An Efficient Sampling Method for Fragment-based Protein Structure Prediction

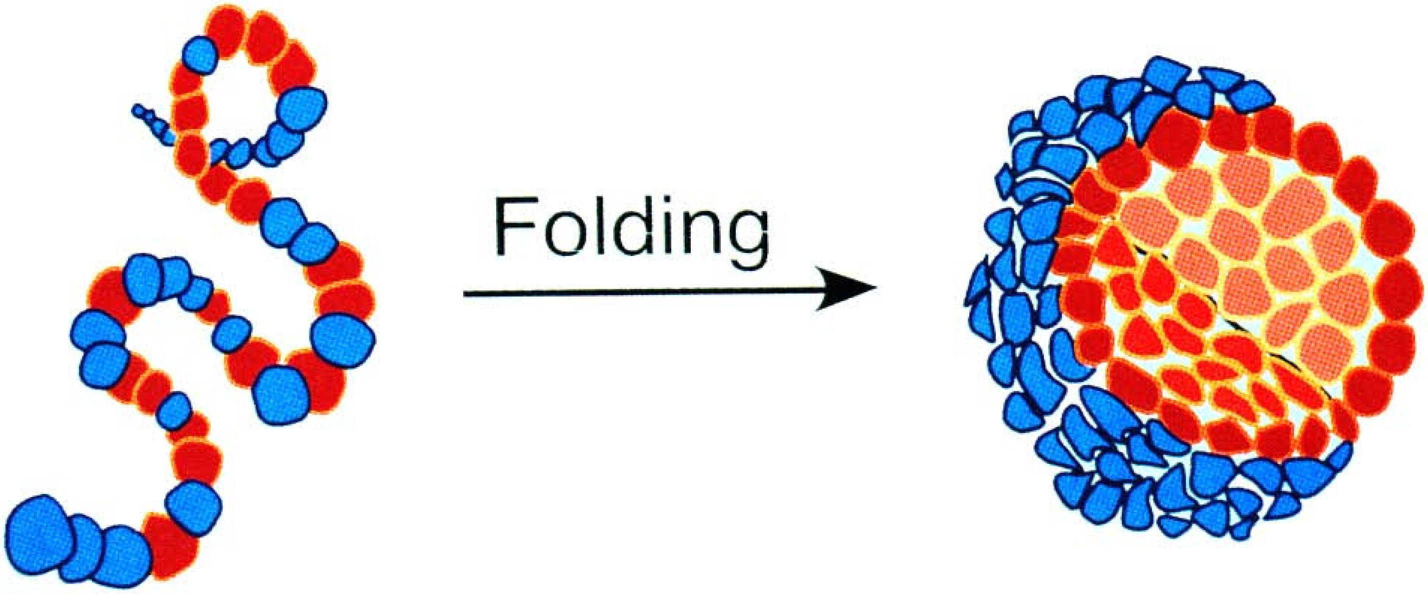
Kam Zhang, Ph.D.
Zhang Initiative Research Unit
Advanced Science Institute
RIKEN

The Protein Folding Problem

- **Anfinsen's Folding Postulate**: *The information needed to specify the complex three-dimensional structure of a protein is contained in its amino acid sequence*. – Anfinsen et al. (1961) *PNAS*, **47**, 1309-1314.

Hydrophobic residues

Hydrophilic residues



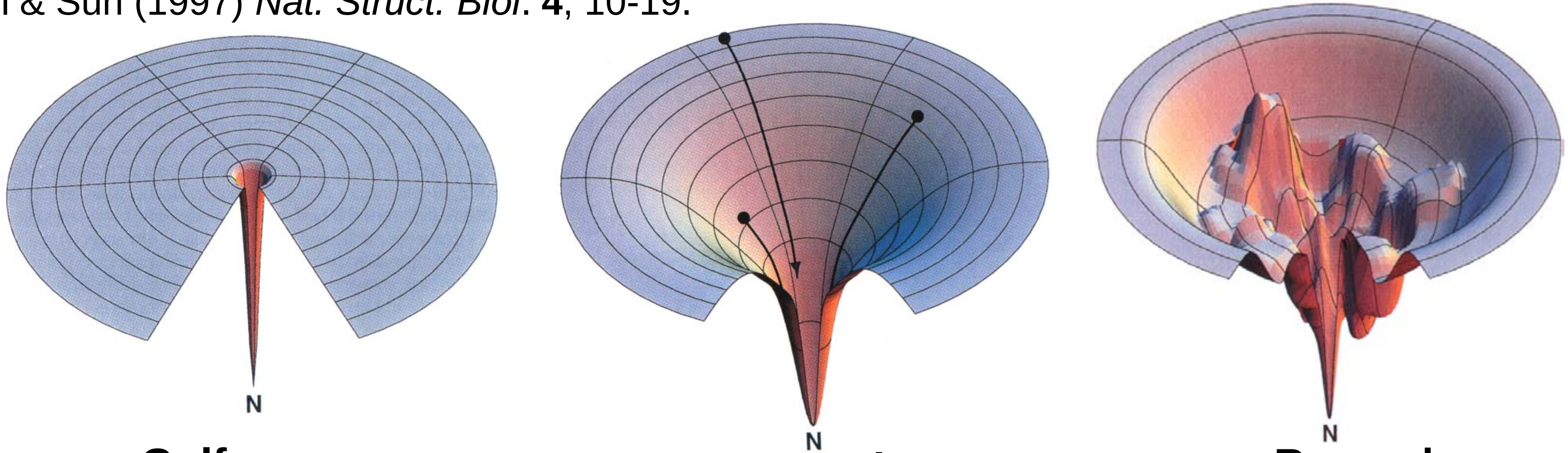
Protein Folding Energy Landscape

- **Levinthal's Paradox**: *It's impossible to search the whole conformation space. The folding must be following a path, therefore under kinetic control and there must be intermediates. But the refolding experiment clearly indicates that the thermodynamic equilibrium has been reached.* – Levinthal (1968) *J. Chem. Phys*., **65**, 44-45.

For example: a protein of 100 aa will take ~10^{74} years to fold assuming 3 possibilities for each dihedral and picosecond sampling rate.

$3^{198}/10^{12} \approx 10^{82}$ S ≈ 3×10^{74}Y >> 1.4×10^{10}Y (age of the universe)

Dill & Sun (1997) *Nat. Struct. Biol*. **4**, 10-19.



