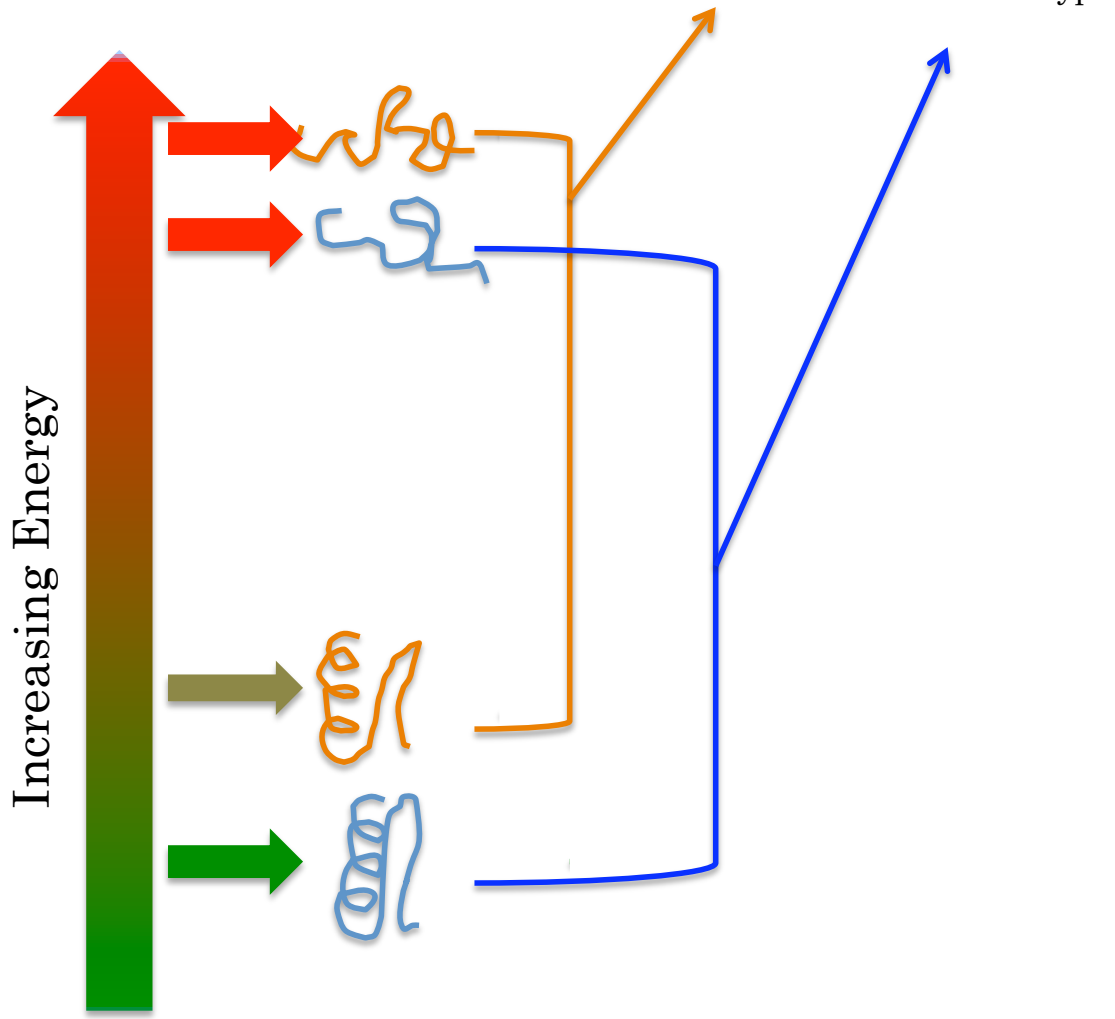
Predicting ΔΔGs

Elizabeth Kellogg
ekellogg@u.washington.edu
8/5/09

Mutations can alter the free-energy landscape, changing free energy of folding

$$ddG = dG_{mutant} - dG_{wildtype}$$

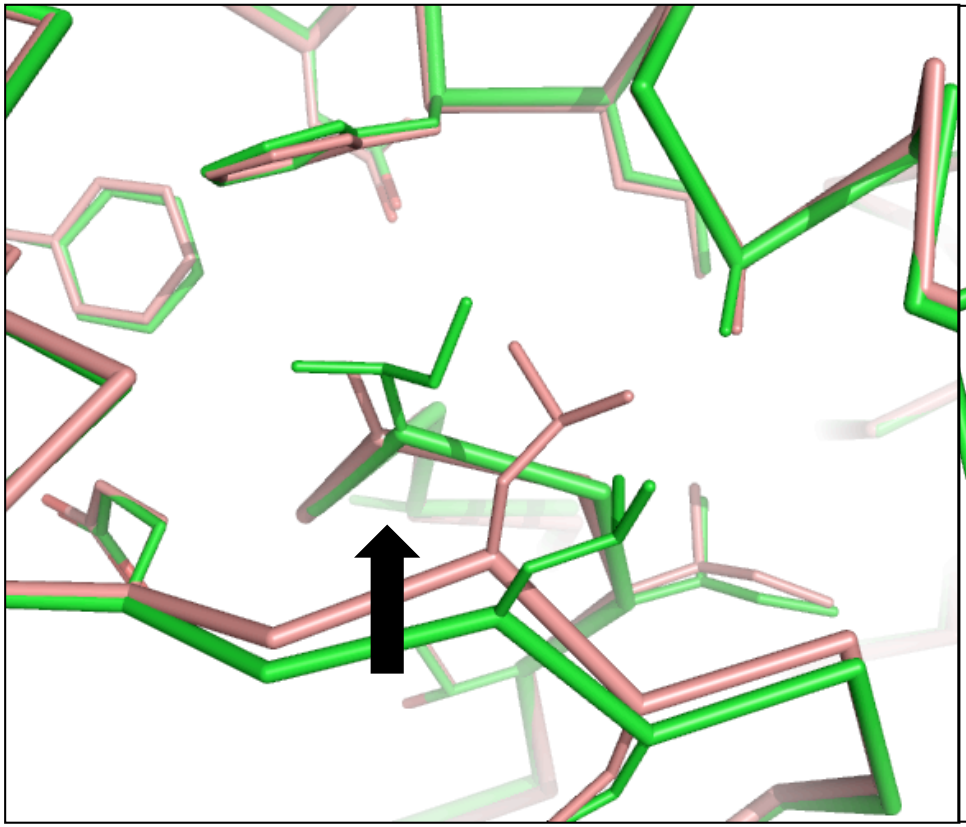


Mutations can alter the folded state….

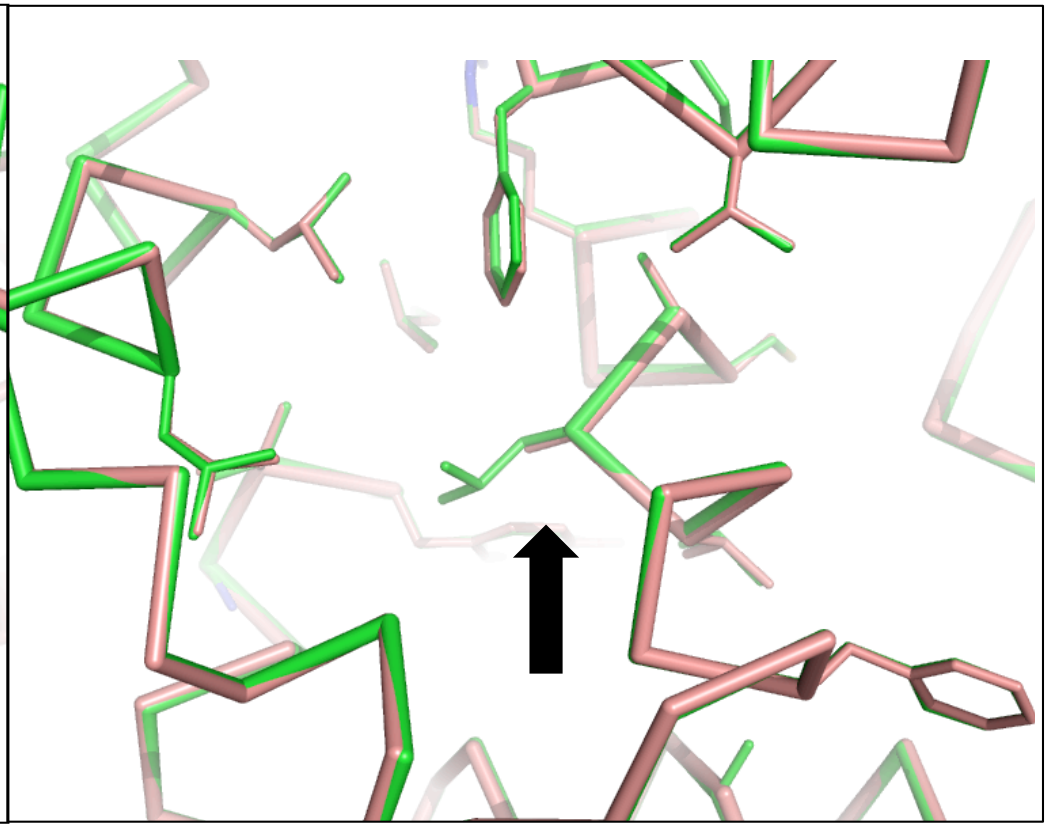
As well as the unfolded state

Why use computational techniques to predict ΔΔGs?

Understand relationship between protein structure and its energy



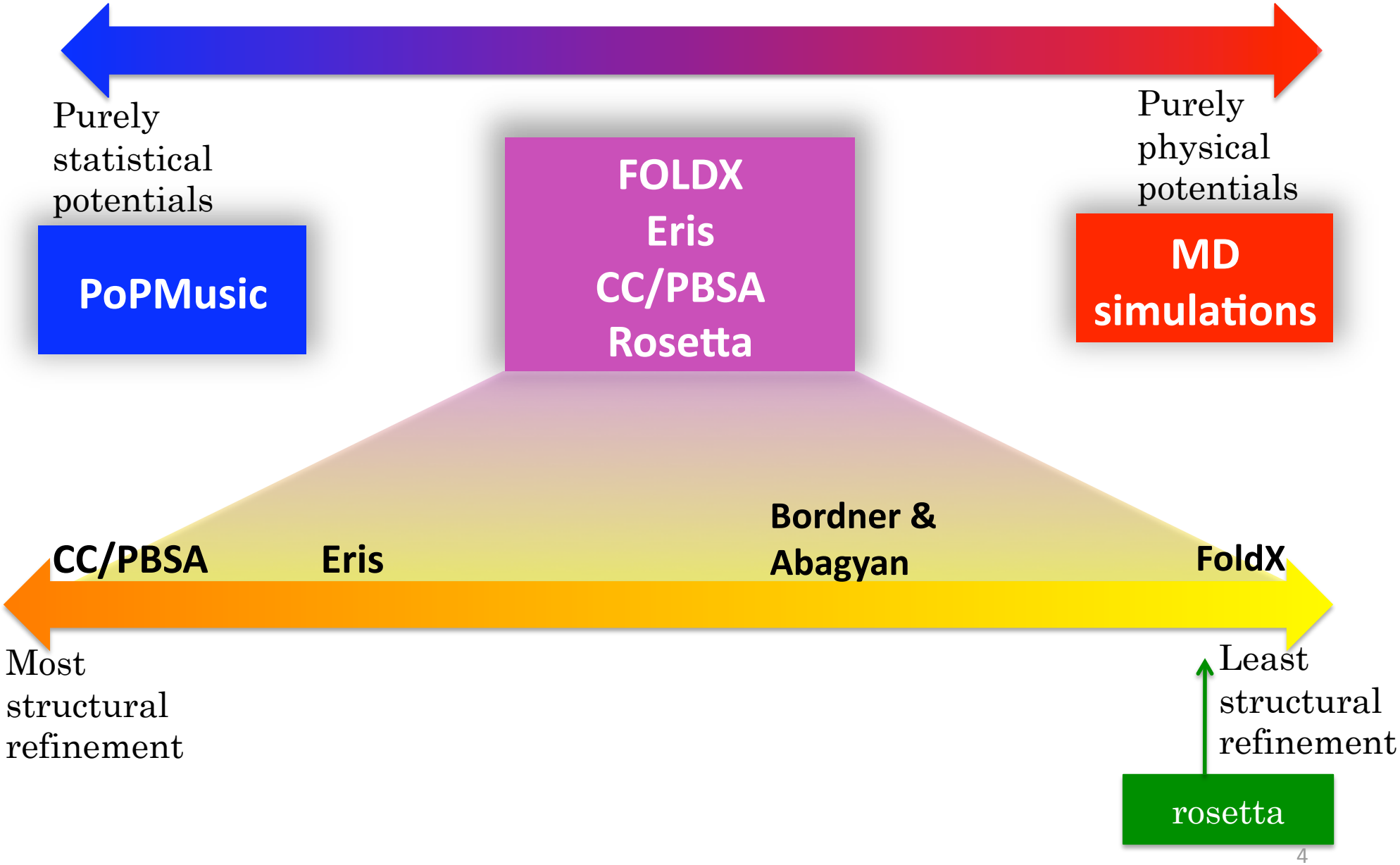
I29A 2.7 kcal/mol



L99A 5 kcal/mol

Wildtype green, **Mutant pink**

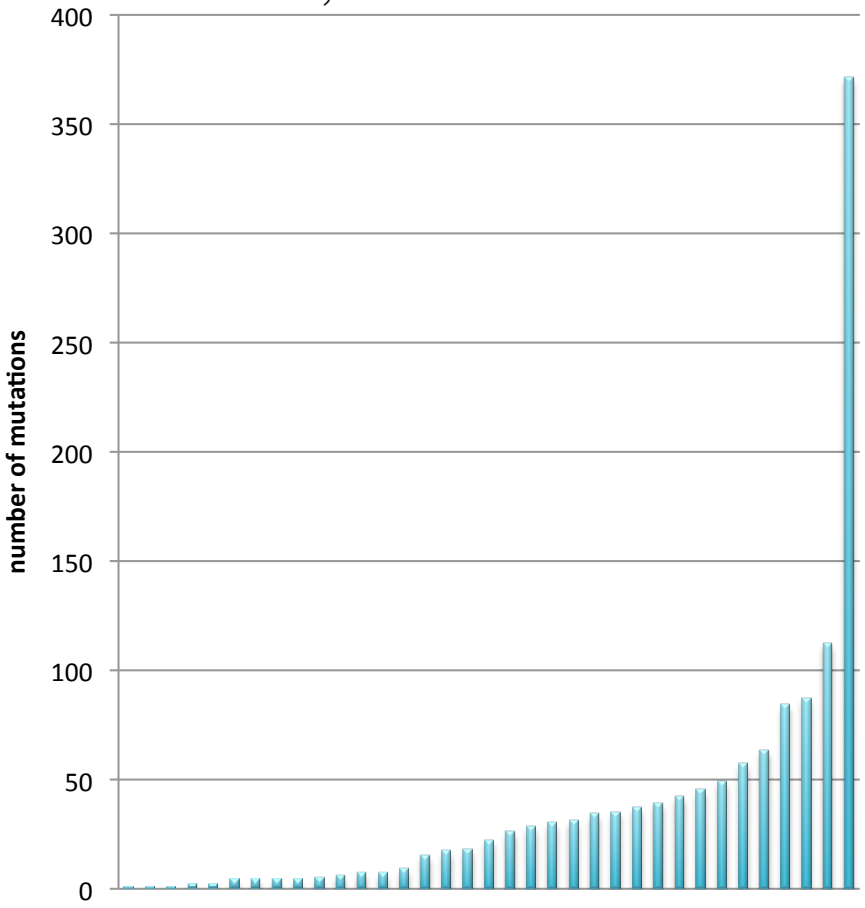
Previous computational algorithms for estimating changes in stability



