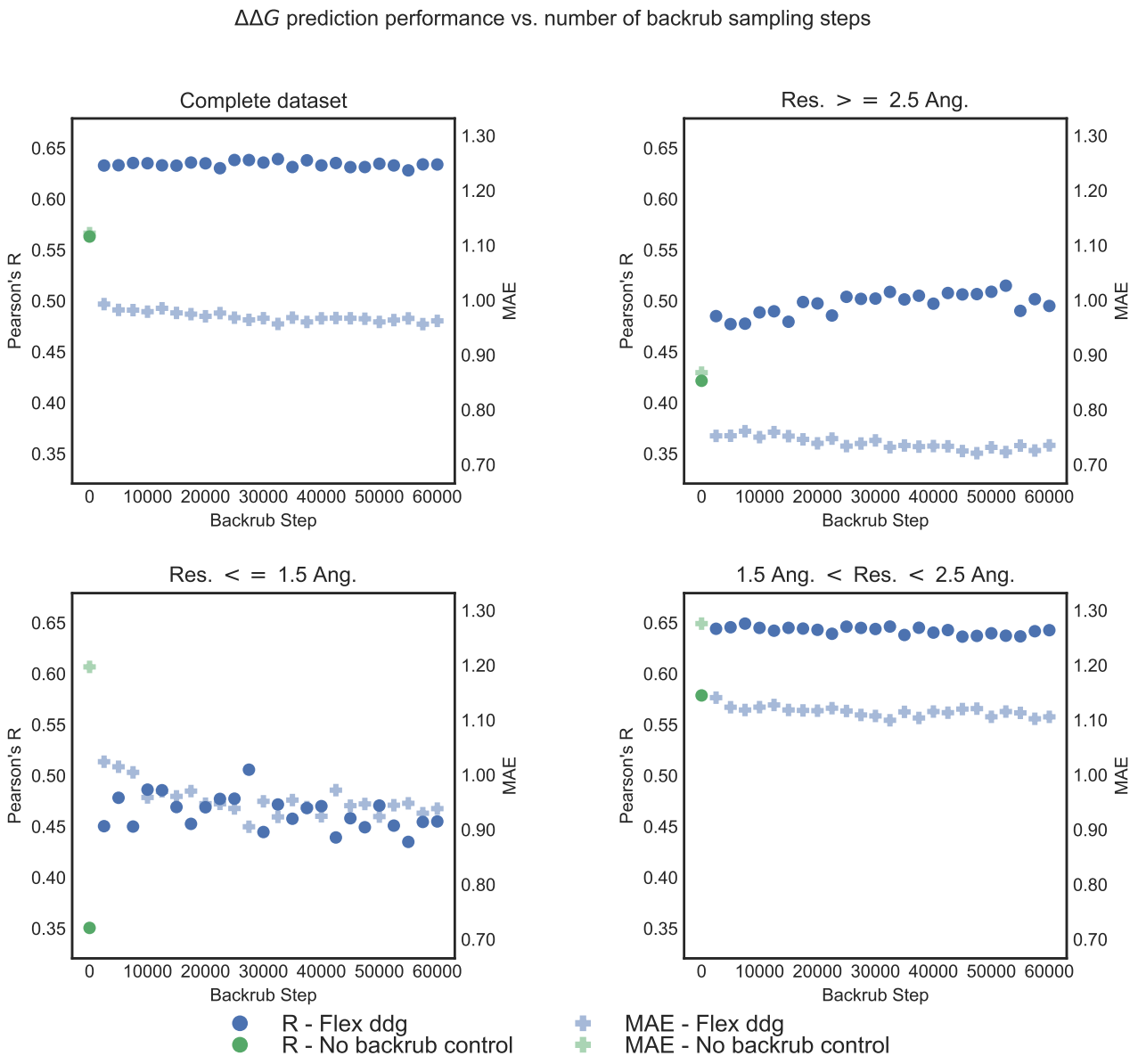


Figure 5: Correlation (Pearson's R) and MAE (Mean Absolute Error) vs. number of backrub steps, on the complete ZEMu set, and subsets. (a) Complete dataset (n=1240) (b) Res. >= 2.5 ang. (n=457) (c) Res. <= 1.5 ang. (n=52) (d) 1.5 ang. < res. < 2.5 ang. (n=731)