Golf course **Funnel** **Rugged**

Challenges in protein structure prediction

- Energy function: Not accurate
  - Quantum mechanics (accurate but too slow)
  - Molecular mechanics (Newtonian)
  - Heuristic (fast but inaccurate)
- Conformational space: Astronomical

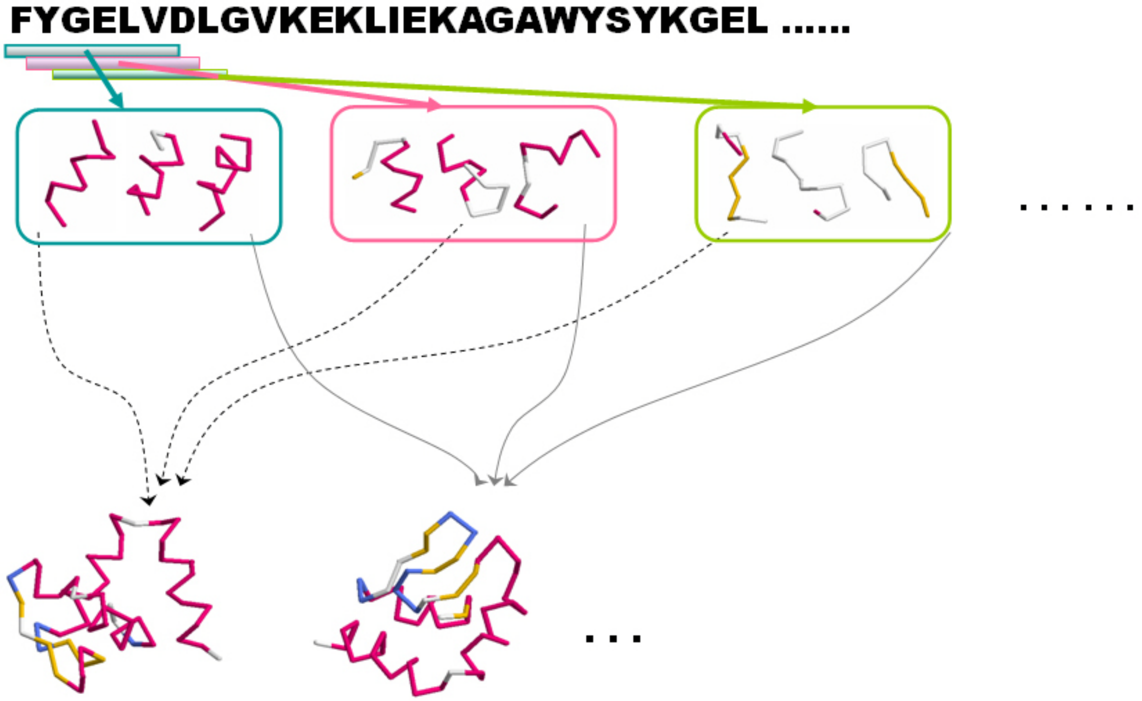
For a 100 aa protein, 99 peptide bonds, 198 ϕ & ψ dihedrals, assume 3 possibilities for each dihedral, one FLOP on K-computer to evaluate each conformation, it will take:

$3^{198} /10^{16} \approx 10^{78}$ S $\approx 3x10^{70}$Y $>> 1.4x10^{10}$Y (age of the universe)

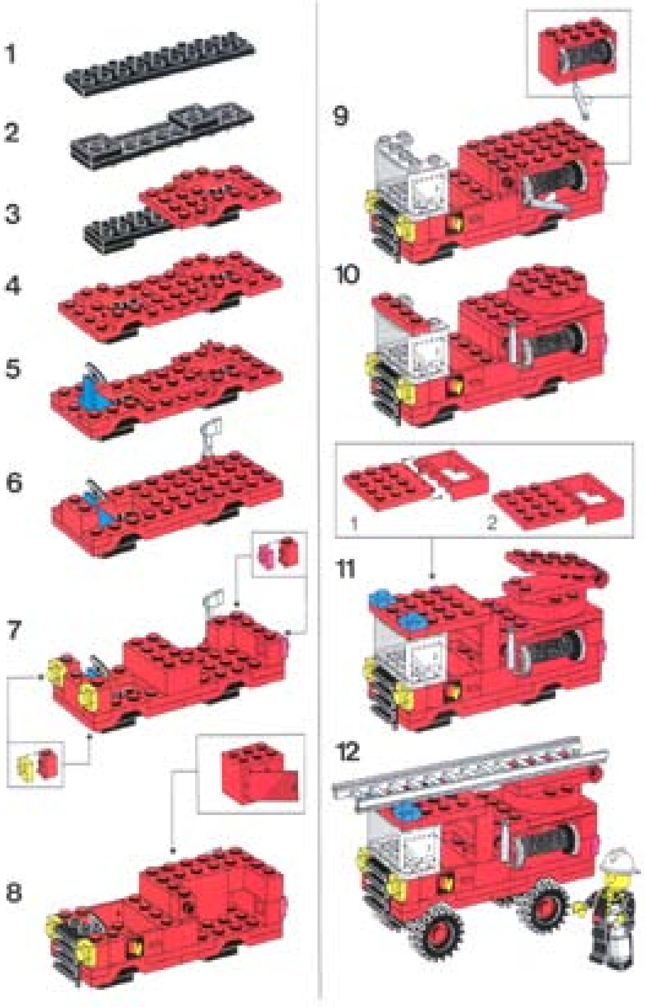
Fragment assembly approach to protein structure prediction

Bowie & Eisenberg (1994) *PNAS*, **91**, 4436-4440.

Simons, *etal*. & Baker (1997) *JMB*, **268**, 209-225.
Kuhlman, *etal*. & Baker (2003) *Science*, **302**, 1364-1368.