A large set of mutations has been gathered to aid method development

76 proteins

1,286 mutations

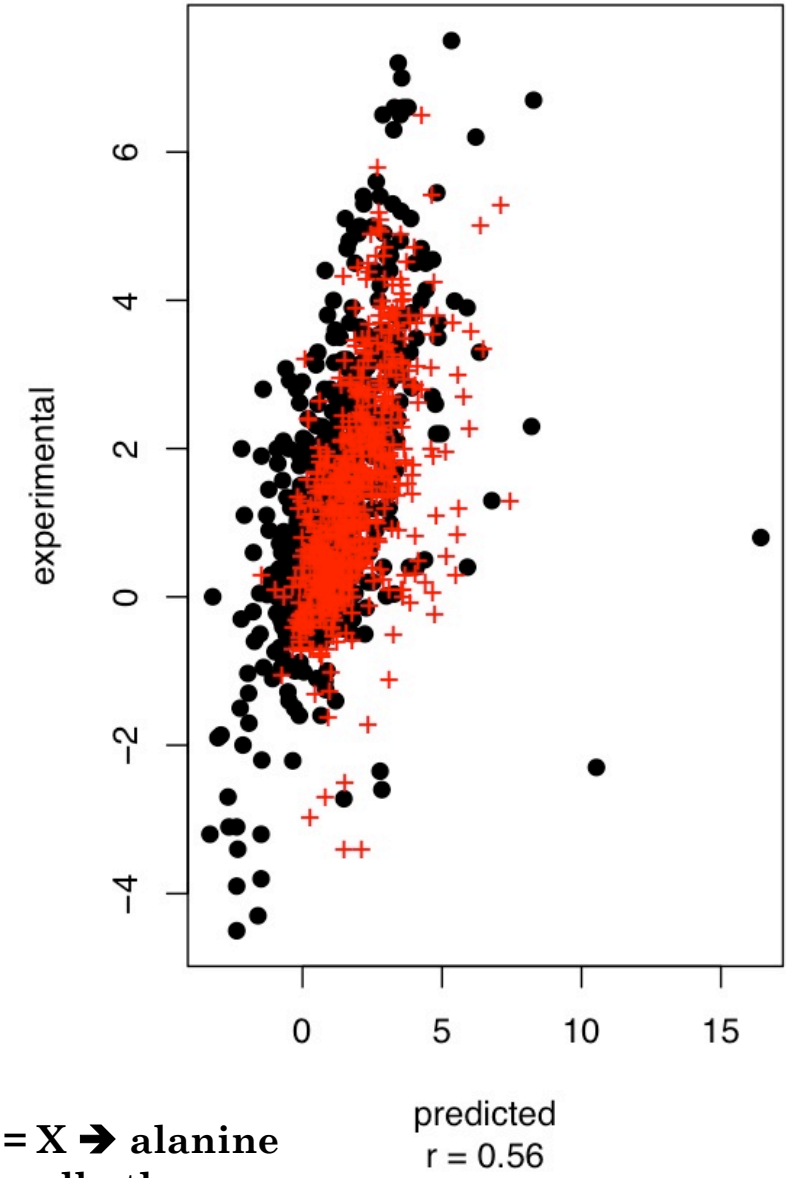


37	FK506 BINDING PROTEIN	26	AZURIN
17	HUMAN PROCARBOXYPEPTIDASE A2	63	Crystal Structure of Repeats 15 and 16 of Chicken Brain Alpha Spectrin
112	BARNASE	30	PROTEIN G
22	CHROMOSOMAL DNA-BINDING PROTEIN SSO7D/ D(GCGAACGC) COMPLEX	28	APOFLAVODOXIN
35	THE STRUCTURE OF BOVINE PANCREATIC TRYPSIN INHIBITOR	45	NATIVE HUMAN LYSOZYME
34	APOMYOGLOBIN	18	GLN 25-RIBONUCLEASE T1
5	CHEMOTAXIS Y PROTEIN	4	SH3 DOMAIN
4	BACILLUS SUBTILIS MAJOR COLD SHOCK PROTEIN, CSPB	15	SRC-HOMOLOGY 3 (SH3) DOMAIN
1	BOVINE CYTOCHROME B(5)	371	STAPHYLOCOCCAL NUCLEASE
6	DIHYDROFOLATE REDUCTASE	1	SUBTILISIN BPN'
49	HUMAN TYROSINE-PROTEIN KINASE C-SRC	39	FIBRONECTIN TYPE III DOMAIN
4	FERREDOXIN	4	YEAST ISO-1-CYTOCHROME C
7	ALPHA-LACTALBUMIN	87	CHYMOTRYPSIN INHIBITOR 2
1	HUMAN GROWTH HORMONE	9	HUMAN LYSOZYME
57	B1 DOMAIN OF PROTEIN L	31	CheY
2	FATTY ACID-BINDING APOPROTEIN	84	BACTERIOPHAGE T4 LYSOZYME
2	TRIOSE PHOSPHATE ISOMERASE	7	HEN EGG-WHITE LYSOZYME
42	T4-lysozyme mutant		

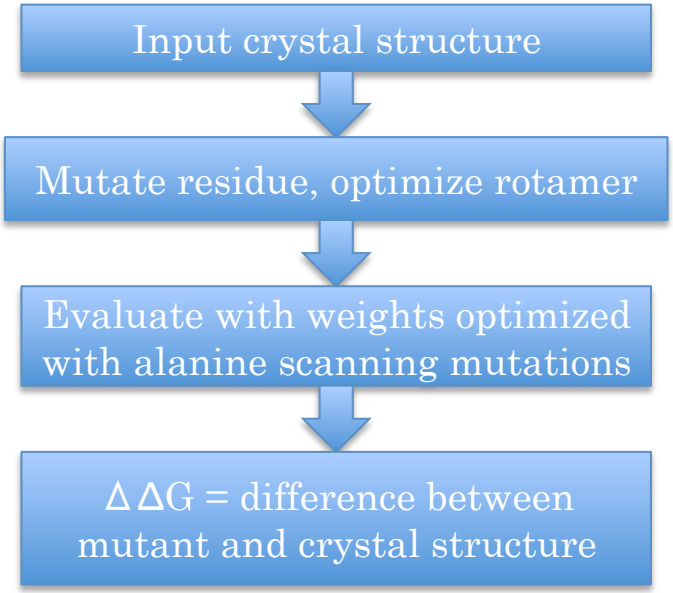
Guerois R. et al. "*Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations*", JMB (2002): 320, 369-387

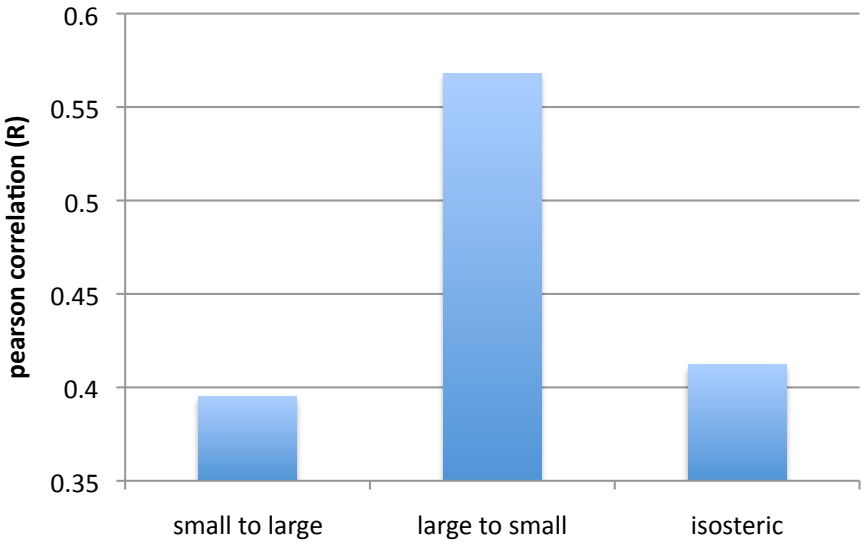
Yin et al. "*Modeling Backbone Flexibility Improves Protein Stability Estimation*" Structure (2007)15: 1567-1576

The previously published ΔΔG algorithm is not general enough to handle all types of mutations

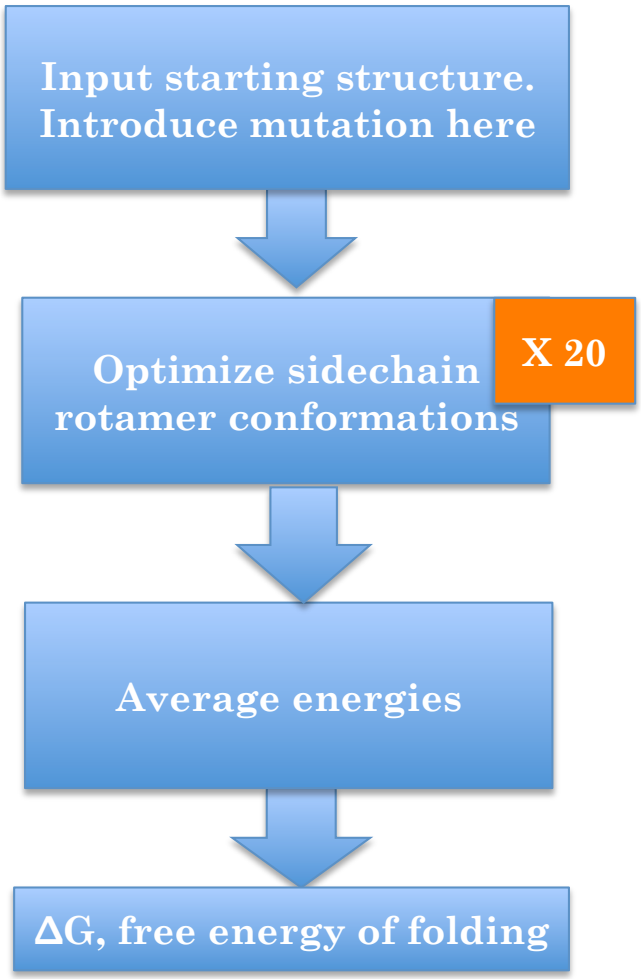


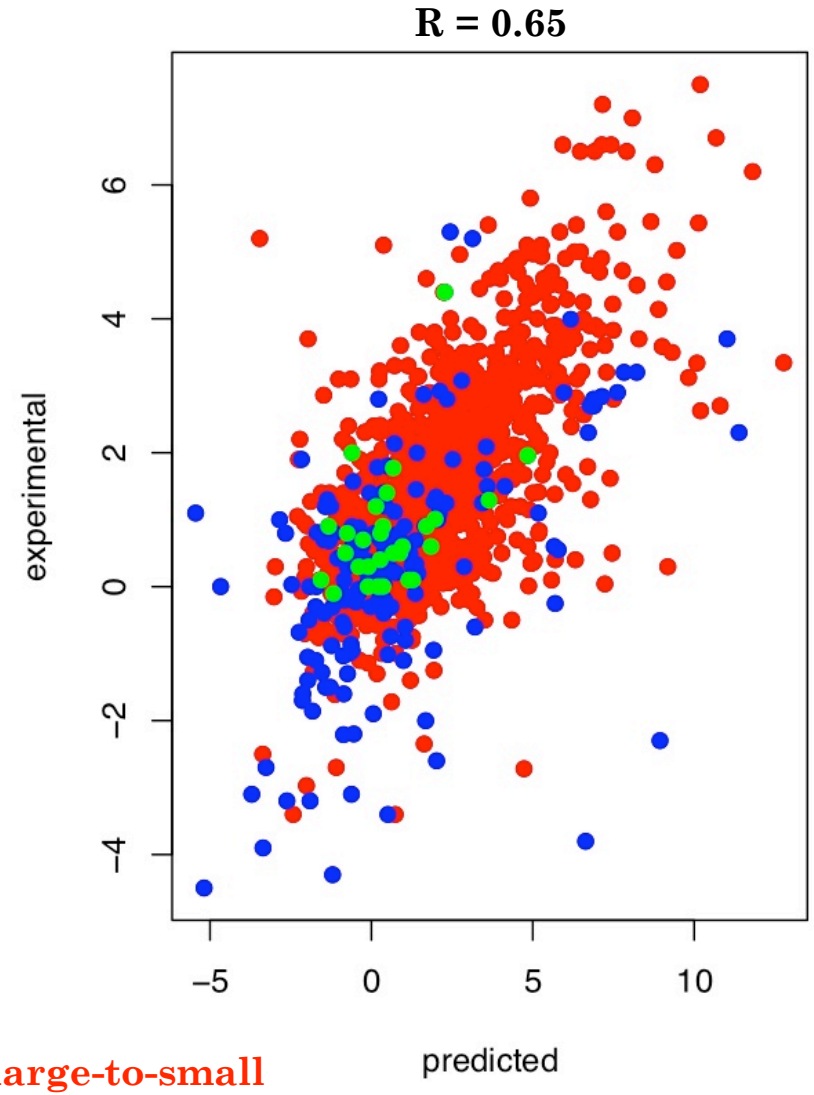
+ = X ➔ alanine
• = all other mutations





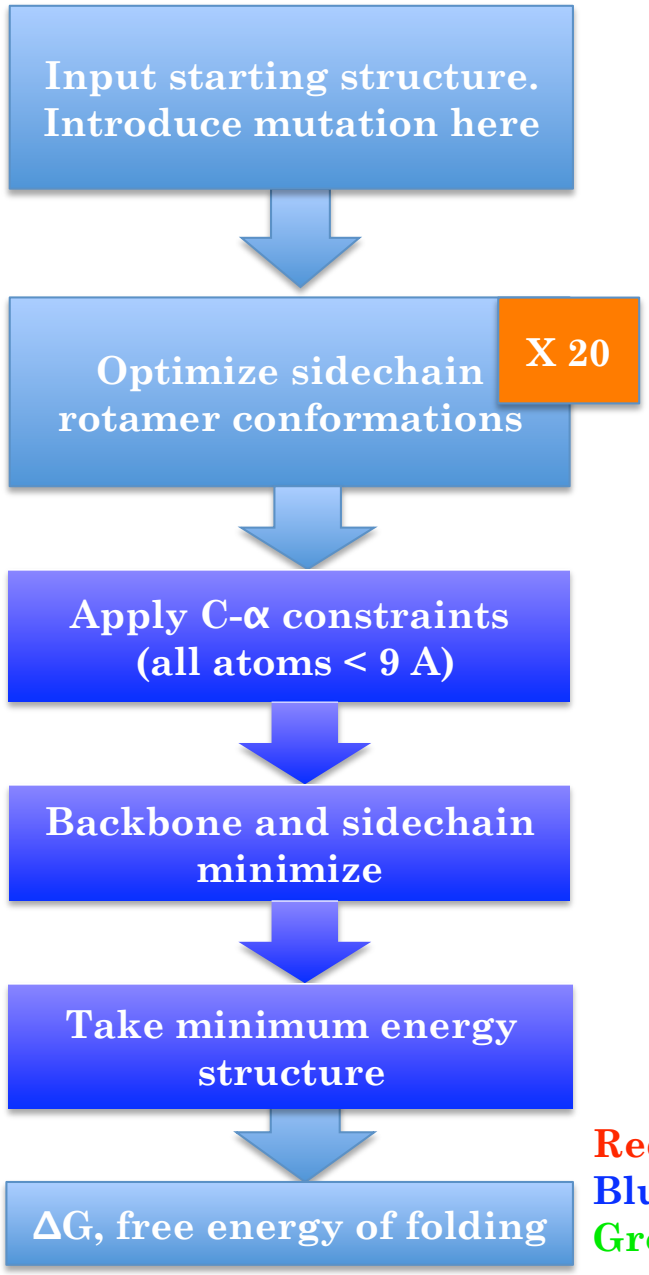
Side-chain optimization and averaging energies improve predictions

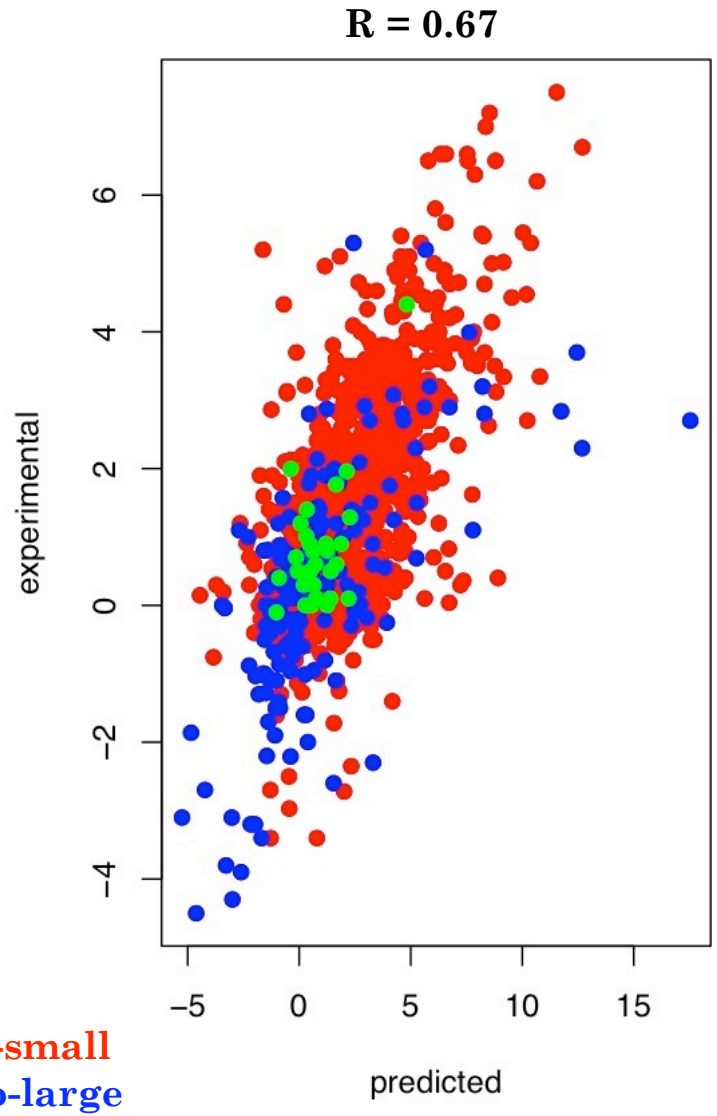




Red = large-to-small
Blue = small-to-large
Green = isosteric

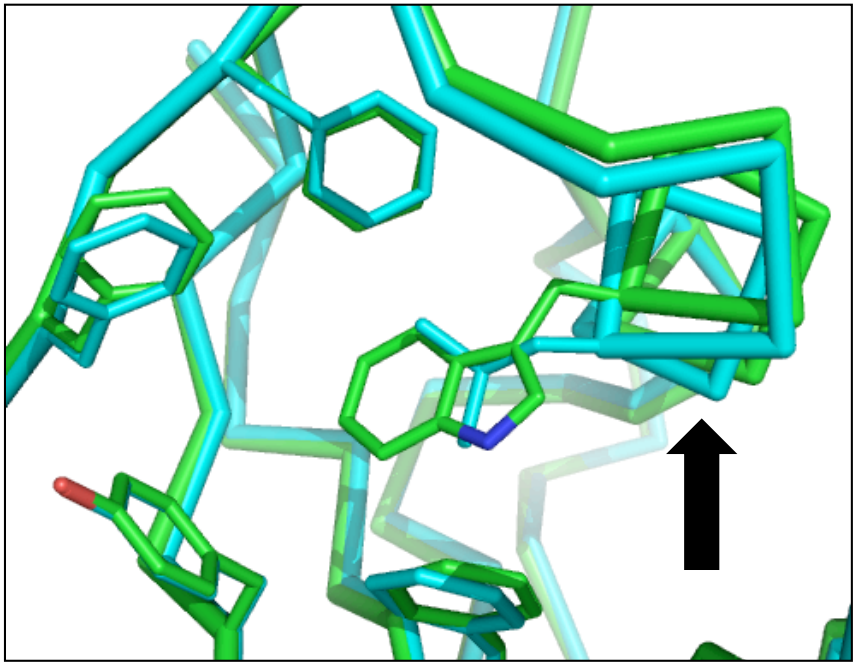
Introduction of minimization improves correlation further