Mutation Category	Prediction Method	N	R	MAE	FC
Complete dataset	flex ddG	1240	**0.63**	**0.97**	**0.76**
	no backrub control		0.56	1.12	0.73
	ddG monomer		0.62	1.16	0.75
	ZEMu paper		0.61	1.08	0.71
Antibodies	flex ddG	355	**0.60**	**0.93**	0.75
	no backrub control		0.49	1.06	0.72
	ddG monomer		0.58	1.07	**0.77**
	ZEMu paper		0.54	1.06	0.67

Table 5: Performance of the Rosetta flex ddG method on the subset of complexes containing an antibody binding partner (backrub steps = 35000). R = Pearson's R. MAE = Mean Absolute Error. FC = Fraction Correct. N = number of mutations in the dataset or subset.

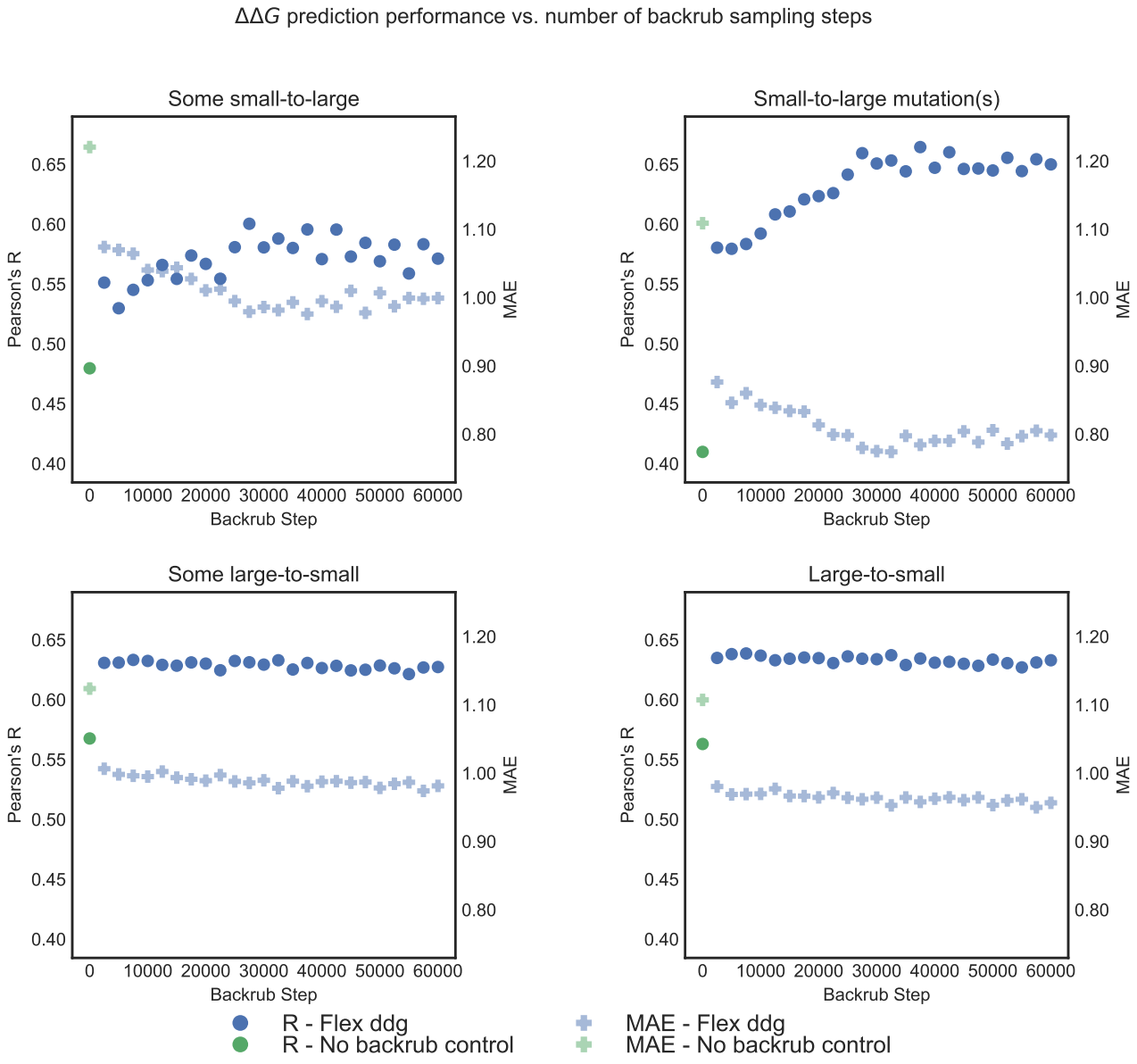


Figure 6: Correlation (Pearson's R) and MAE (Mean Absolute Error) vs. number of backrub steps, on the complete ZEMu set, and subsets. (a) Some small-to-large (n=164) (b) Small-to-large mutation(s) (n=130) (c) Some large-to-small (n=1110) (d) Large-to-small (n=1076)

Git SHA1	Protocol
69aa5266f0d5	flex ddG
69aa5266f0d5	no backrub control
3b2aa5cc3798	ddG monomer