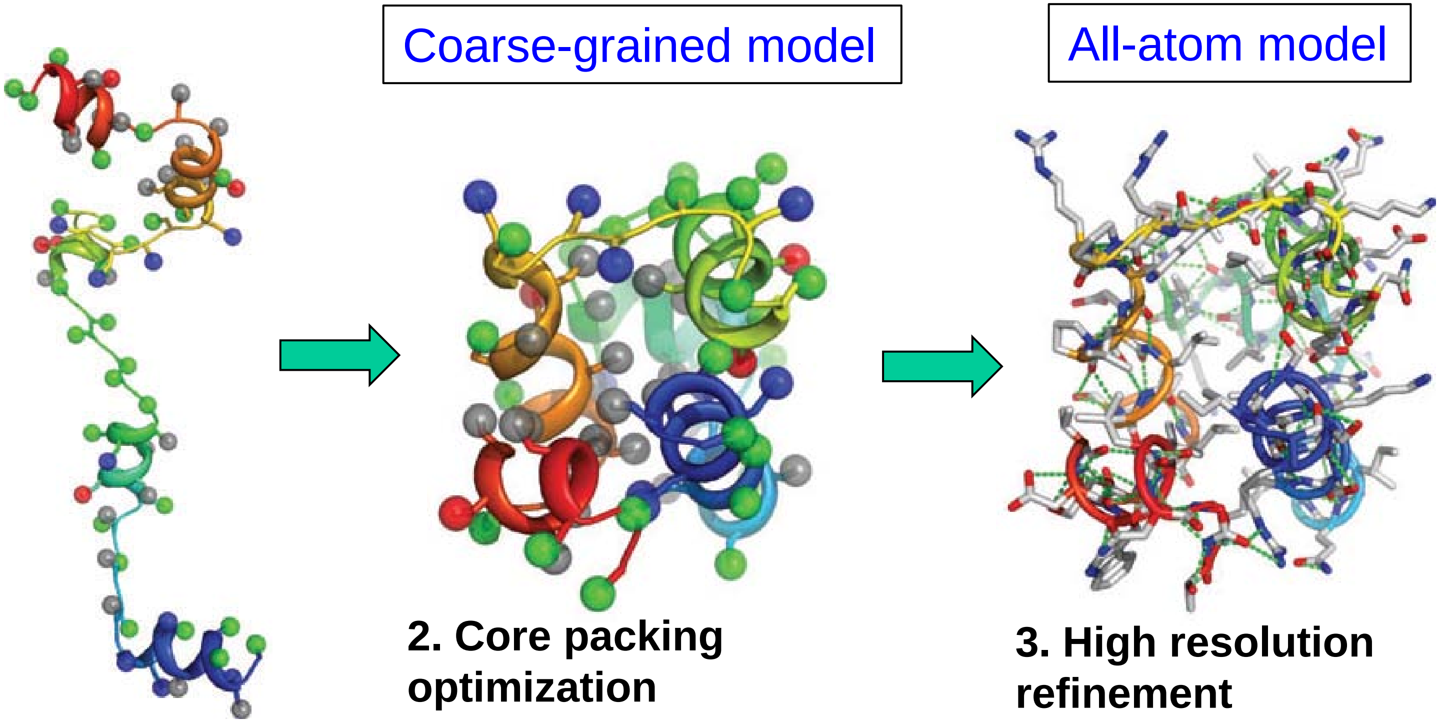
FYGELVDLGVKEKLIEKAGAWYSYKGEL ......



http://www.bi.a.u-tokyo.ac.jp/~shugo/3d_prediction.html

The Rosetta Structure Prediction Protocol

Das & Baker (2008) *Ann Rev Biochem*, **77**, 363-382.