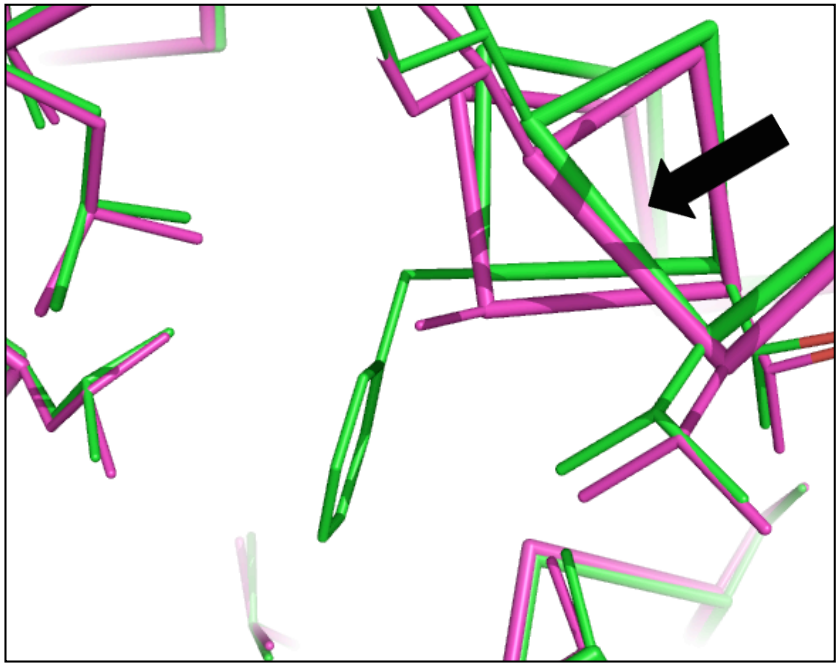




Red = large-to-small
Blue = small-to-large
Green = isosteric

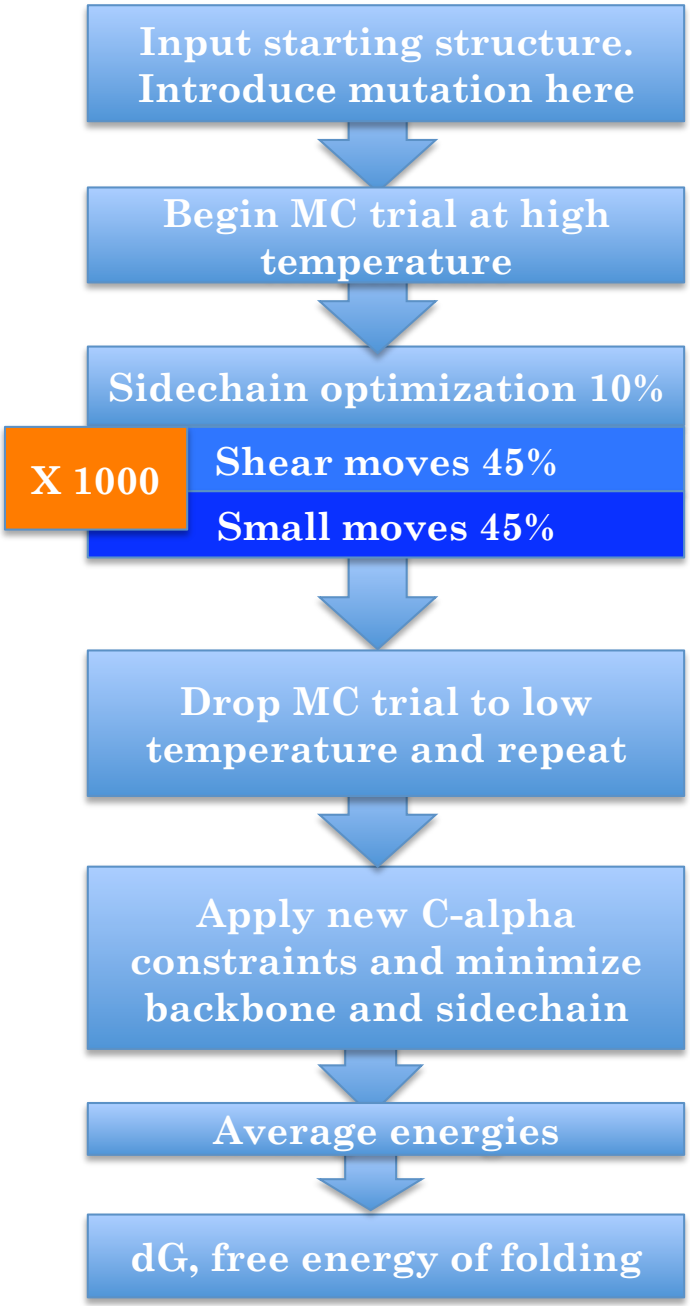
However, perhaps we could improve predictions if we allow more relaxation for non-conservative mutations

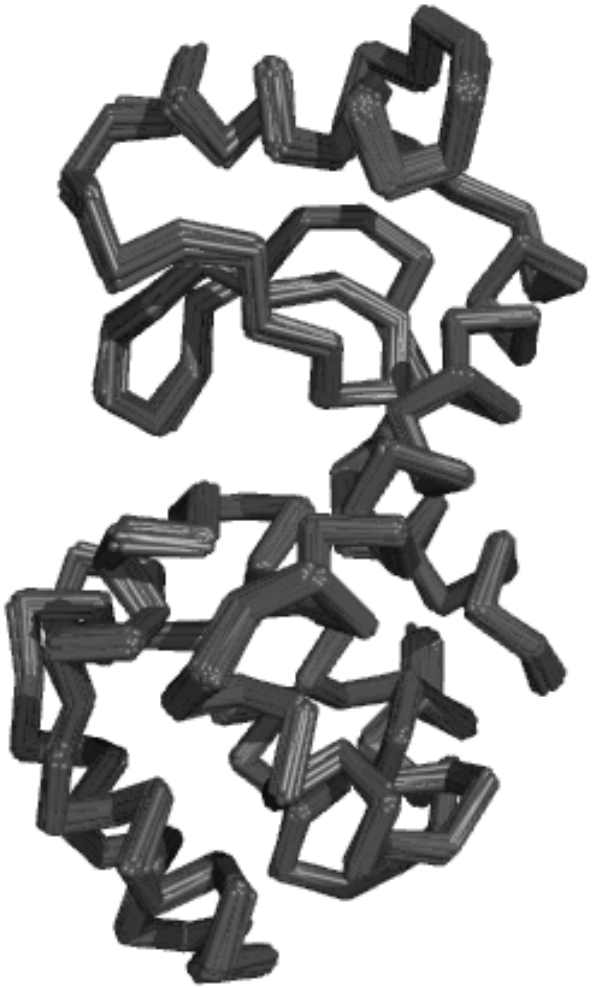




Left: FK506 binding protein W59L
Right: T4-lysozyme F153A

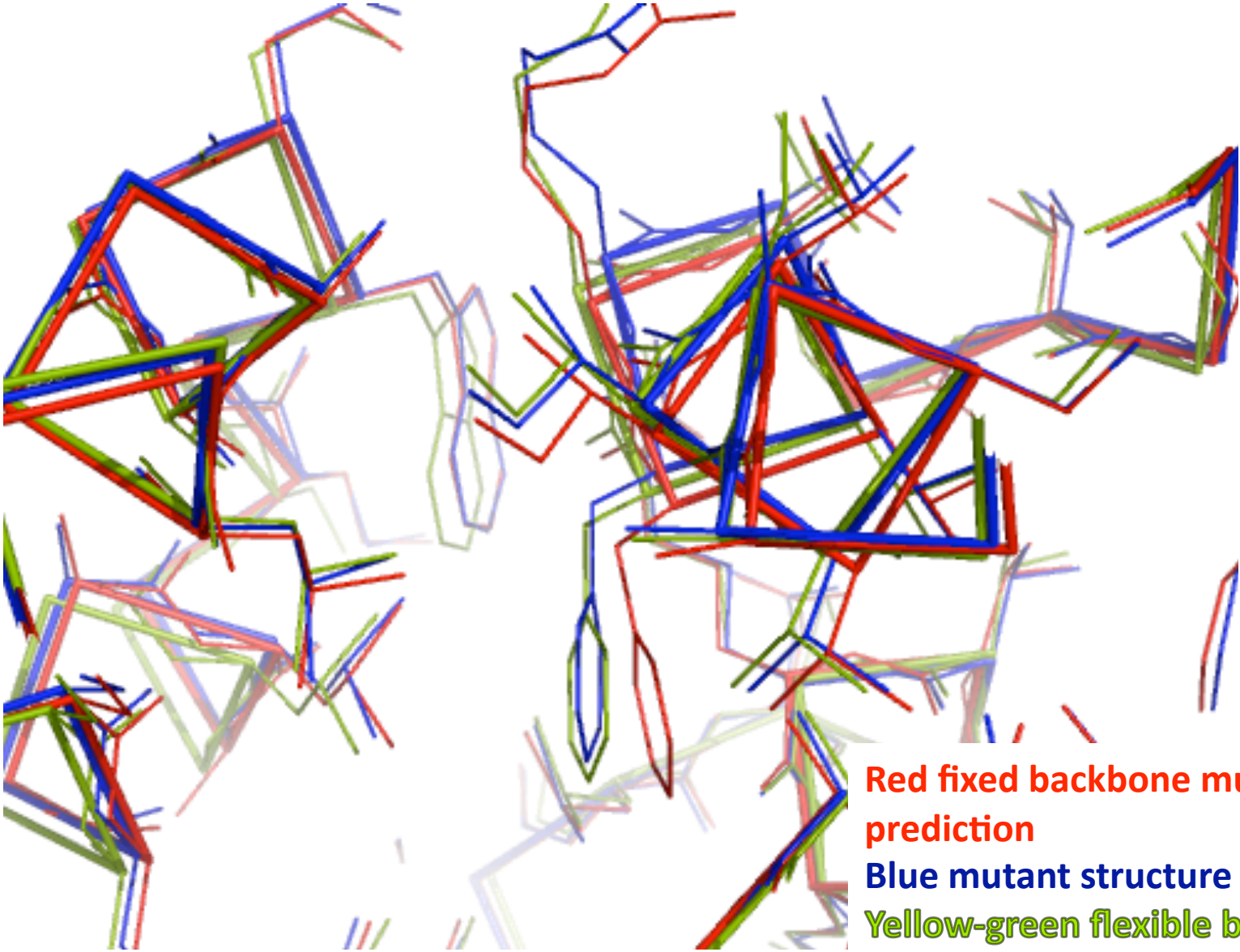
An aggressive backbone protocol produces uniform variation across the length of the protein





Superposition of wild-type ensemble of T4-lysozyme

By allowing backbone changes, we recapitulate the backbone shift upon mutation



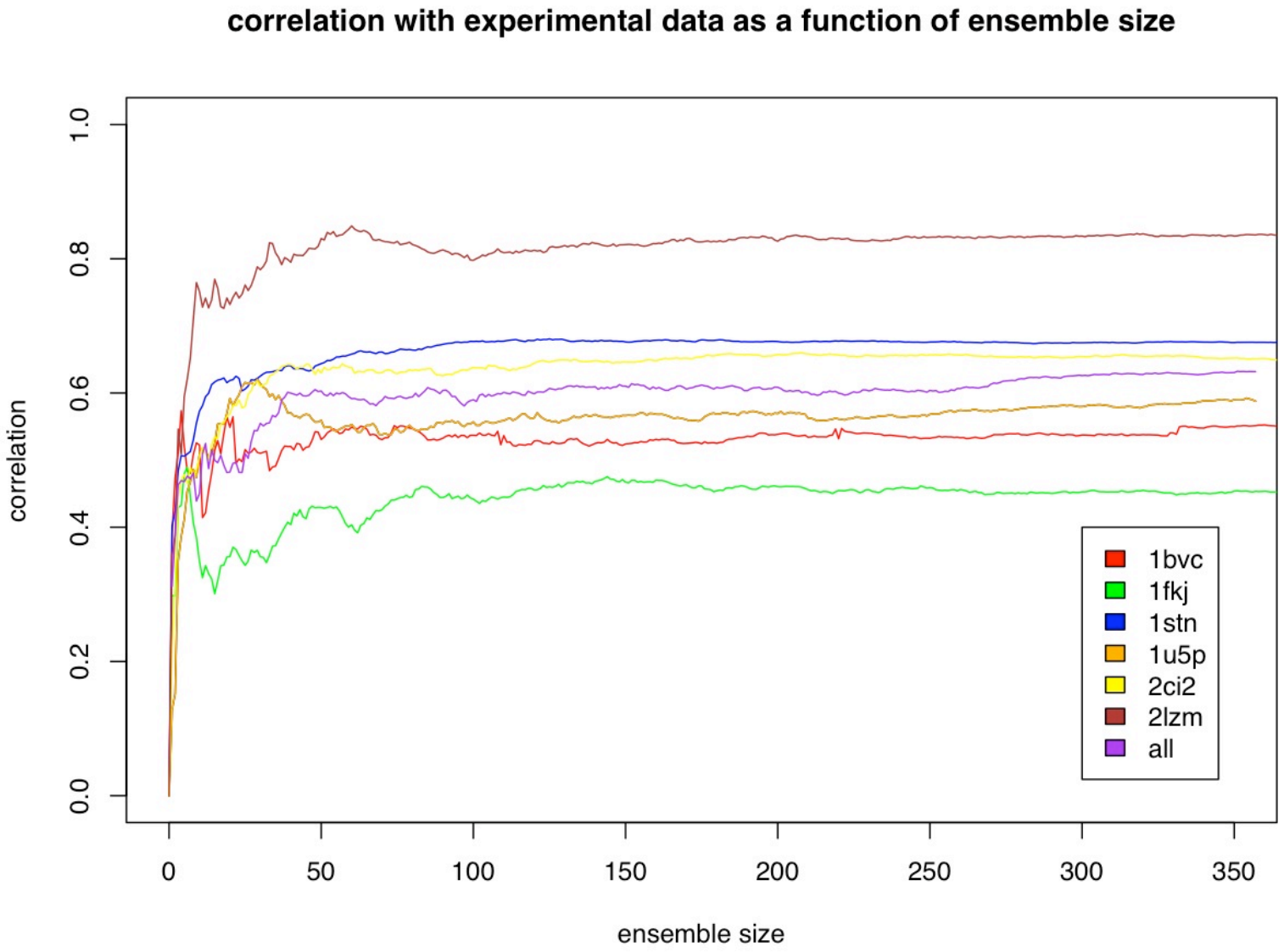
Red fixed backbone mutant prediction

Blue mutant structure

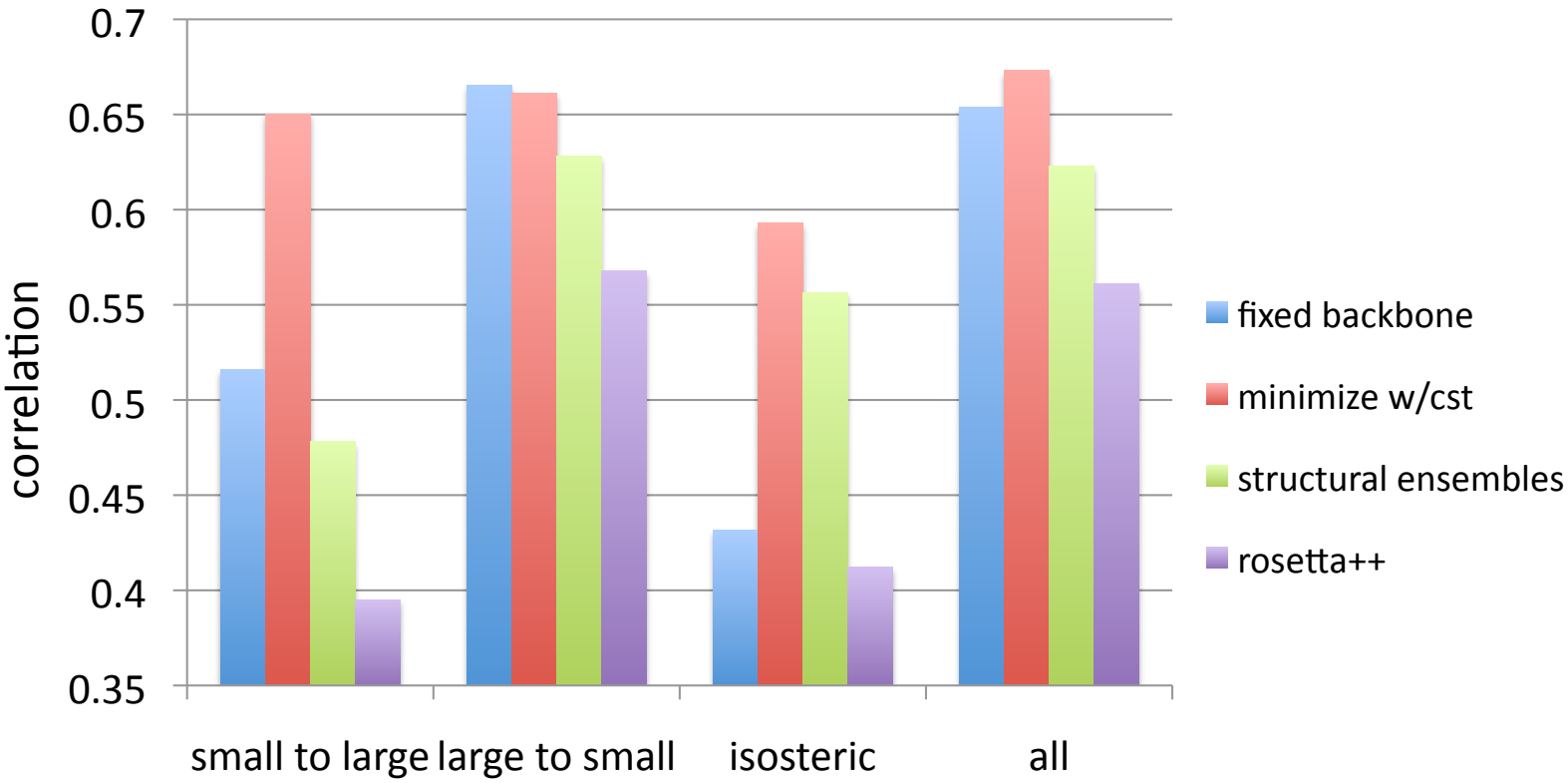
Yellow-green flexible backbone mutant prediction