Table 6: SHA1 Git version of Rosetta used for benchmarking

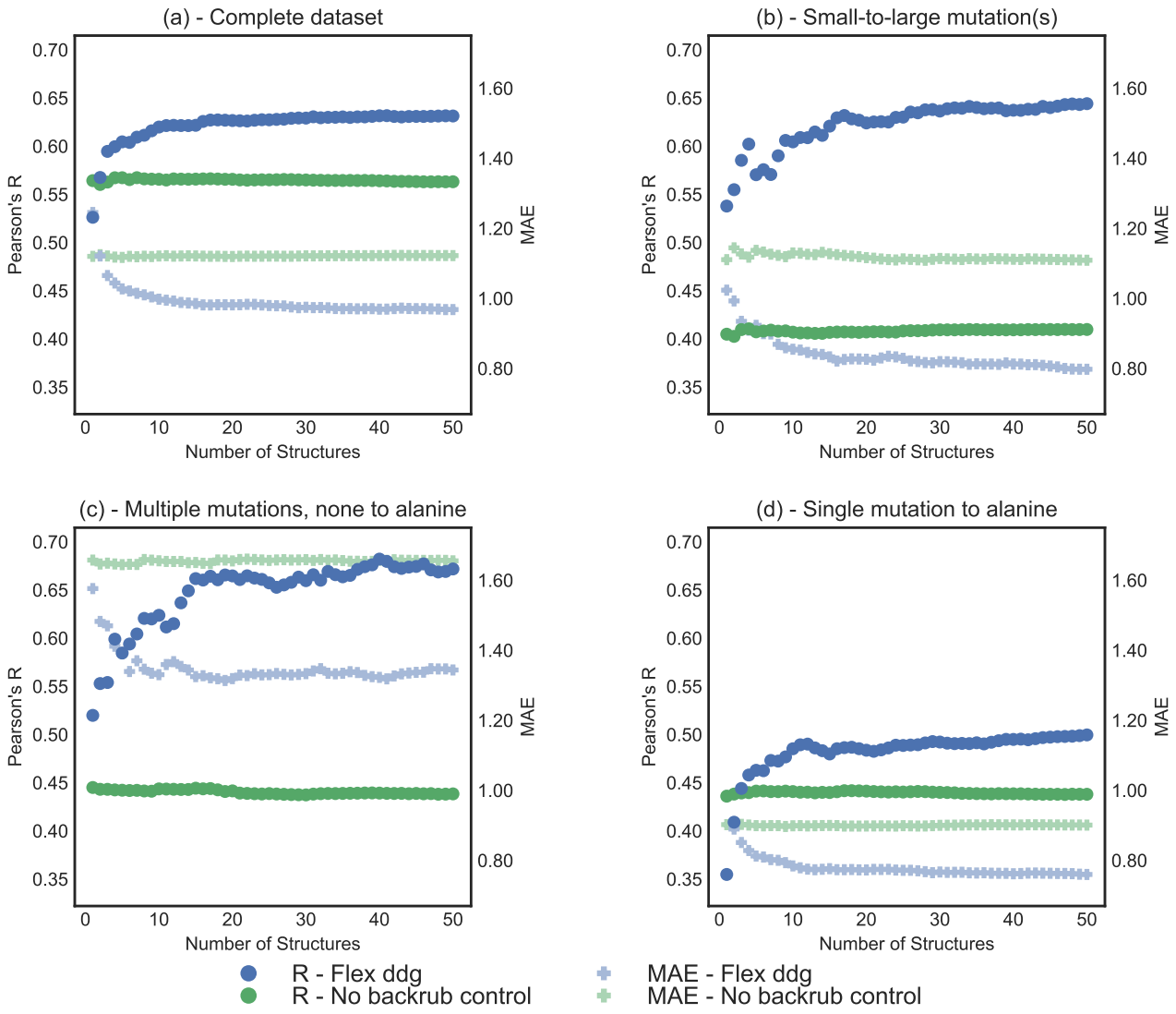


Figure 7: Correlation (Pearson's R, left y-axis) and MAE (Mean Absolute Error, right y-axis) vs. number of averaged structures (x-axis), on the complete ZEMu set, and subsets. Pearson's R is shown with circular points, and MAE with faded plus-shaped points. (a) Complete dataset (n = 1240, backrub steps = 35000) (b) Small-to-large mutation(s) (n = 130, backrub steps = 35000) (c) Multiple mutations, none to alanine (n = 45, backrub steps = 35000) (d) Single mutation to alanine (n = 748, backrub steps = 35000)