Coarse-grained model

All-atom model

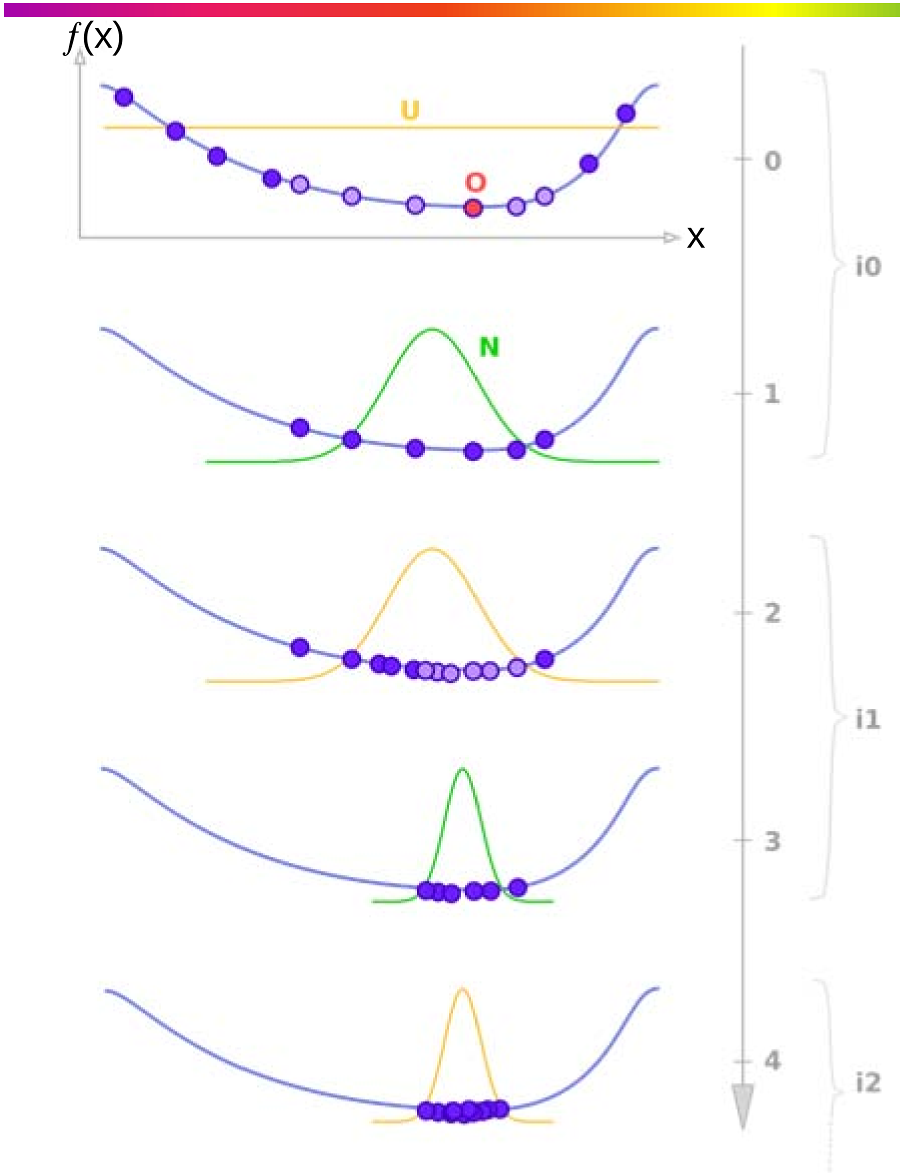
1. Fragment Assembly

2. Core packing optimization

3. High resolution refinement

Energy function:
Heuristic
Conformational search:
Monte Carlo

Estimation of Distribution Algorithm



● O: global minimum

▬ U: uniform distribution

▬ N: normal distribution

● P: a population of solutions

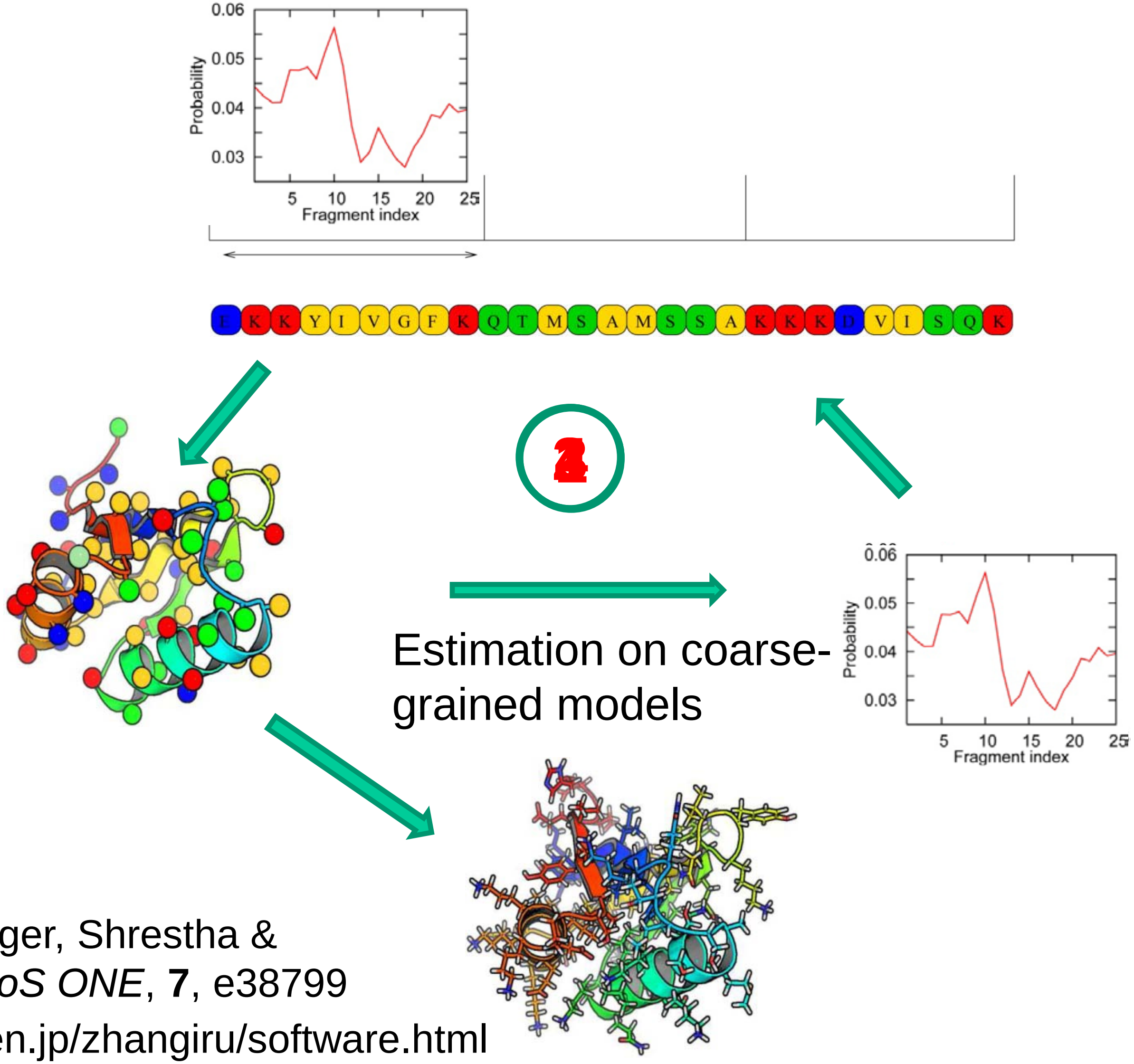
● PS: a sub-population of good solutions

▬ PDe: estimated distribution with PS

▬ PDu: new distribution used for next round of sampling

Johann Dréo, http://en.wikipedia.org/wiki/File:Eda_mono-variant_gauss_iterations.svg