T4-lysozyme A153F

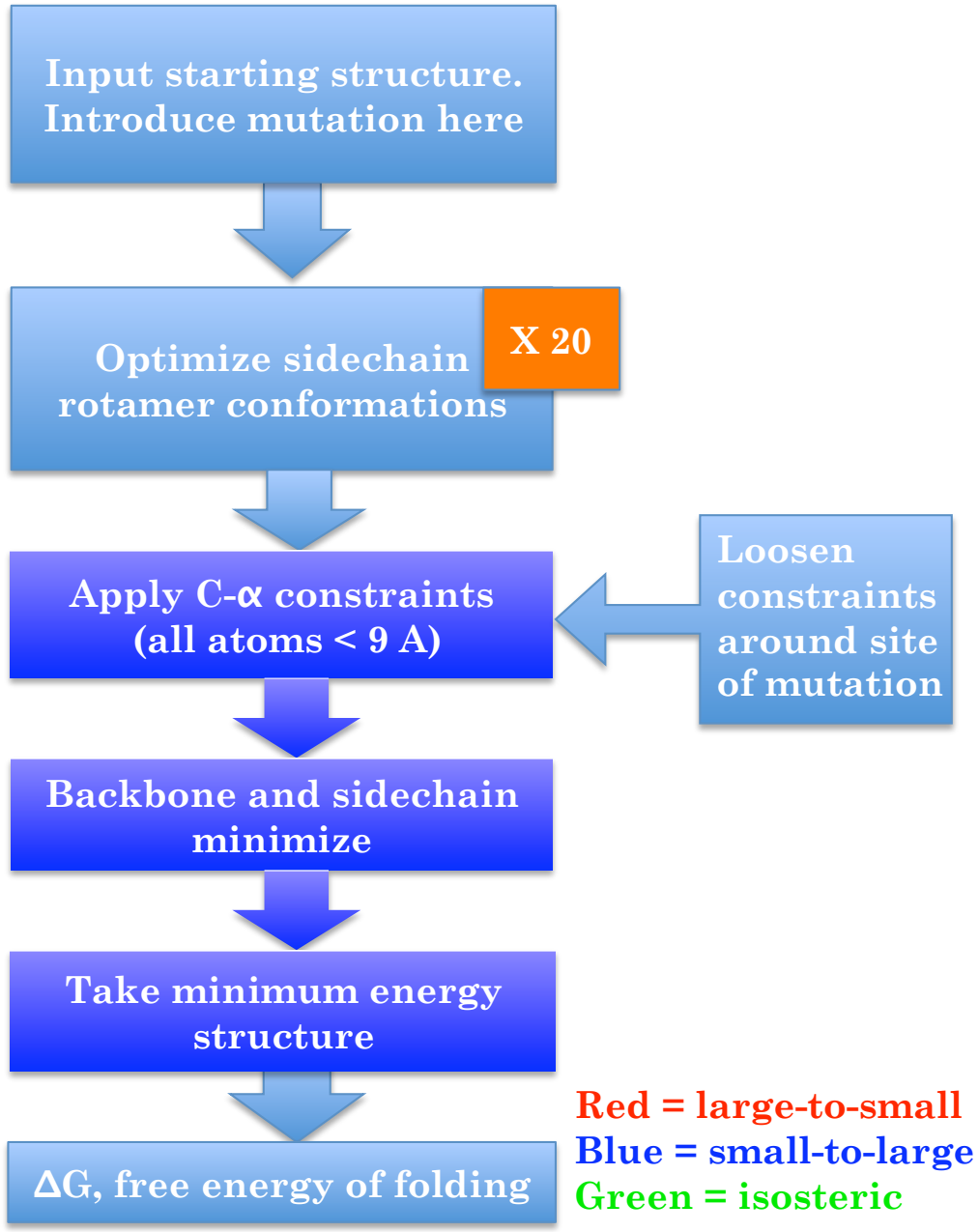
Correlation improves with increasing ensemble size



The minimization protocol performs the best of the four protocols in all size categories



To allow the structure to locally relax, we loosen constraints around the site of mutation



New constraint strategy:

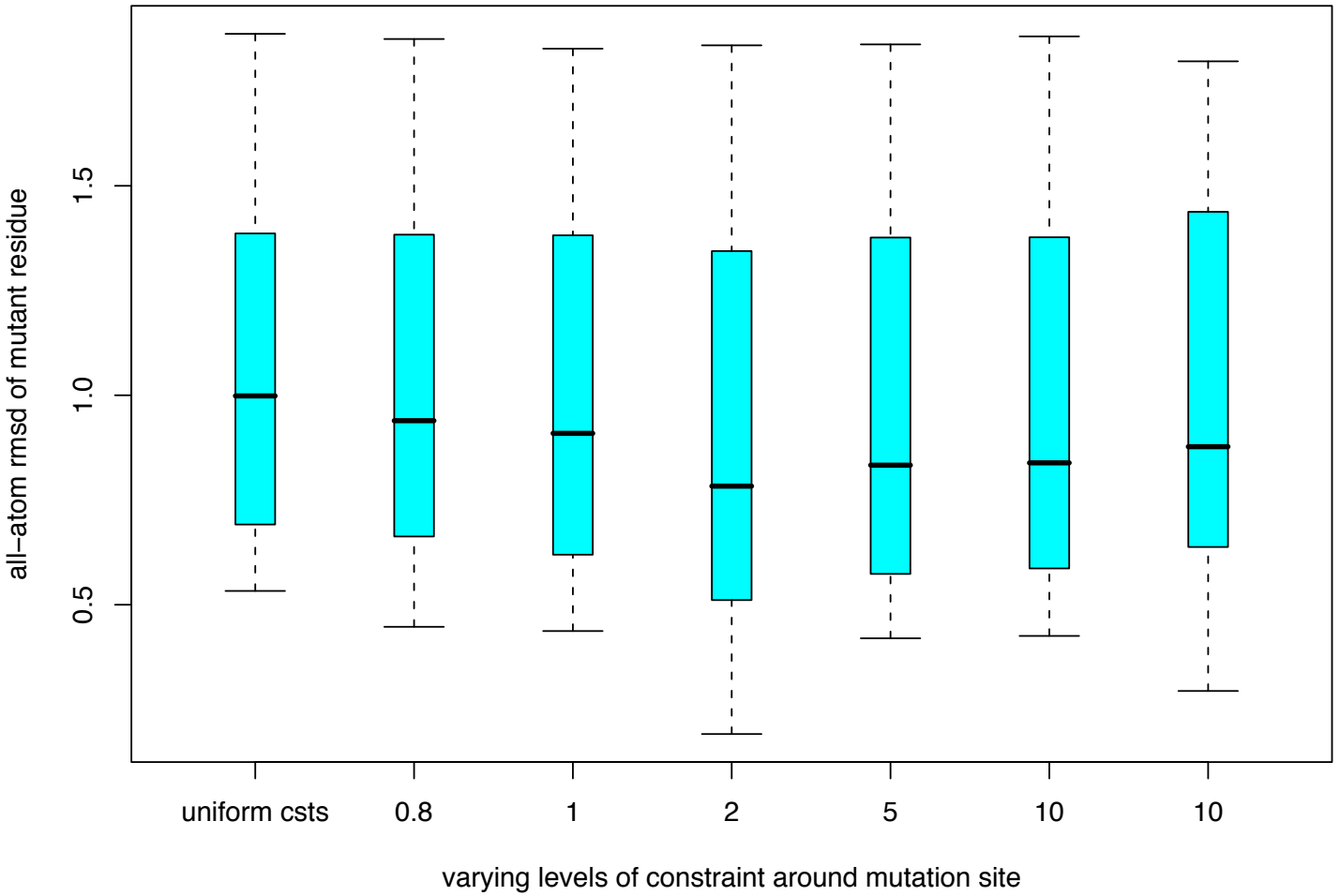
- Define neighborhood of mutation site as any sidechain center of mass lying within 8 Angstrom of mutant residue

- All constraints to any of these neighborhood residues are loosened

- All other residues have a harmonic constraint with standard deviation of 0.5

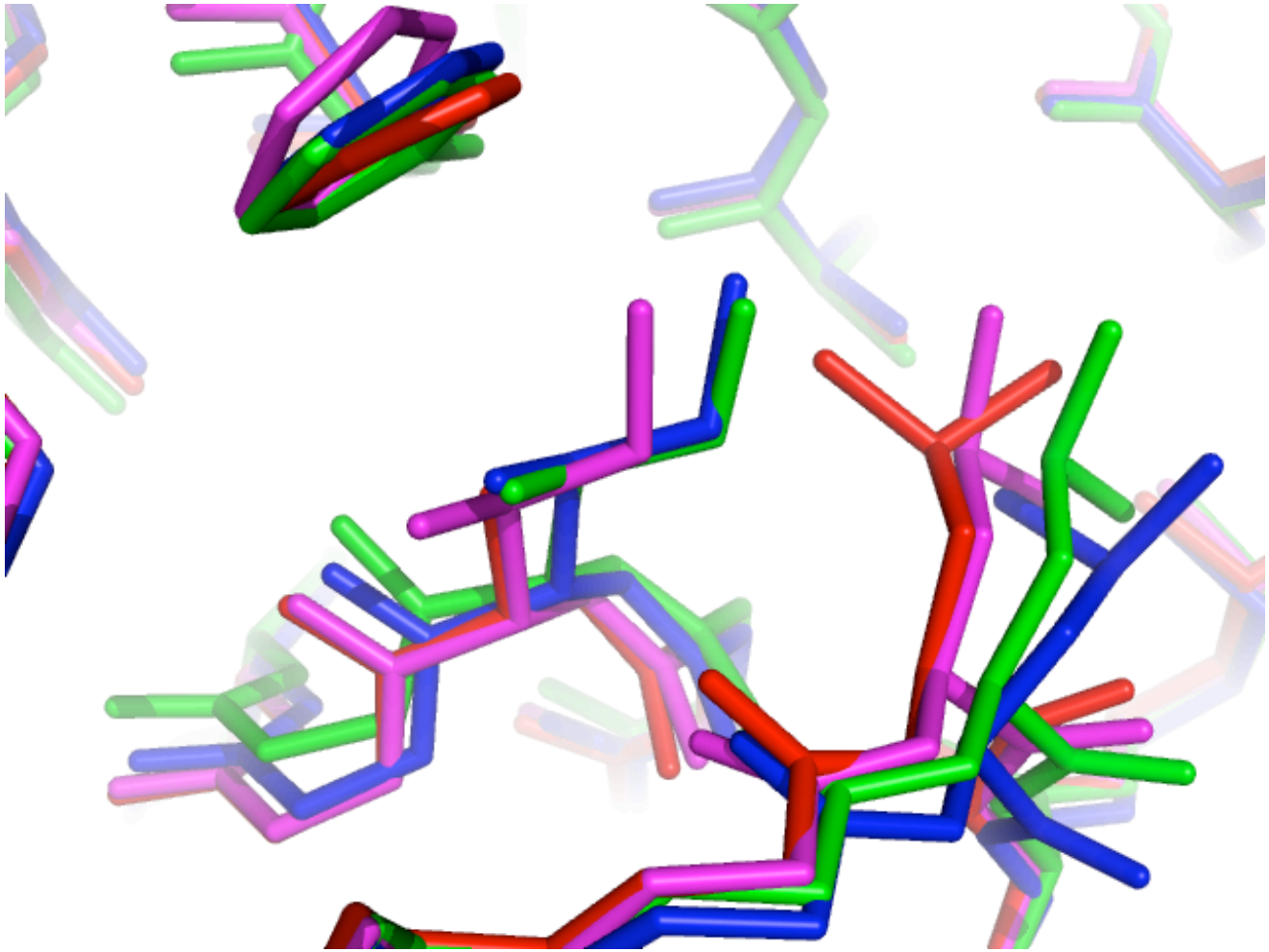
Loosening constraints improves modeling accuracy for buried small-to-large mutations

Tight constraints → Loose constraints



Buried small to large mutations (n = 11). All-atom rmsd of mutant residue calculated after optimal C-alpha superposition with mutant crystal structure

In some cases, loosening constraints around the site of mutation dramatically improves modeling results



Starting structure

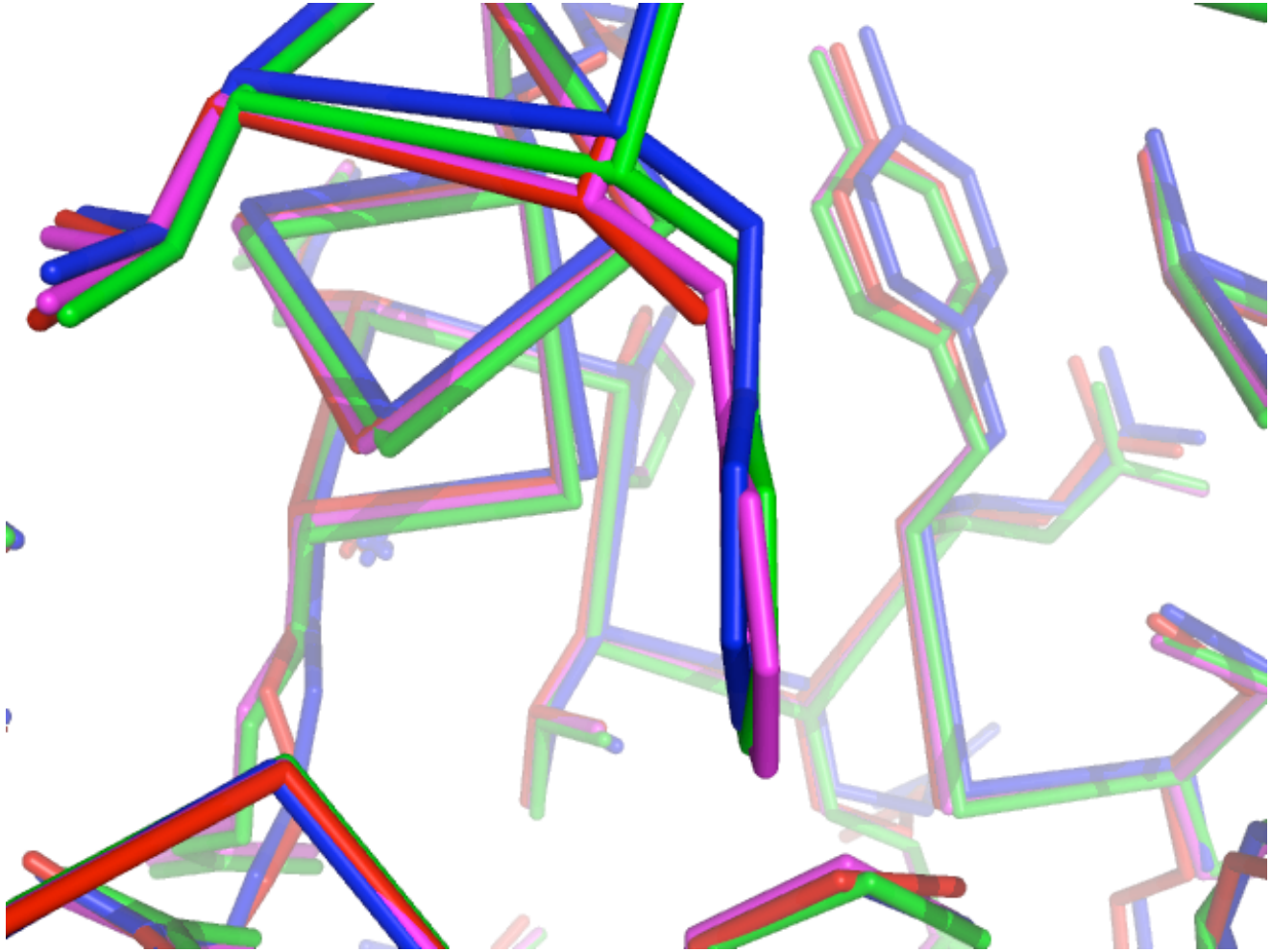
Mutant crystal structure

Min with uniform constraints

Minimization with loose-local constraints

Predicted ddg uniform constraints: 1.132
Predicted ddg loose-local constraints: -2.352
Experimentally determined value: -2.6 kcal/mol