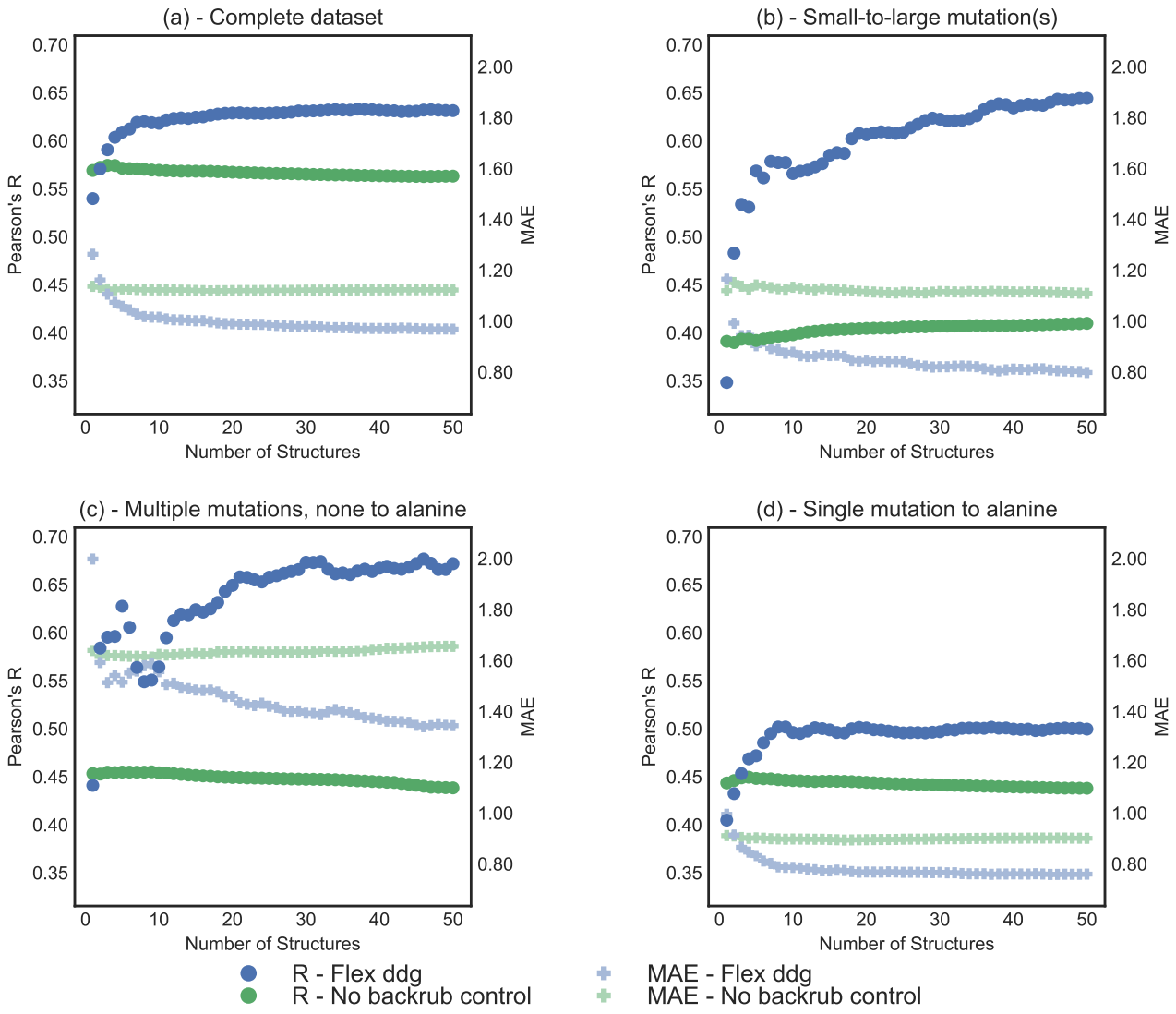


Figure 8: Correlation (Pearson's R, left y-axis) and MAE (Mean Absolute Error, right y-axis) vs. number of averaged structures (x-axis), on the complete ZEMu set, and subsets. Pearson's R is shown with circular points, and MAE with faded plus-shaped points. (a) Complete dataset (n = 1240, backrub steps = 35000) (b) Small-to-large mutation(s) (n = 130, backrub steps = 35000) (c) Multiple mutations, none to alanine (n = 45, backrub steps = 35000) (d) Single mutation to alanine (n = 748, backrub steps = 35000)