EdaFold – Protein Folding with Estimation of Distribution Algorithm

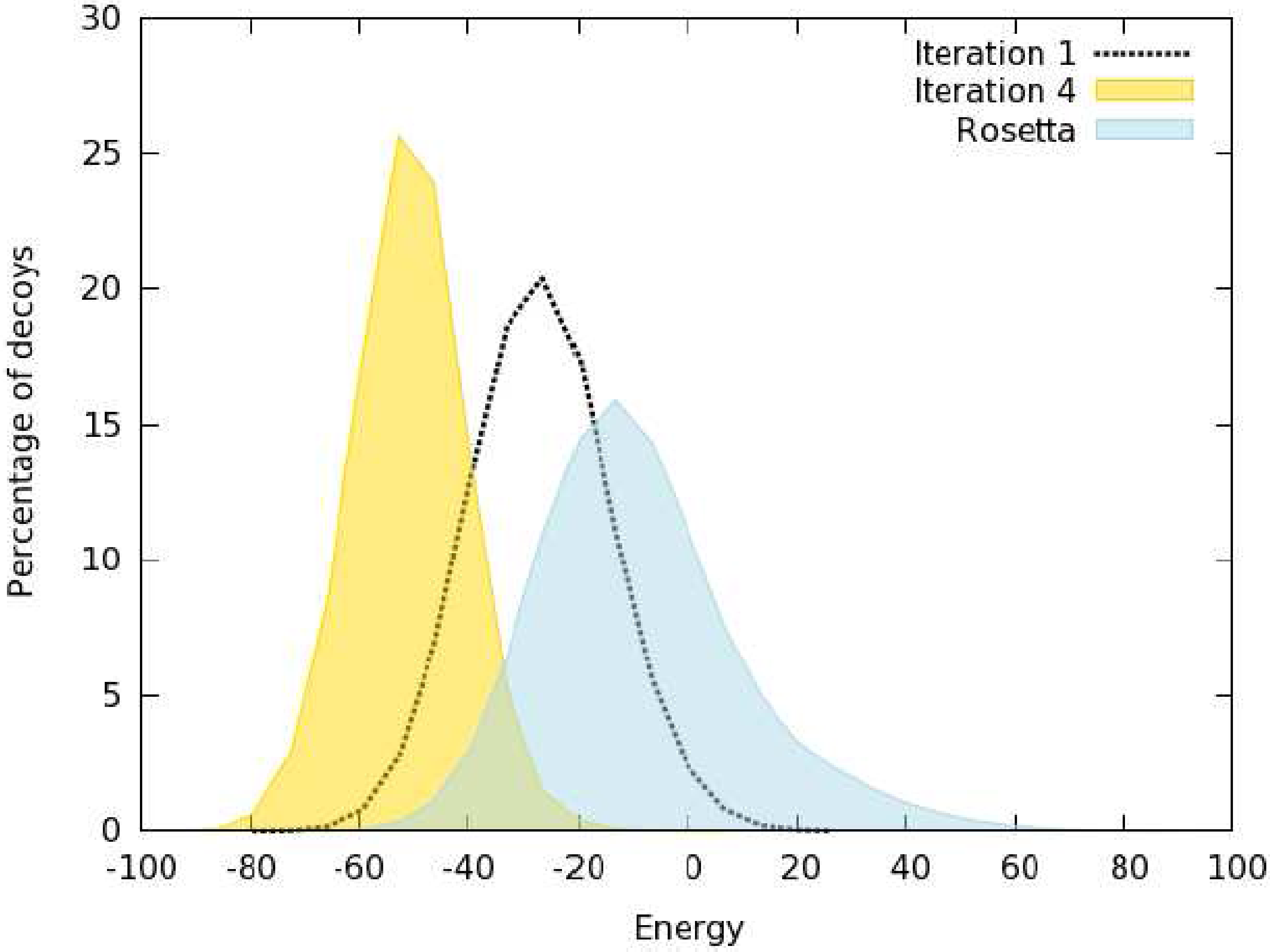


Estimation on coarse-grained models

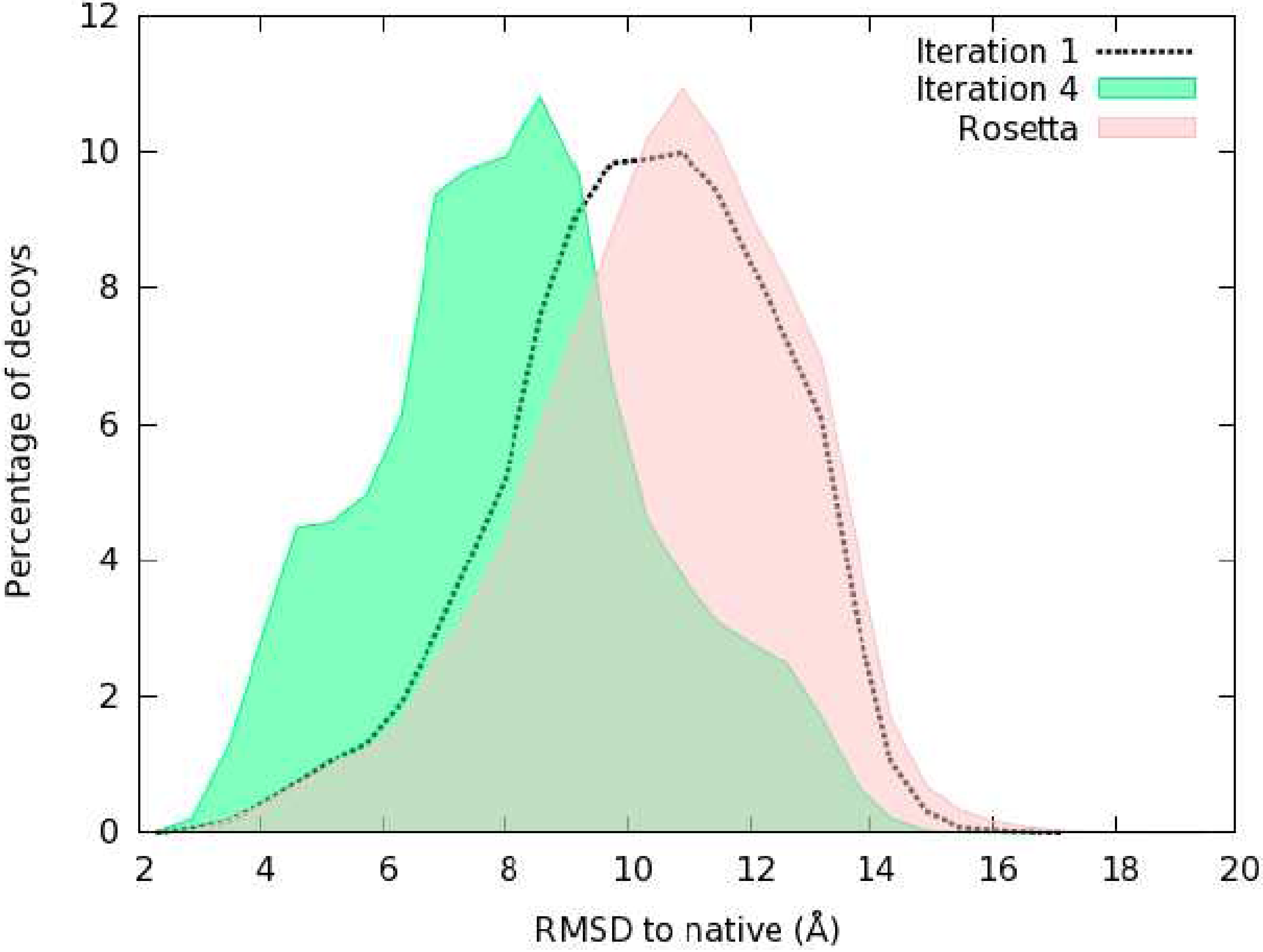
Simoncini, Berenger, Shrestha & Zhang (2012) *PLoS ONE*, **7**, e38799