In some cases, modeling accuracy is improved slightly



Starting structure

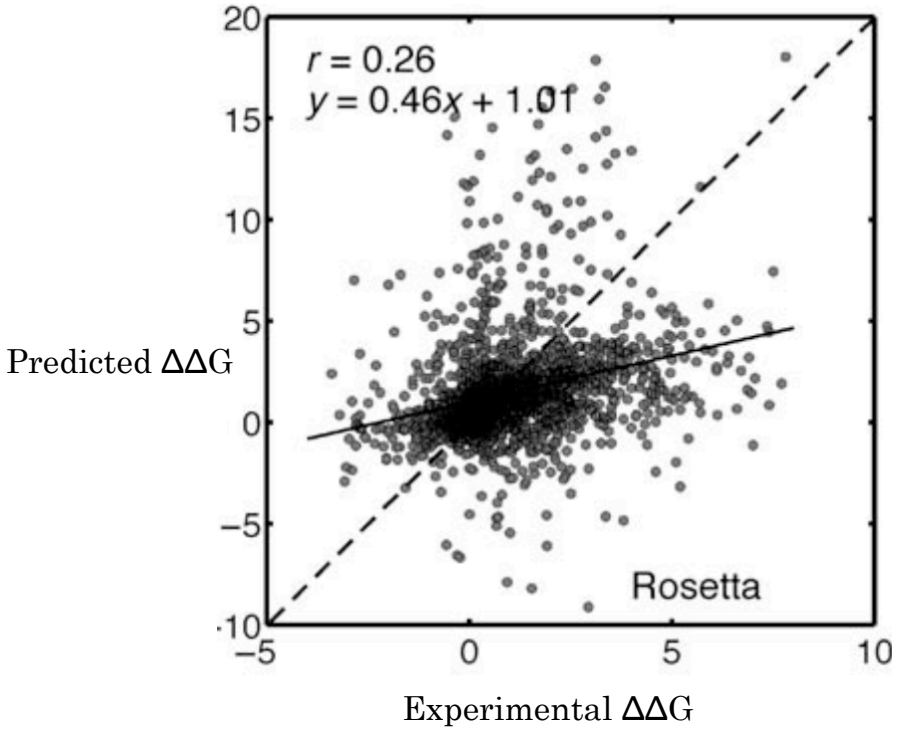
Mutant crystal structure

Min with uniform constraints

Minimization with loose-local constraints

Predicted ddg uniform constraints: -2.331
Predicted ddg loose-local constraints: -3.67
Experimental ddg: -3.8 kcal/mol

An independent assessment of computational ddg algorithms ranks rosetta as the poorest performing algorithm

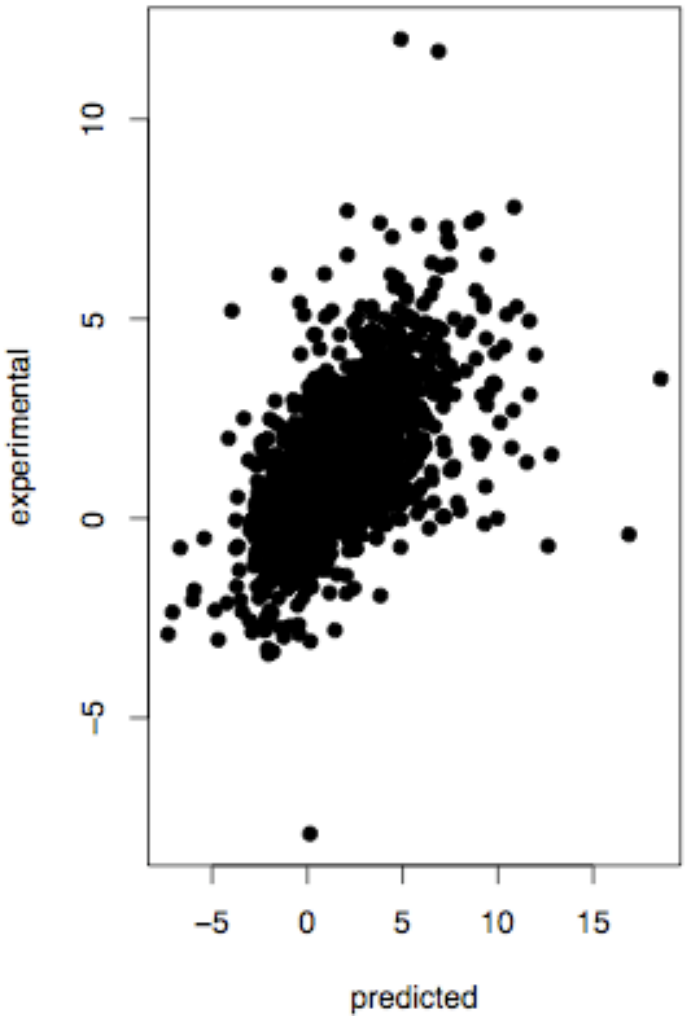


Method	r	n	outliers
CC/PBSA	0.56	478	-
EGAD	0.59	1065	1091
FoldX	0.50	1200	-
Hunter	0.45	1594	-
I-Mutant2.0	0.54	933	-
Rosetta	**0.26**	**1913**	**243**

Used fixed backbone design, All residues within 5 Å of mutation allowed to repack. Input structures were idealized and minimized.

Potapov et al. "Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details" PEDS 2009

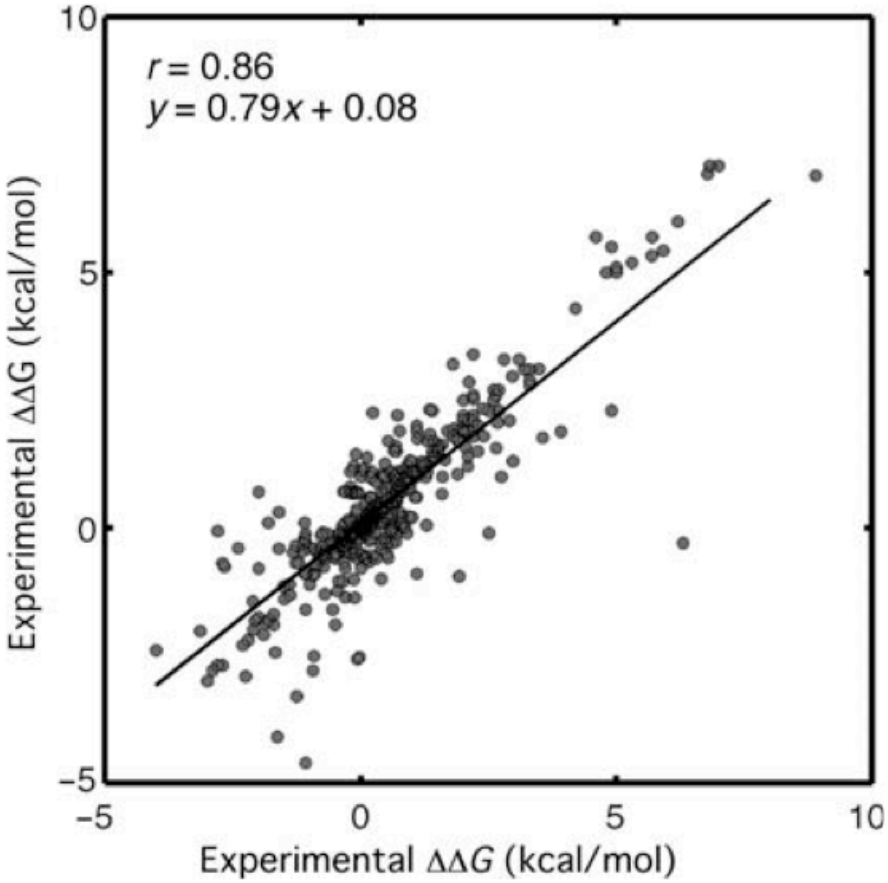
On an independent benchmark the best method ranks on par with all other methods



Method	*r*	*n*	outliers
CC/PBSA	0.56	478	-
EGAD	0.59	1065	1091
FoldX	0.50	1200	-
Hunter	0.45	1594	-
I-Mutant2.0	0.54	933	-
Rosetta	0.53	1871	0
Rosetta (remove clashes)	0.59	1857	14

Potapov et al. "Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details" PEDS 2009

The maximum correlation we can expect to achieve is around r = 0.86

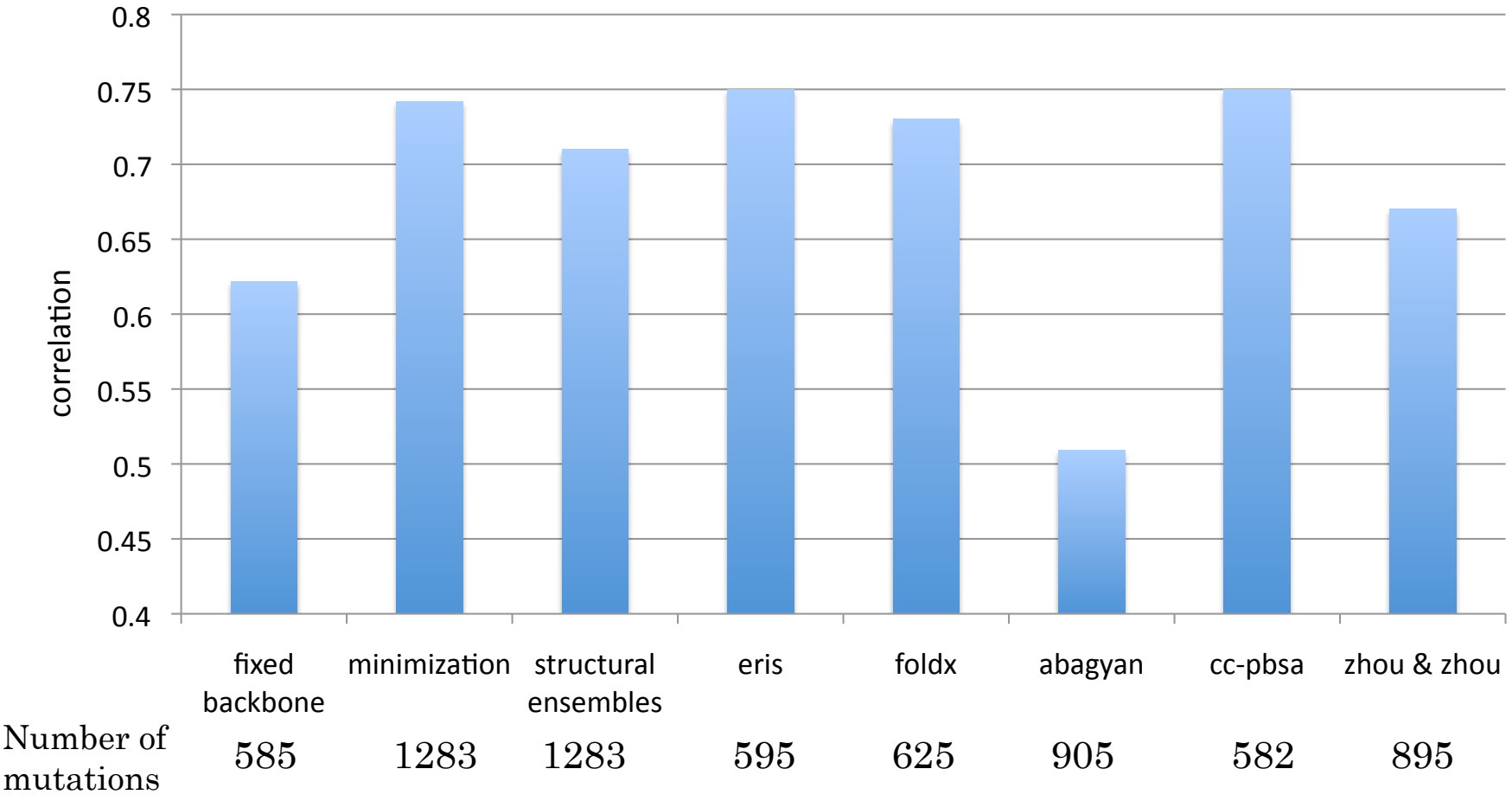


n = 406 mutations

Potapov et al. "Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details" PEDS 2009

Optimizing weights show that maximum correlation possible is 0.74, matching the results produced by other algorithms

Minimize f, where $f = \sum_{\text{mutation}\ \ i=1\ldots N} \left(\Delta\Delta G_{experimental} - \left(\sum_{j} w_j \Delta E_{ij} + \Delta ref_i \right) \right)^2$



Conclusions/Future work

- Protocol adjustments without weight optimization is sufficient to produce a 0.67 correlation on a comprehensive set of 1,287 mutations.
- Further weight optimization can increase correlation with experimental data to 0.74
- Judicial choice of constraints can increase accuracy of predicting mutant sidechain conformations

Future Work:

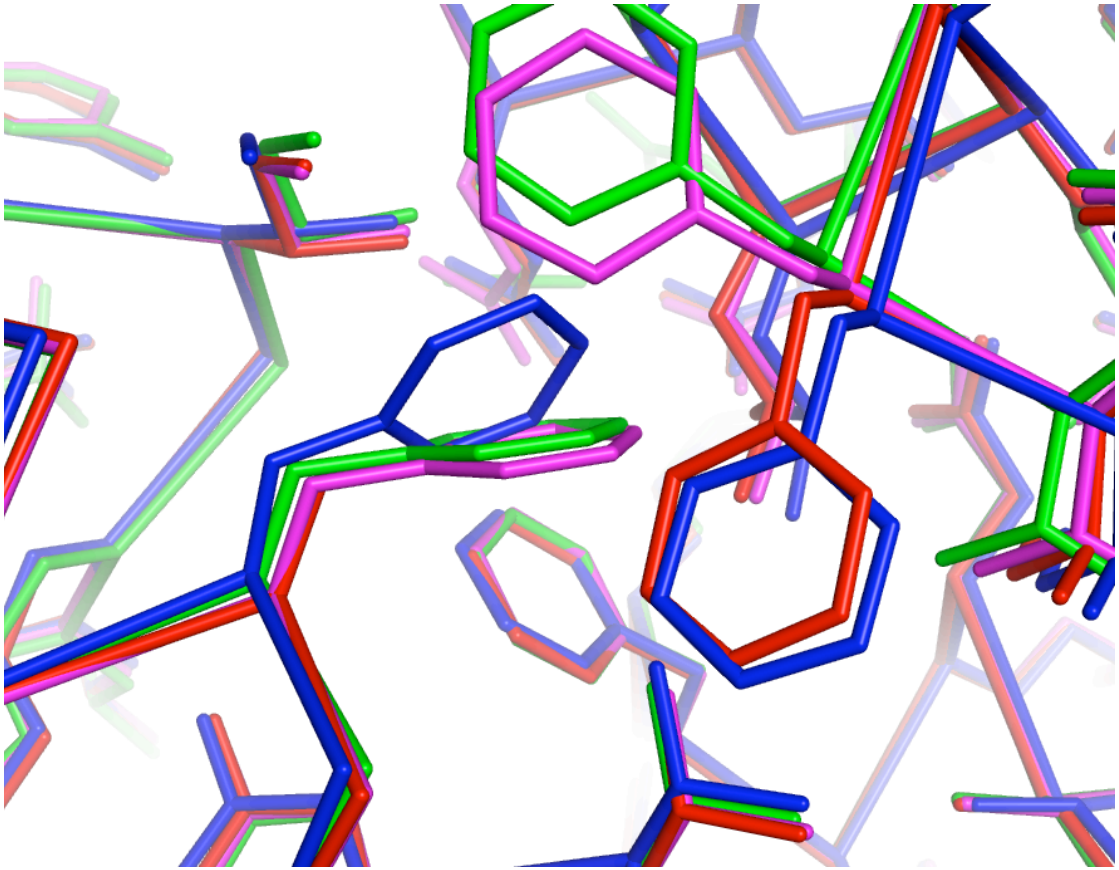
- Cross-validation of weight optimization

Acknowledgements

Andrew Leaver-Fay
David Baker
The Baker Lab

Funding:
Genome Training Grant

And in other cases, loosening constraints around the site of mutation does not increase modeling accuracy



Starting structure

Mutant crystal structure

Min with uniform constraints

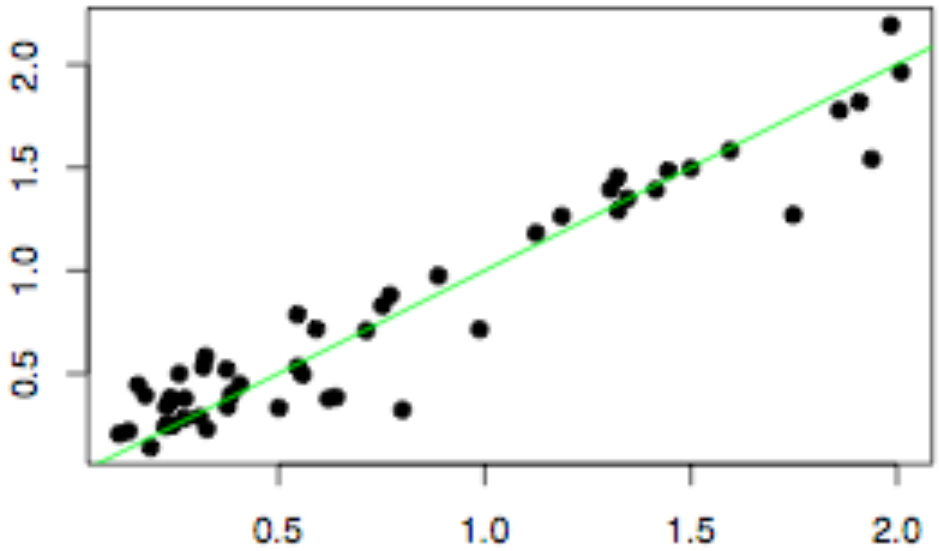
Minimization with loose-local constraints

outline

- Data-set: curation
- Fixed bb results
- Flexible bb results
  - Backrub (make slides but put at end in case anyone asks)
  - Ensembles (don't talk about this?)
  - Minimization
  - Local minimization
- Energetic results
- Structural results: tying energy and structure together?
  - In some cases we can predict significant backbone changes very well
  - In other cases, our energy function is unable to guide sampling to the correct answer, and does not predict the ddg as well
- An independent test:
  - Introduce potapov set, explain results
  - Show results of running algorithm on this set
- Opte results: what is the upper limit for our combination of scoring and sampling?