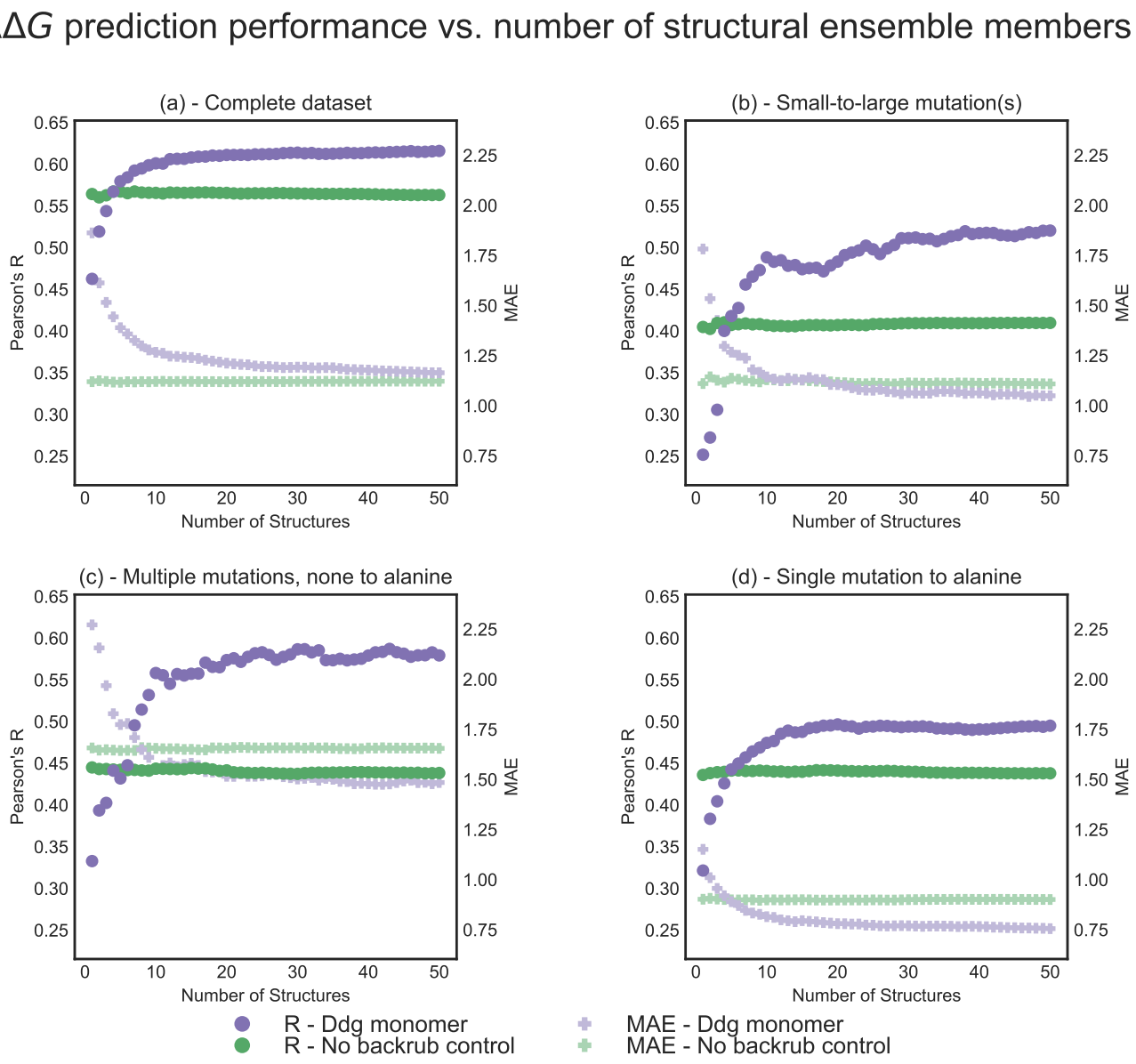


Figure 9: Correlation (Pearson's R, left y-axis) and MAE (Mean Absolute Error, right y-axis) vs. number of averaged structures (x-axis), on the complete ZEMu set, and subsets. Pearson's R is shown with circular points, and MAE with faded plus-shaped points. (a) Complete dataset (n = 1240) (b) Small-to-large mutation(s) (n = 130) (c) Multiple mutations, none to alanine (n = 45) (d) Single mutation to alanine (n = 748)