http://www.riken.jp/zhangiru/software.html

Energy improvement on coarse-grained models

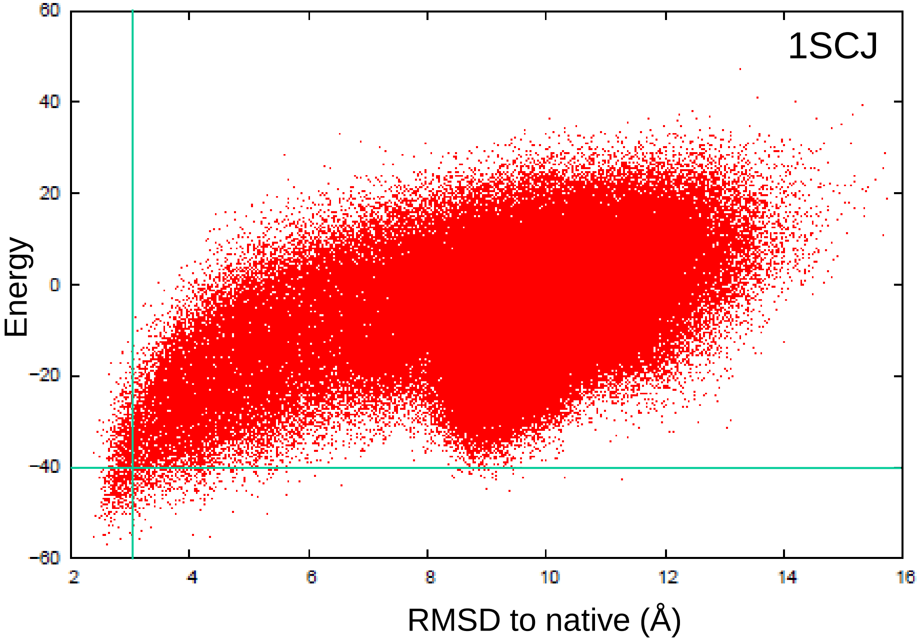


RMSD improvement on coarse-grained models

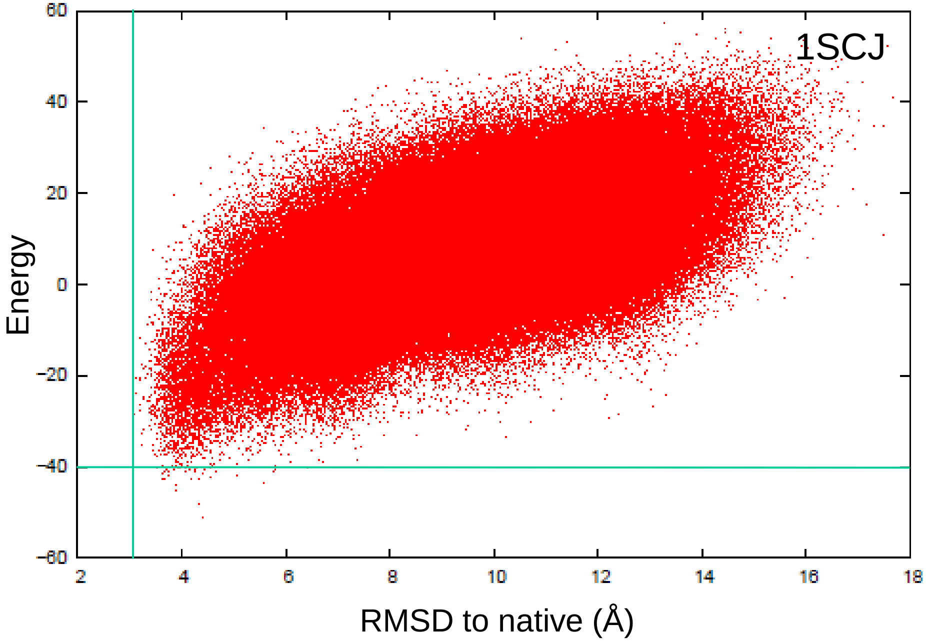


Energy landscape of coarse-grained models

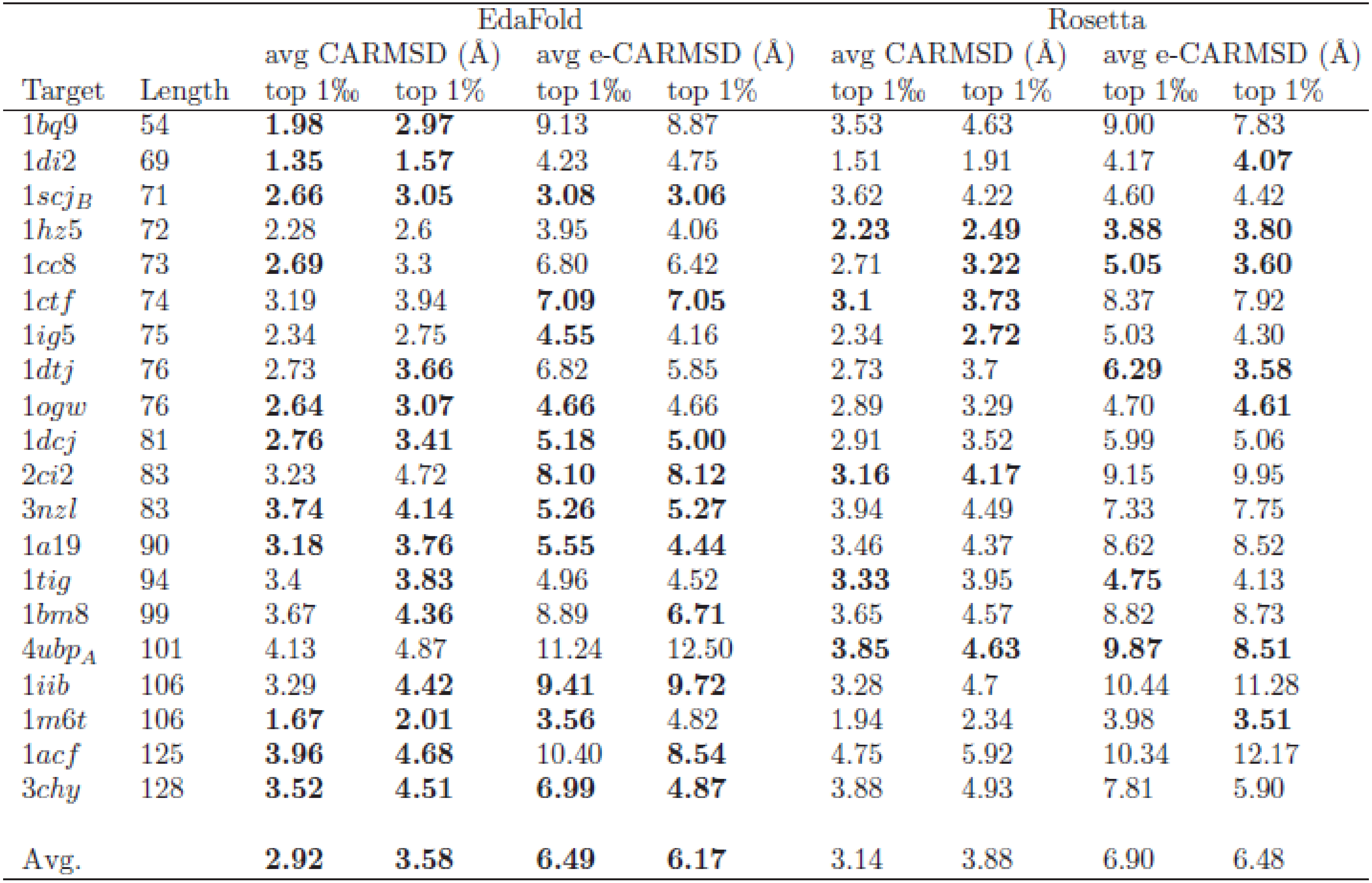
EdaFold



Rosetta

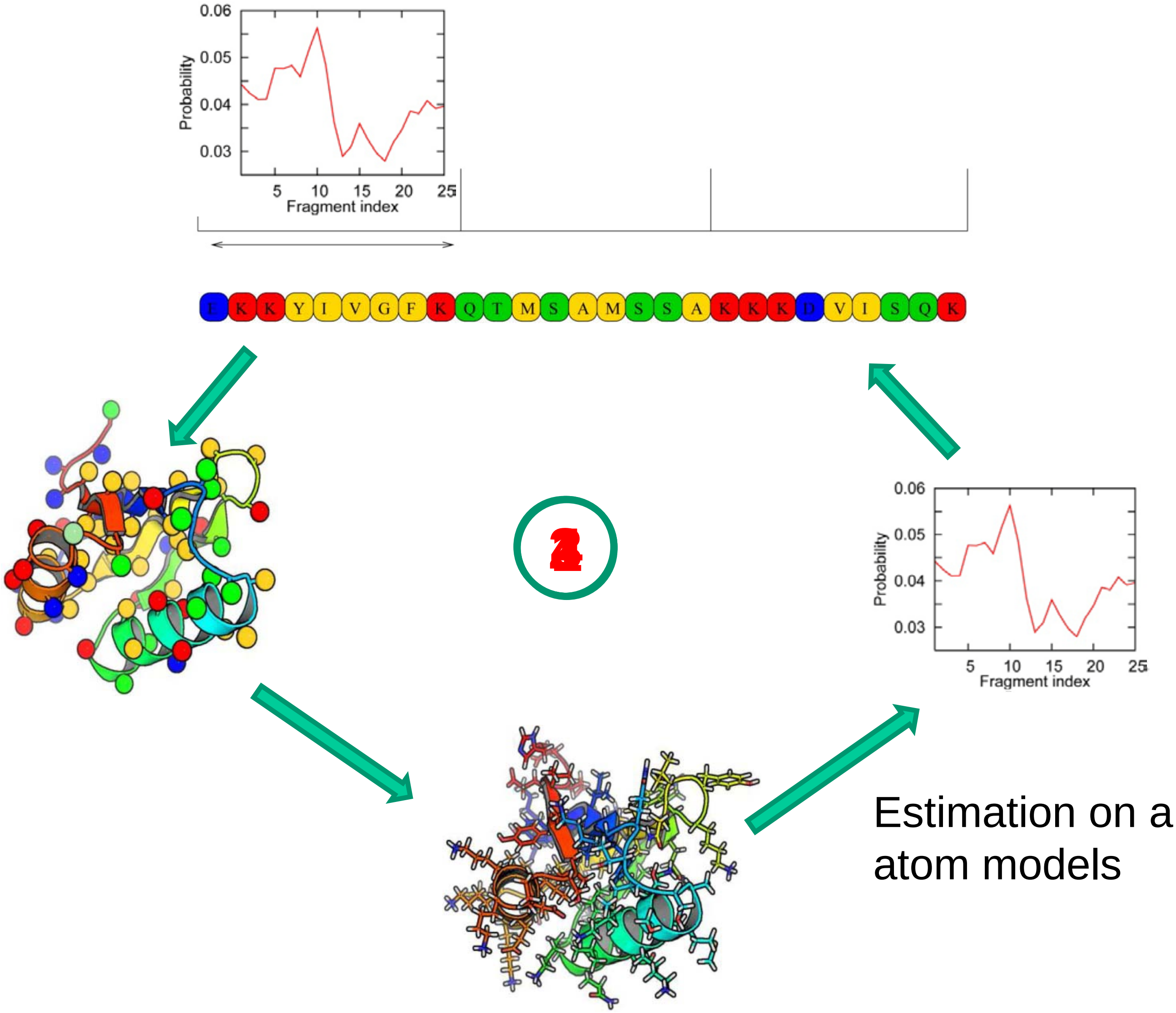


Compare coarse-grained models generated by EdaFold and Rosetta



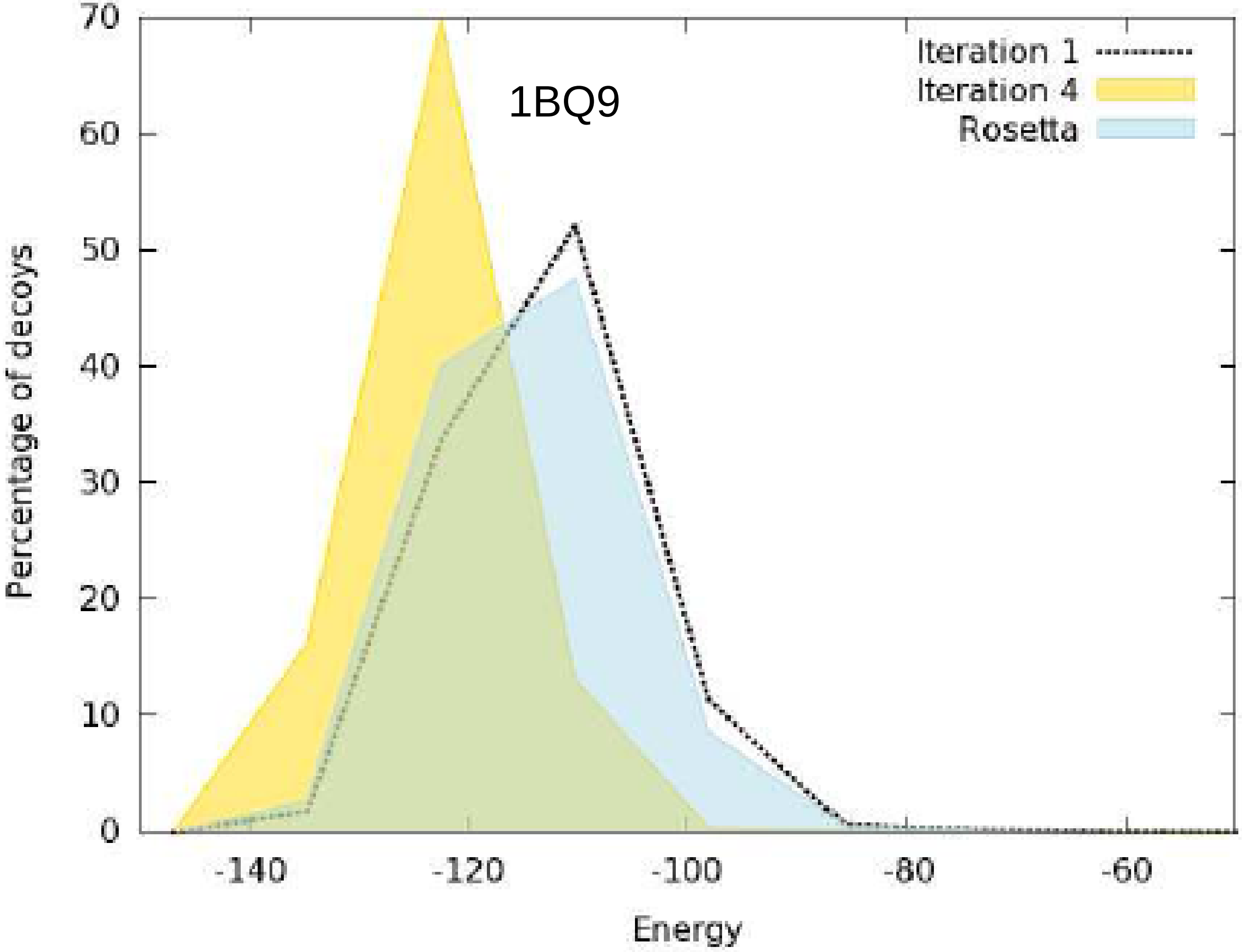
		EdaFold				Rosetta			
		avg CARMSD (Å)		avg e-CARMSD (Å)		avg CARMSD (Å)		avg e-CARMSD (Å)	
Target	Length	top 1‰	top 1%	top 1‰	top 1%	top 1‰	top 1%	top 1‰	top 1%
1*bq*9	54	**1.98**	**2.97**	9.13	8.87	3.53	4.63	9.00	7.83