As well as lower rmsd predictions

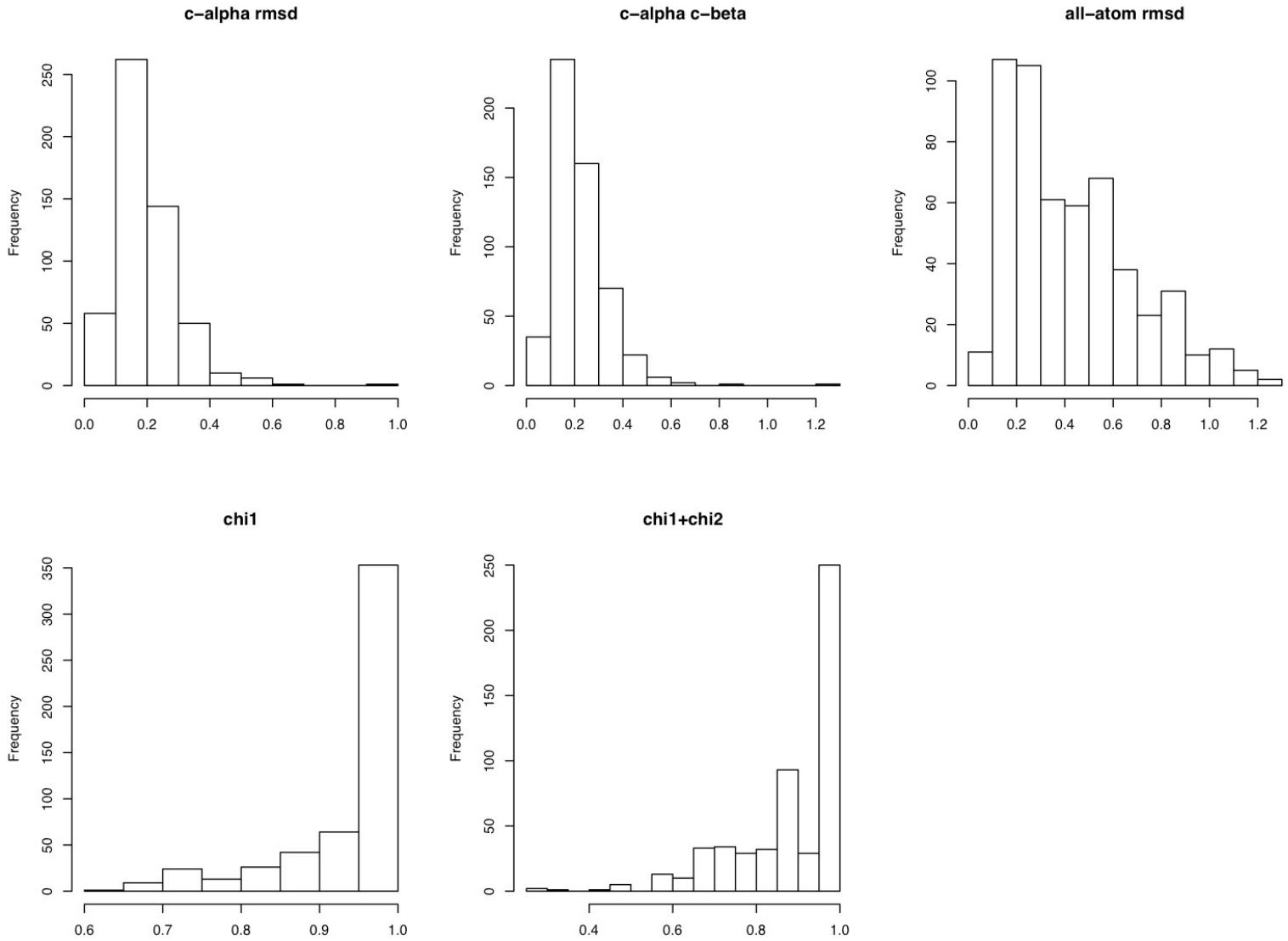
Rmsd to mutant xstal, loose-local constraints



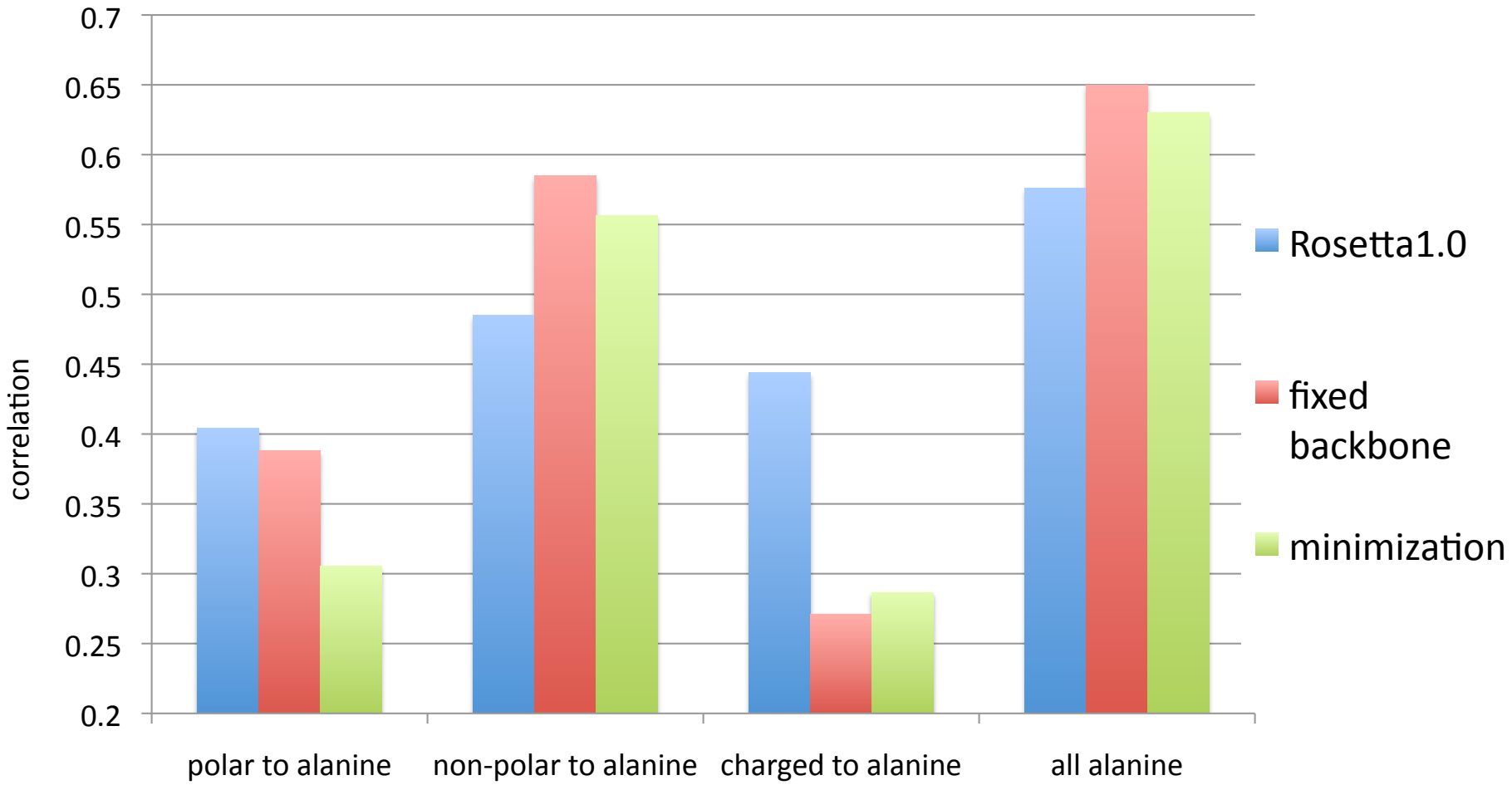
Rmsd to mutant xstal, uniform constraints

Rmsd measured as all-atom rmsd at site of mutation

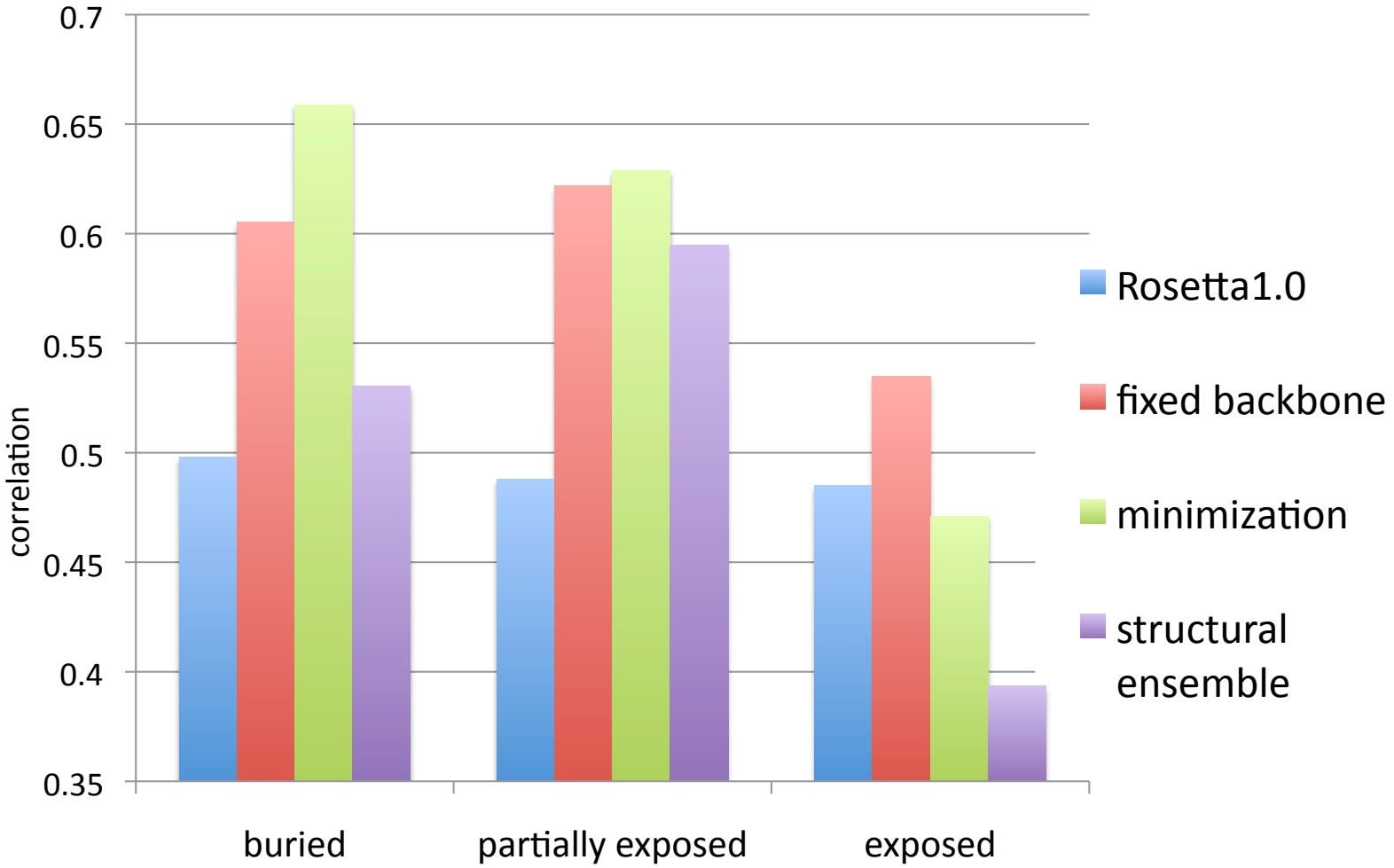
Examining wt and mutant crystal structures show that most of the time structures essentially stay the same, but also can change significantly



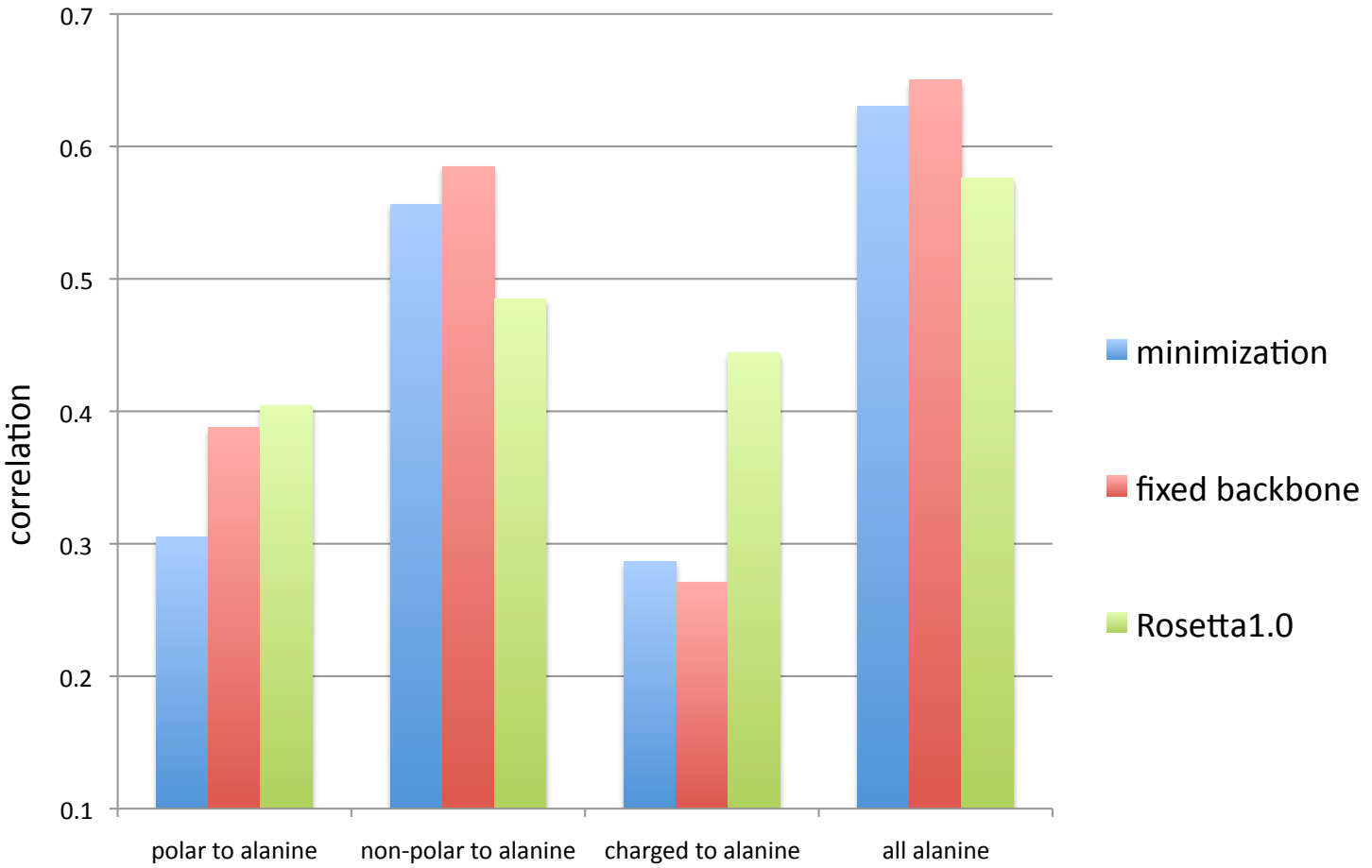
Mutations involving polar or charged residues are predicted less accurately



Buried residues are predicted more accurately than exposed residues



Mutations involving only non-polar residues are predicted more accurately than those involving polar or charged residues

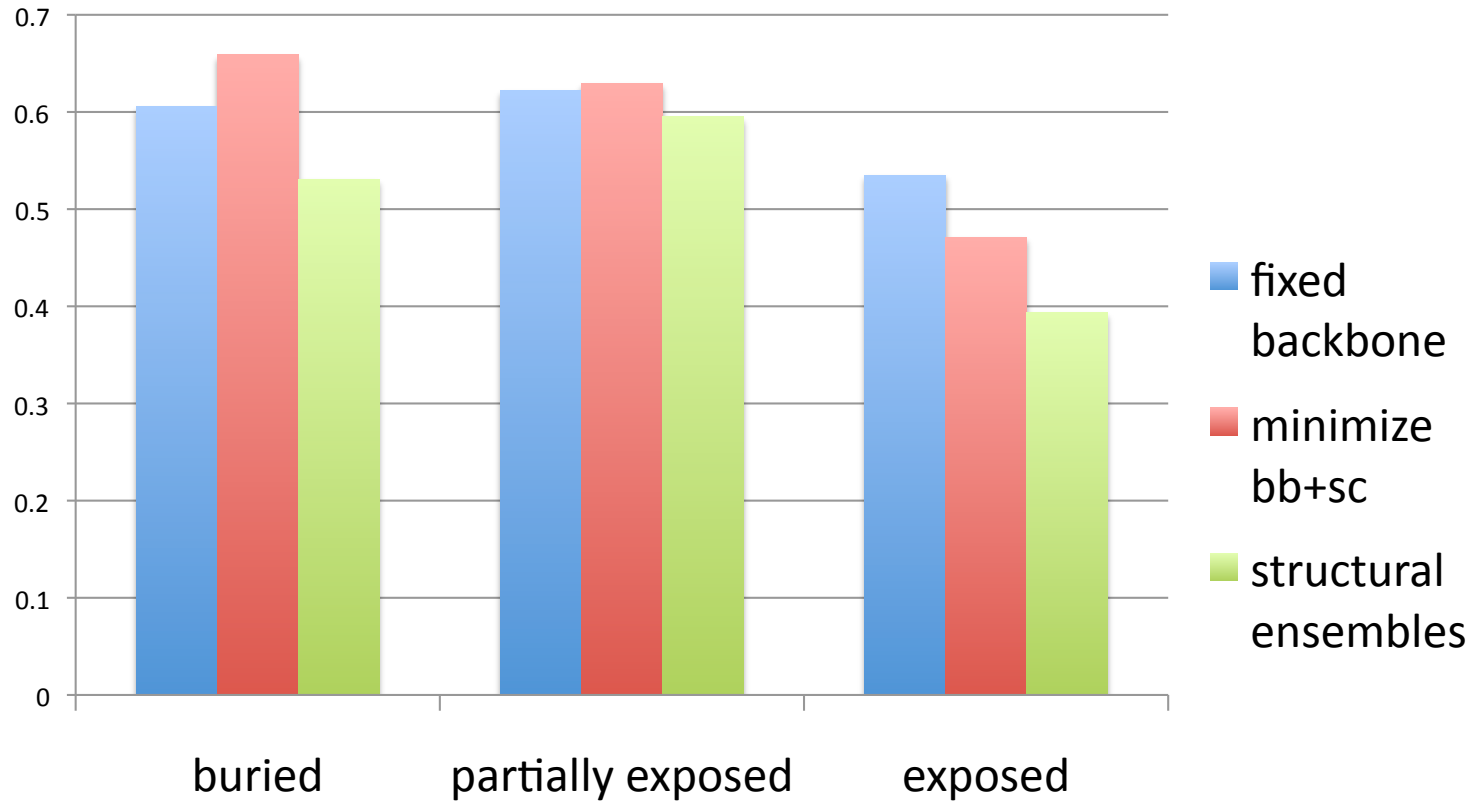


Structural changes upon mutation are not well understood or predicted

- Often mutations cause no structural changes at all
- However, mutations can cause large backbone shifts and structural re-arrangements
- Most mutations affect the local conformation of the protein structure
- An experimental "CASP" for small sequence changes was held in October 2007 to assess the current state of the art.