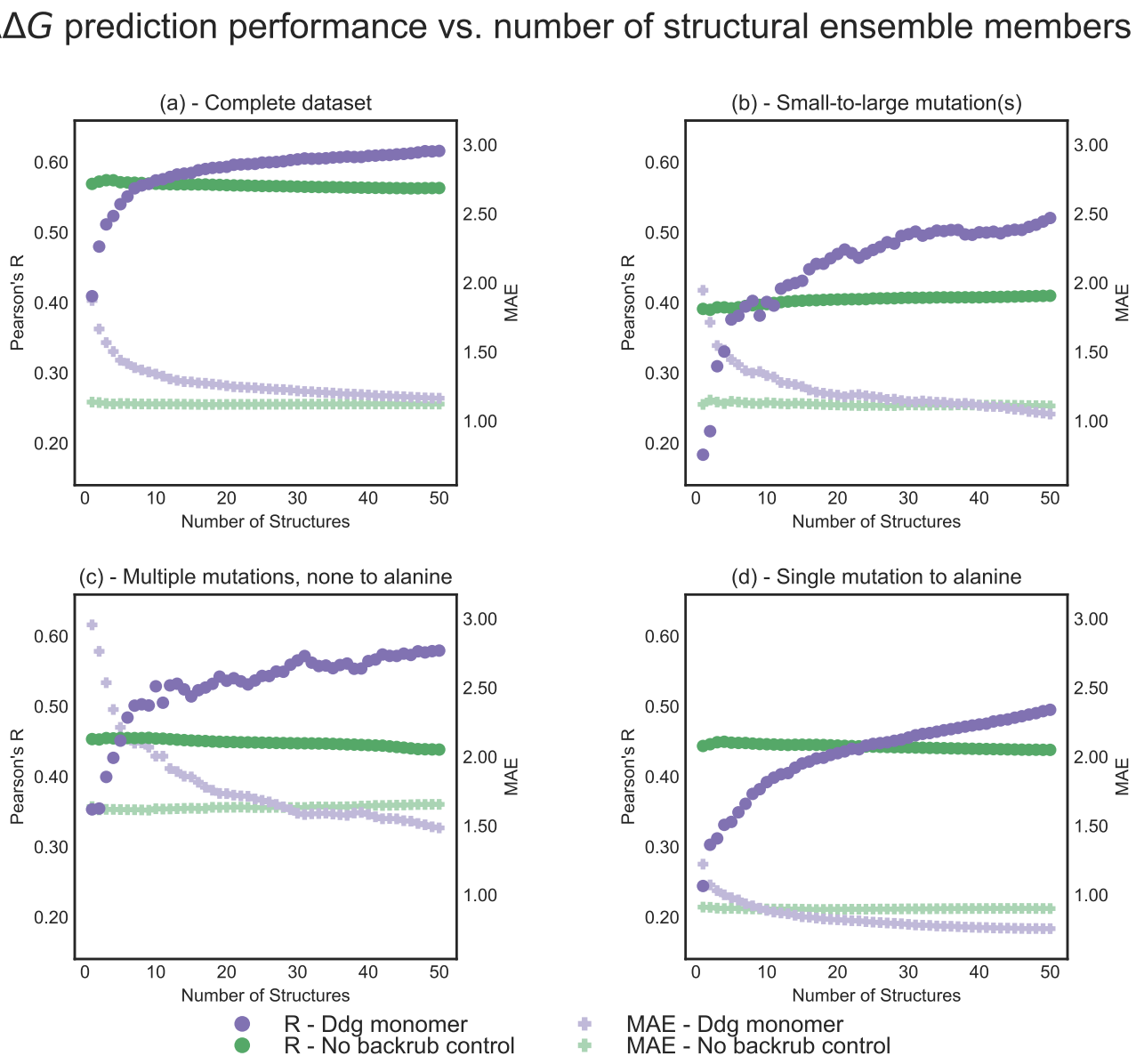


Figure 10: Correlation (Pearson's R, left y-axis) and MAE (Mean Absolute Error, right y-axis) vs. number of averaged structures (x-axis), on the complete ZEMu set, and subsets. Pearson's R is shown with circular points, and MAE with faded plus-shaped points. (a) Complete dataset (n = 1240) (b) Small-to-large mutation(s) (n = 130) (c) Multiple mutations, none to alanine (n = 45) (d) Single mutation to alanine (n = 748)

Mutation Category	Prediction Method	N	R	MAE	FC
Stabilizing	flex ddG	32	**0.62**	2.11	0.09
	no backrub control		0.50	2.31	**0.31**
	ddG monomer		0.39	2.18	0.19
	ZEMu paper		0.31	**2.01**	**0.31**
Neutral	flex ddG	719	**0.19**	**0.52**	**0.87**
	no backrub control		0.10	0.72	0.78
	ddG monomer		0.13	0.75	0.80
	ZEMu paper		0.16	0.66	0.79
Positive	flex ddG	489	**0.48**	**1.55**	0.64
	no backrub control		0.44	1.63	0.67
	ddG monomer		0.47	1.71	**0.72**
	ZEMu paper		**0.48**	1.63	0.62

Table 7: Performance of the Rosetta flex ddG method on the subset of mutations experimentally determined to be stabilizing ($\Delta\Delta G <= -1$), neutral ($-1 < \Delta\Delta G < 1$), or destabilizing ($\Delta\Delta G >= 1$). Backrub steps = 35000. R = Pearson's R. MAE = Mean Absolute Error. FC = Fraction Correct. N = number of mutations in the dataset or subset.

Mutation Category	Prediction Method	N	R	MAE	FC
pdb-1A22	flex ddG	142	**0.33**	**0.62**	0.77
	no backrub control		0.18	0.77	0.74
	ddG monomer		0.12	0.91	0.73
	ZEMu paper		0.19	0.68	**0.78**
pdb-1A4Y	flex ddG	45	0.80	1.38	**0.78**
	no backrub control		0.79	1.47	**0.78**
	ddG monomer		0.77	1.91	0.62
	ZEMu paper		**0.87**	**1.12**	0.73
pdb-1ACB	flex ddG	6	0.25	2.80	0.83
	no backrub control		0.23	2.37	0.83
	ddG monomer		0.58	**1.57**	**1.00**
	ZEMu paper		**0.79**	2.17	0.83
pdb-1AHW	flex ddG	10	-0.77	1.19	0.4
	no backrub control		-0.42	1.42	0.4
	ddG monomer		-0.34	1.26	0.5
	ZEMu paper		**0.30**	**0.93**	**0.6**
pdb-1AK4	flex ddG	15	**0.69**	**0.59**	**0.73**
	no backrub control		0.35	1.01	0.47
	ddG monomer		0.63	1.35	0.60
	ZEMu paper		0.44	1.63	0.53
pdb-1CBW	flex ddG	15	-0.08	**0.65**	0.60
	no backrub control		**0.05**	0.83	**0.67**
	ddG monomer		-0.09	0.72	**0.67**
	ZEMu paper		-0.26	0.71	**0.67**
pdb-1CSE	flex ddG	6	0.53	1.83	0.67
	no backrub control		0.37	2.03	0.67
	ddG monomer		0.46	1.88	0.67
	ZEMu paper		**0.87**	**0.81**	**1.00**
pdb-1DAN	flex ddG	118	0.65	**0.53**	0.84
	no backrub control		**0.69**	0.59	**0.85**
	ddG monomer		0.61	0.71	0.83
	ZEMu paper		0.32	0.88	0.76
pdb-1DFJ	flex ddG	20	0.71	1.23	**0.70**
	no backrub control		**0.83**	**1.04**	0.60
	ddG monomer		0.69	1.38	0.55
	ZEMu paper		0.55	1.40	0.55
pdb-1DQJ	flex ddG	34	**0.44**	**1.69**	0.79
	no backrub control		0.39	1.93	0.65
	ddG monomer		0.37	1.87	**0.82**
	ZEMu paper		0.28	2.08	0.59
pdb-1DVF	flex ddG	38	**0.65**	1.54	0.61
	no backrub control		**0.65**	**1.50**	0.66
	ddG monomer		0.61	1.54	**0.71**
	ZEMu paper		0.57	1.54	0.53
pdb-1E96	flex ddG	6	**0.80**	**0.83**	0.50
	no backrub control		0.51	0.91	0.50
	ddG monomer		0.45	0.96	0.50
	ZEMu paper		0.50	0.85	**0.67**
pdb-1EAW	flex ddG	27	-0.02	0.62	0.89
	no backrub control		0.07	0.73	0.81
	ddG monomer		**0.13**	0.61	0.89
	ZEMu paper		0.00	**0.49**	**0.93**
pdb-1EMV	flex ddG	51	**0.89**	**0.87**	**0.86**
	no backrub control		0.84	0.98	0.84
	ddG monomer		0.84	0.96	0.80
	ZEMu paper		0.87	0.89	0.84
pdb-1F47	flex ddG	12	0.52	**0.78**	0.50
	no backrub control		0.58	0.87	**0.58**
	ddG monomer		**0.60**	0.87	**0.58**
	ZEMu paper		0.51	1.02	0.42
	flex ddG		-0.15	0.90	0.56

Mutation Category	Prediction Method	N	R	MAE	FC
Complete dataset	flex ddG	1240	**0.63**	**0.97**	**0.76**
	flex ddG (REF energy)		**0.63**	1.19	0.75
Small-to-large mutation(s)	flex ddG	130	**0.64**	**0.80**	0.71
	flex ddG (REF energy)		0.57	1.01	**0.72**
Single mutation to alanine	flex ddG	748	**0.50**	**0.76**	**0.77**
	flex ddG (REF energy)		0.49	0.90	0.74
Multiple mutations	flex ddG	273	**0.62**	**1.62**	0.78
	flex ddG (REF energy)		0.59	2.12	**0.80**
Res. <= 1.5 Ang.	flex ddG	52	0.46	**0.95**	0.73
	flex ddG (REF energy)		**0.65**	1.10	**0.75**
Res. >= 2.5 Ang.	flex ddG	457	**0.50**	**0.74**	**0.76**
	flex ddG (REF energy)		0.48	0.90	0.75

Table 9: Performance comparison of the standard flex ddG protocol (using Rosetta's Talaris energy function) with flex ddG run with the REF score function. "res <= 1.5 Ang." indicates data points for which the resolution of the input wild-type crystal structure is less than or equal to 1.5 Å. Backrub steps = 35000. R = Pearson's R. MAE = Mean Absolute Error. FC = Fraction Correct. N = number of mutations in the dataset or subset.