Xu, Baase, Baldwin et al. Protein Sci **1998** 1:158-77
Eriksson, Baase, et al. J. Mol. Biol. **1993** 3:747-69
Baldwin, Xu et al. J. Mol. Biol. **1996** 3:542-59
Fulton et al. Biochemistry **2003** 42: 2364-72
Cuneo , et al. BMC Struc Biol. **2008** 8:20

Separation by solvent exposure shows that buried residues are predicted accurately, but exposed residues are not.

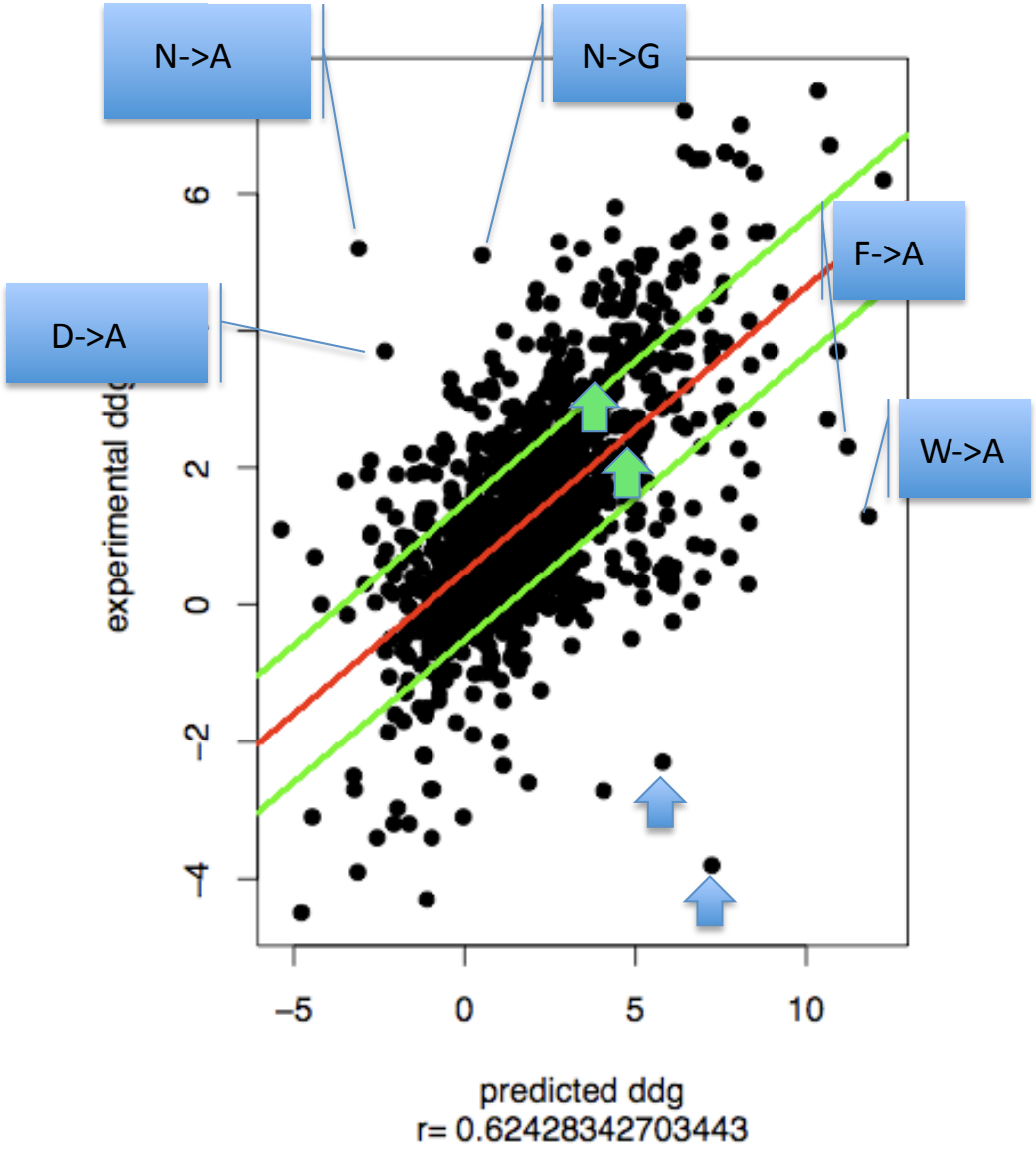


	#	buried	#	partially exposed	#	exposed
fixed backbone	416	0.6054	415	0.6217	452	0.535
minimization	416	0.6587	415	0.6289	452	0.47111
structural ensembles	171	0.5304	183	0.59489	232	0.39343

- Ddgs:
  - Important for design
  - Mutations measuring thermodynamic stability are time-consuming experiments, could use computational to speed it up
- What has been done (previous algorithms)
- What was done previously in rosetta++ (and limitations of the algorithm)
- New algorithm: fixed backbone allow sidechain optimization. Used different energy function to reduce effect of steric clashes
- This still doesn't full reduce the effect of steric clashes
- Then, introduce backbone and sidechain minimization
- This works much better, but real mutations can cause large backbone shifts in structure, and our minimization protocol would not capture this currently.
- So introduce a new protocol which would randomly sample the structure space surrounding the original structure very closely
- Comparison of the three protocols side-by-side: minimization does the best out of the three
- But this isn't the best we could do. In order to further increase our performance we can optimize the weights
- Furthermore, the weights indicate the best we could do with the current protocol in the context of the current scorefunction. What can we do now to improve our algorithm?
- We could look at outliers and try to find systemic failures in our energy function.
- What energy function issues have come up so far?
- Observation that buried polar residues are usually predicted more stabilizing than they actually are, and exposed are more destabilizing than they actually are
- Also intra-residue hydrogen bonds
- Electrostatics and the work you've done towards modeling through hackelec?

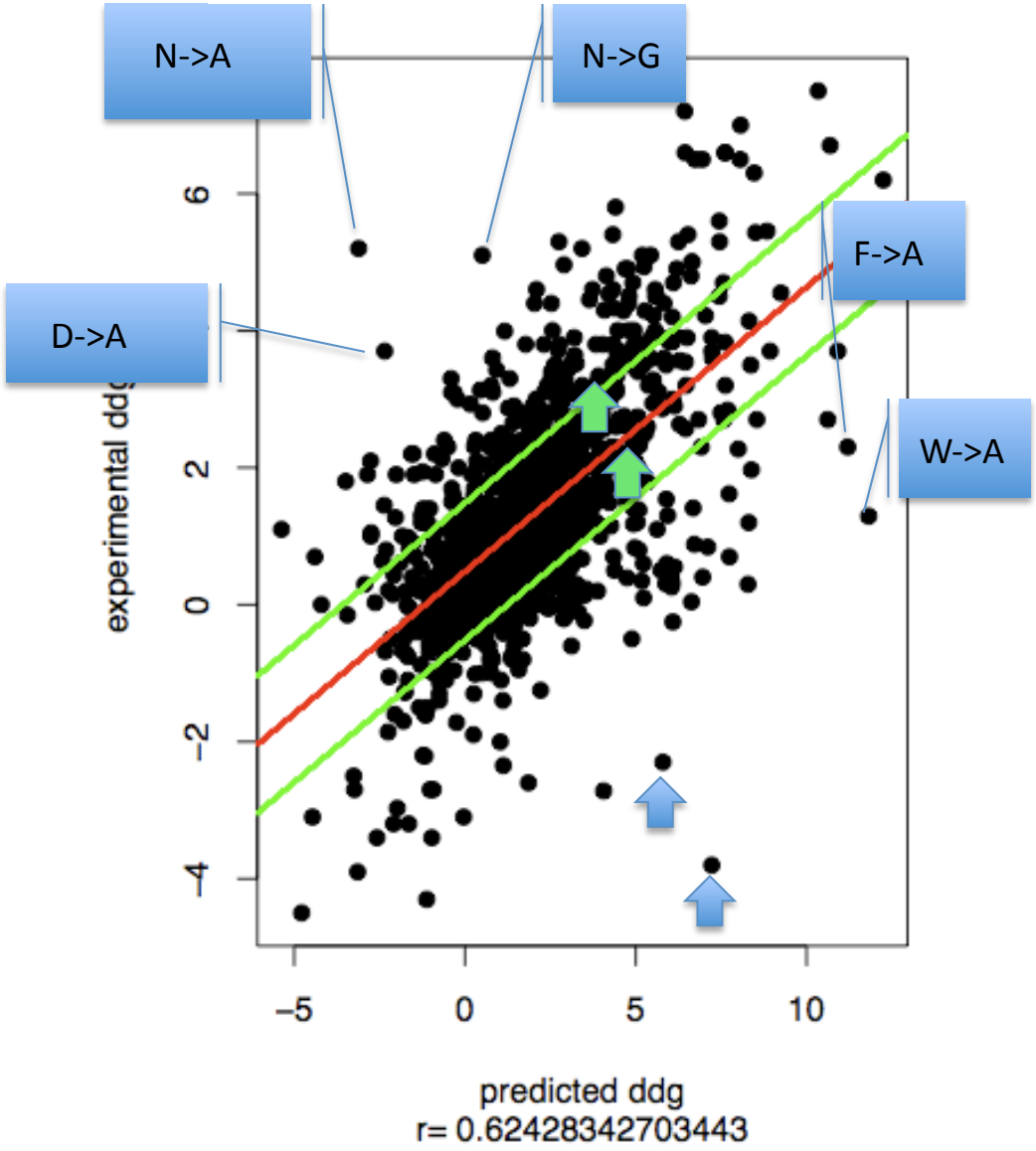
In some cases, we can accurately predict mutations in one way, but not the other

- Green arrows:
  - T4-lysozyme(wildtype):
    - F* 153 A
    - L* 121 A
- Blue arrows:
  - T4-lysozyme(mutant):
    - A 153 F*
    - A 121 L*

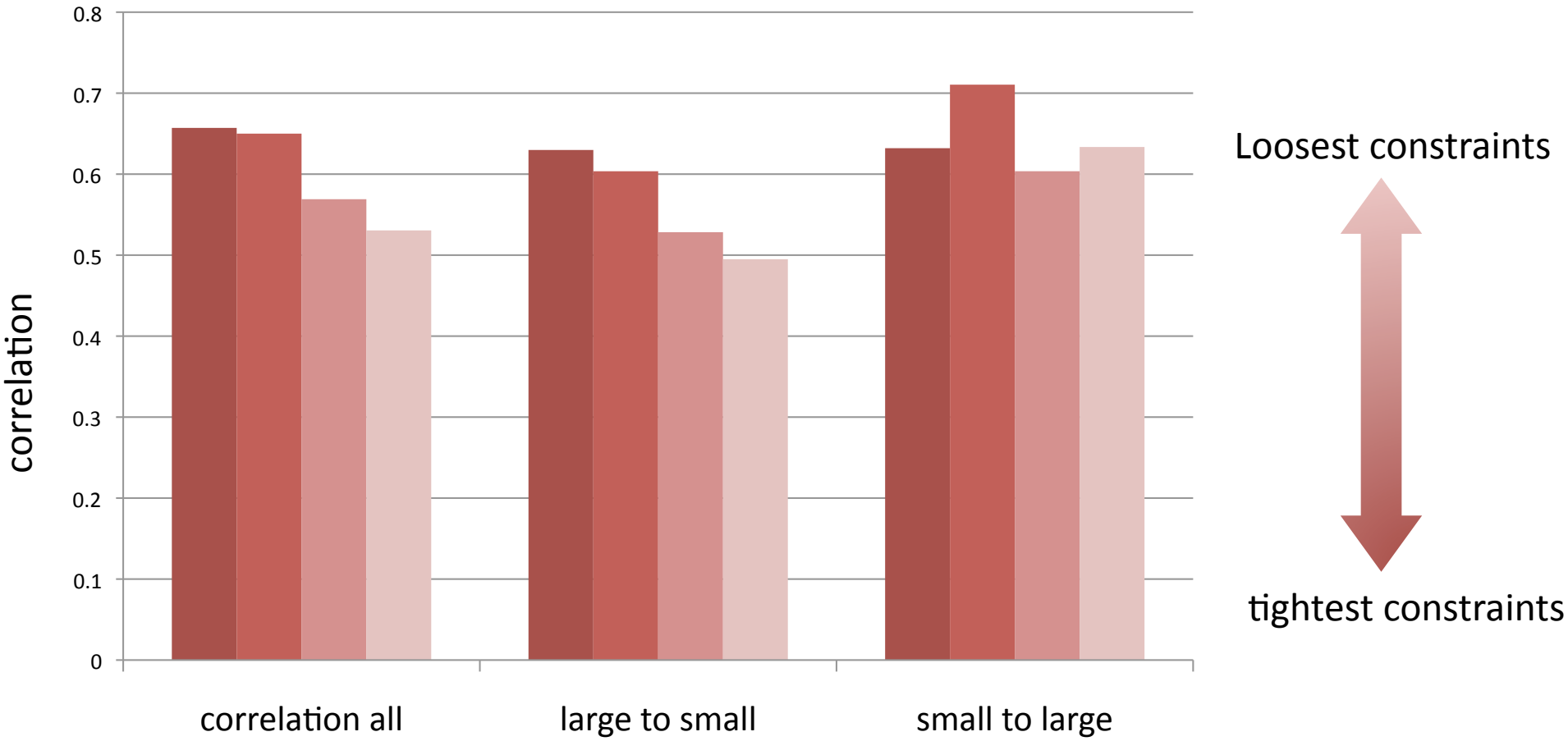


In some cases, we can accurately predict mutations in one way, but not the other

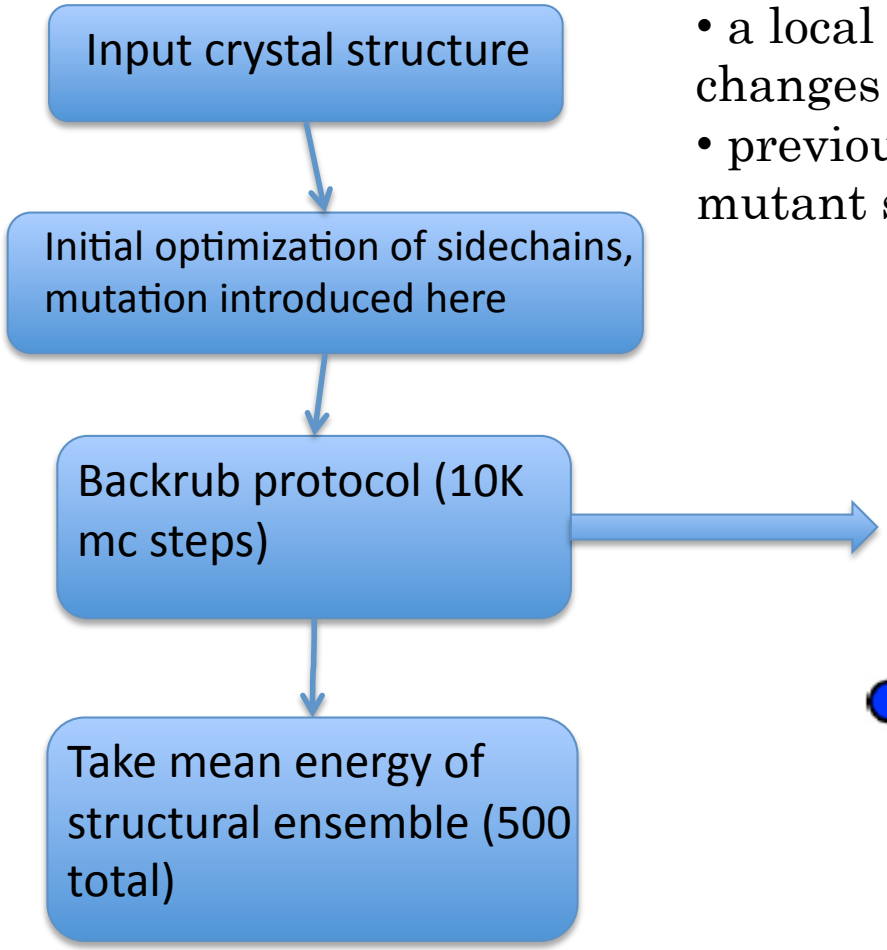
- Green arrows:
  - T4-lysozyme(wildtype):
    - F* 153 A
    - L* 121 A
- Blue arrows:
  - T4-lysozyme(mutant):
    - A 153 F*
    - A 121 L*



Allowing more movement around the site of mutation produces better energy predictions…

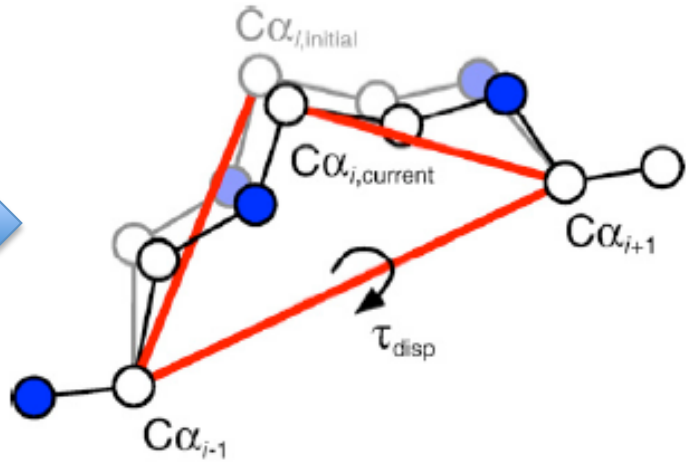


Backrub protocol



Advantages:

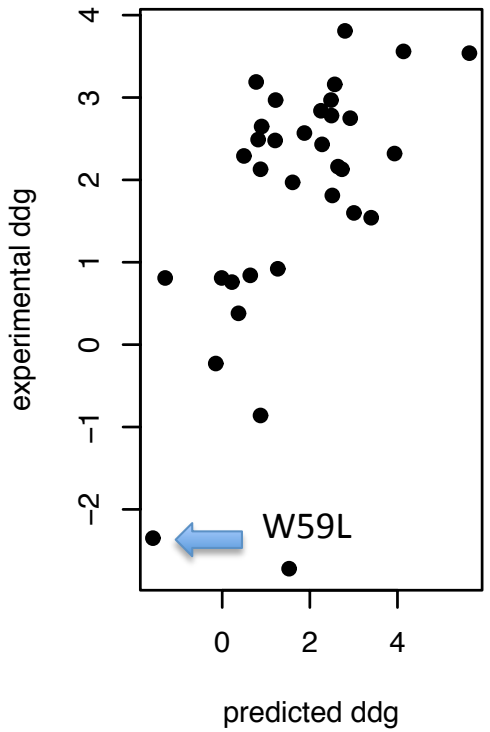
- a local move that does not propagate changes in the protein structure
- previously shown to recapitulate mutant sidechain conformations



Smith CA, Kortemme T. "*Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction.*" J Mol Biol. 2008 Jul 18;380(4):742-56.

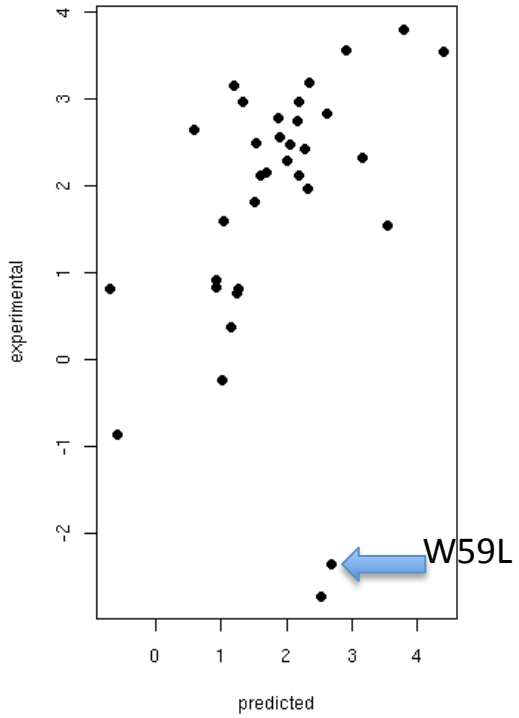
Using a backrub ensemble can improve correlation over using a single-structure fixed backbone protocol

Ensemble of structures generated using the backrub protocol



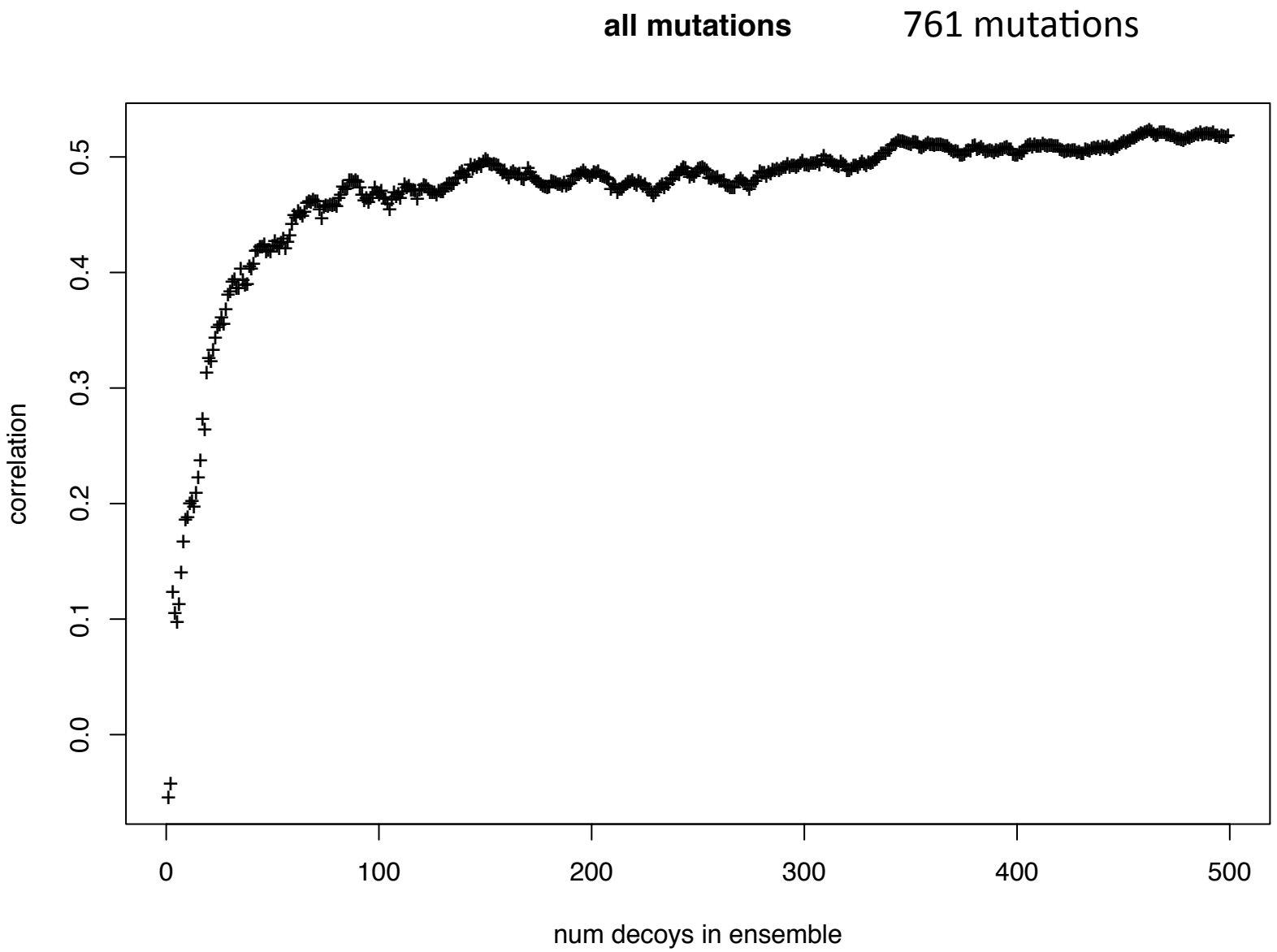
R = 0.58

Single-structure fixed backbone protocol

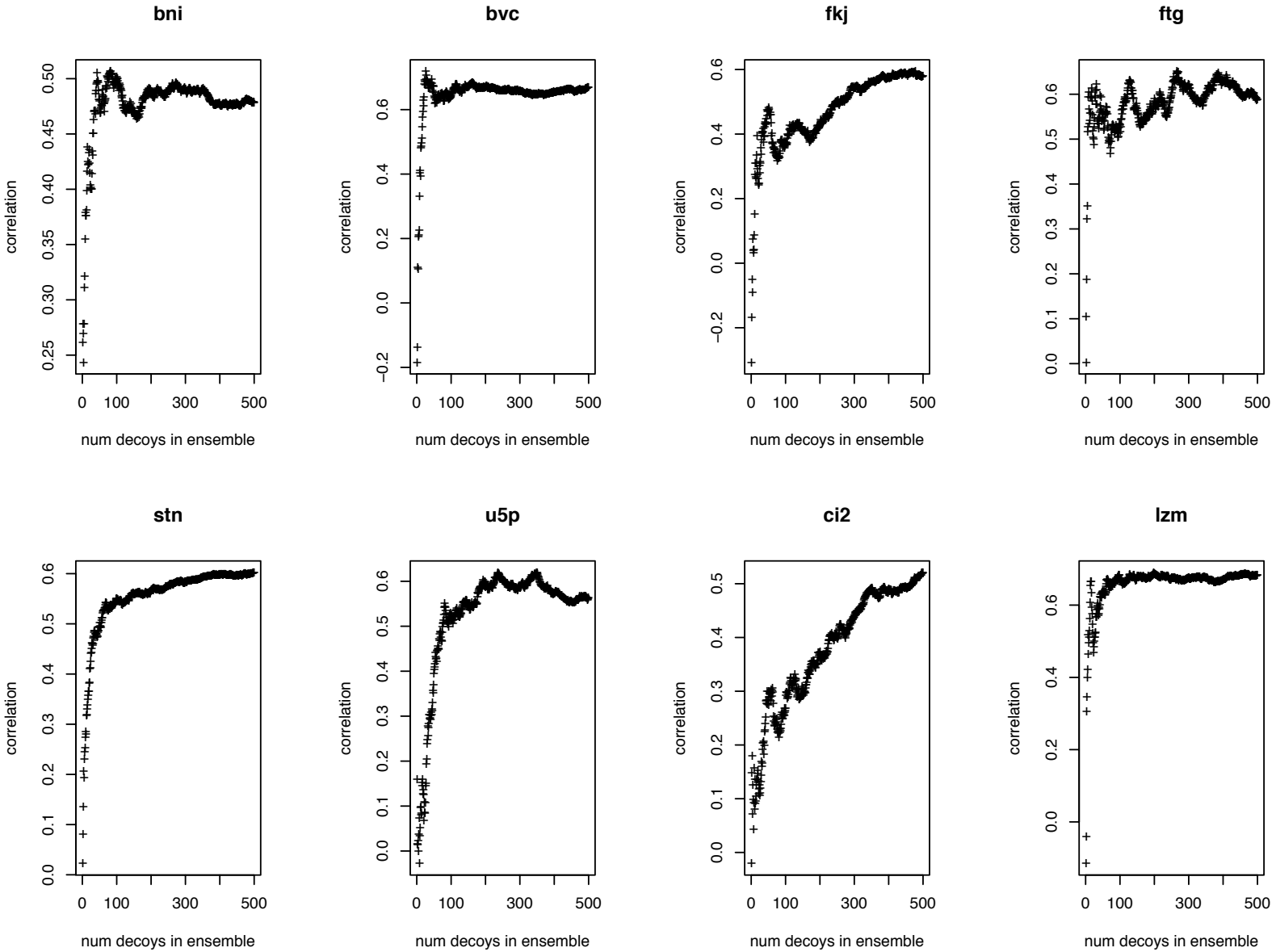


R = 0.34

The more decoys in the ensemble, the better the correlation



Correlation for individual proteins seems to generally improve with inclusion of more decoys in the ensemble.



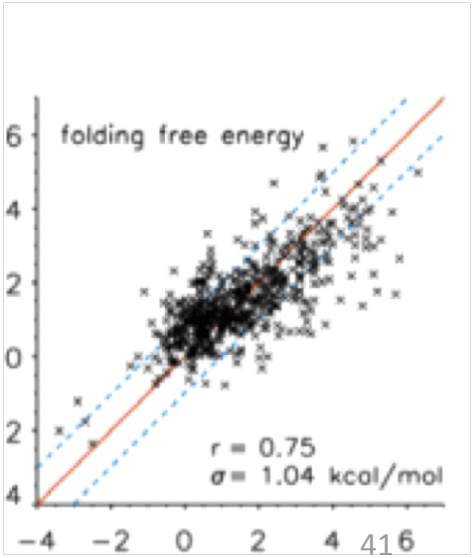
Previous work in estimating ddgs using flexible backbone protocols

Eris:

- Detected strain and relaxed dihedrals accordingly
- Trained reference weights to more closely match experimental ddGs
- On a set of 595 mutations, pearson R = 0.66

CC/PBSA:

- Produced structural ensembles with distance constraints and minimized with GROMACS
- Trained physics-based energy function based on structural ensemble
- Pearson R = 0.75 over 582 mutations



Benedix et al. "*Prediction of Mutational Free Energy Changes Using Ensembles of Structures*", Nature Methods.(2009) 6: 3-4
Yin et al. "*Modeling Backbone Flexibility Improves Protein Stability Estimation*" Structure (2007)15: 1567-1576

Previous work in fixed backbone protocols

Foldx:

- Mutations modeled with WHATIF, energy function based on physical terms was trained on a set of 339 mutants with a pearson R = 0.7
- On 625 single mutations a pearson R = 0.73

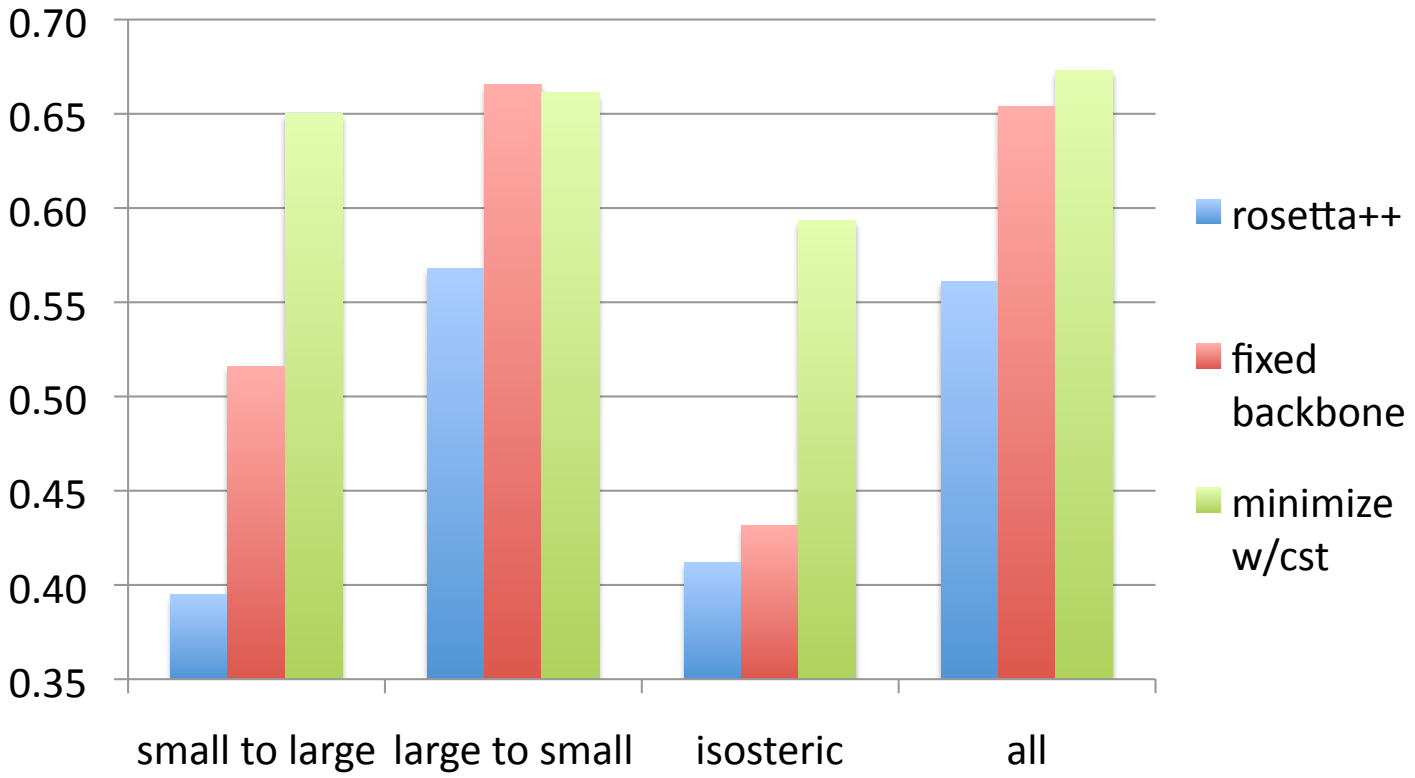
Guerois R. et al. "*Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations*", JMB (2002): 320, 369-387

Previous work from the Baker Lab in estimating ddgs using fixed backbone protocols

- Fixed backbone protocol, allow sidechain re-arrangements
- fit weights using monomeric protein alanine scanning data
- 743 X-> A mutations in monomeric proteins with pearson R = 0.75
- 1,584 mutations X -> Y where Y is smaller or same size as X yields pearson R = 0.70

Kortemme T., and Baker D., "*A Simple Physical Model for Binding Energy Hot Spots in Protein-Protein Complexes*", PNAS., 2002 (99)22:14116-14121

Separating mutations based on size differences shows that the minimization protocol predicts non-conservative mutations best



	small to large	large to small	isosteric	all
rosetta++	0.40	0.57	0.41	0.56
fixed backbone	0.52	0.67	0.43	0.65
minimize w/cst	0.65	0.66	0.59	0.67

Optimization of weights drastically improves correlation

$$\text{Minimize } f, \text{ where } f = \sum_{\text{mutation}\ \ i=1\ldots N} \left(\Delta\Delta G_{experimental} - \left(\sum_{j} w_j \Delta E_{ij} + \Delta ref_i \right) \right)^2$$