1*di*2	69	**1.35**	**1.57**	4.23	4.75	1.51	1.91	4.17	**4.07**
1*scj*$_B$	71	**2.66**	**3.05**	**3.08**	**3.06**	3.62	4.22	4.60	4.42
1*hz*5	72	2.28	2.6	3.95	4.06	**2.23**	**2.49**	**3.88**	**3.80**
1*cc*8	73	**2.69**	3.3	6.80	6.42	2.71	**3.22**	**5.05**	**3.60**
1*ctf*	74	3.19	3.94	**7.09**	**7.05**	**3.1**	**3.73**	8.37	7.92
1*ig*5	75	2.34	2.75	**4.55**	4.16	2.34	**2.72**	5.03	4.30
1*dtj*	76	2.73	**3.66**	6.82	5.85	2.73	3.7	**6.29**	**3.58**
1*ogw*	76	**2.64**	**3.07**	**4.66**	4.66	2.89	3.29	4.70	**4.61**
1*dcj*	81	**2.76**	**3.41**	**5.18**	**5.00**	2.91	3.52	5.99	5.06
2*ci*2	83	3.23	4.72	**8.10**	**8.12**	**3.16**	**4.17**	9.15	9.95
3*nzl*	83	**3.74**	**4.14**	**5.26**	**5.27**	3.94	4.49	7.33	7.75
1*a*19	90	**3.18**	**3.76**	**5.55**	**4.44**	3.46	4.37	8.62	8.52
1*tig*	94	3.4	**3.83**	4.96	4.52	**3.33**	3.95	**4.75**	4.13
1*bm*8	99	3.67	**4.36**	8.89	**6.71**	3.65	4.57	8.82	8.73
4*ubp*$_A$	101	4.13	4.87	11.24	12.50	**3.85**	**4.63**	**9.87**	**8.51**
1*iib*	106	3.29	**4.42**	**9.41**	**9.72**	3.28	4.7	10.44	11.28
1*m*6*t*	106	**1.67**	**2.01**	**3.56**	4.82	1.94	2.34	3.98	**3.51**
1*acf*	125	**3.96**	**4.68**	10.40	**8.54**	4.75	5.92	10.34	12.17
3*chy*	128	**3.52**	**4.51**	**6.99**	**4.87**	3.88	4.93	7.81	5.90
Avg.		**2.92**	**3.58**	**6.49**	**6.17**	3.14	3.88	6.90	6.48

EdaFold II – Protein Folding with Estimation of Distribution Algorithm

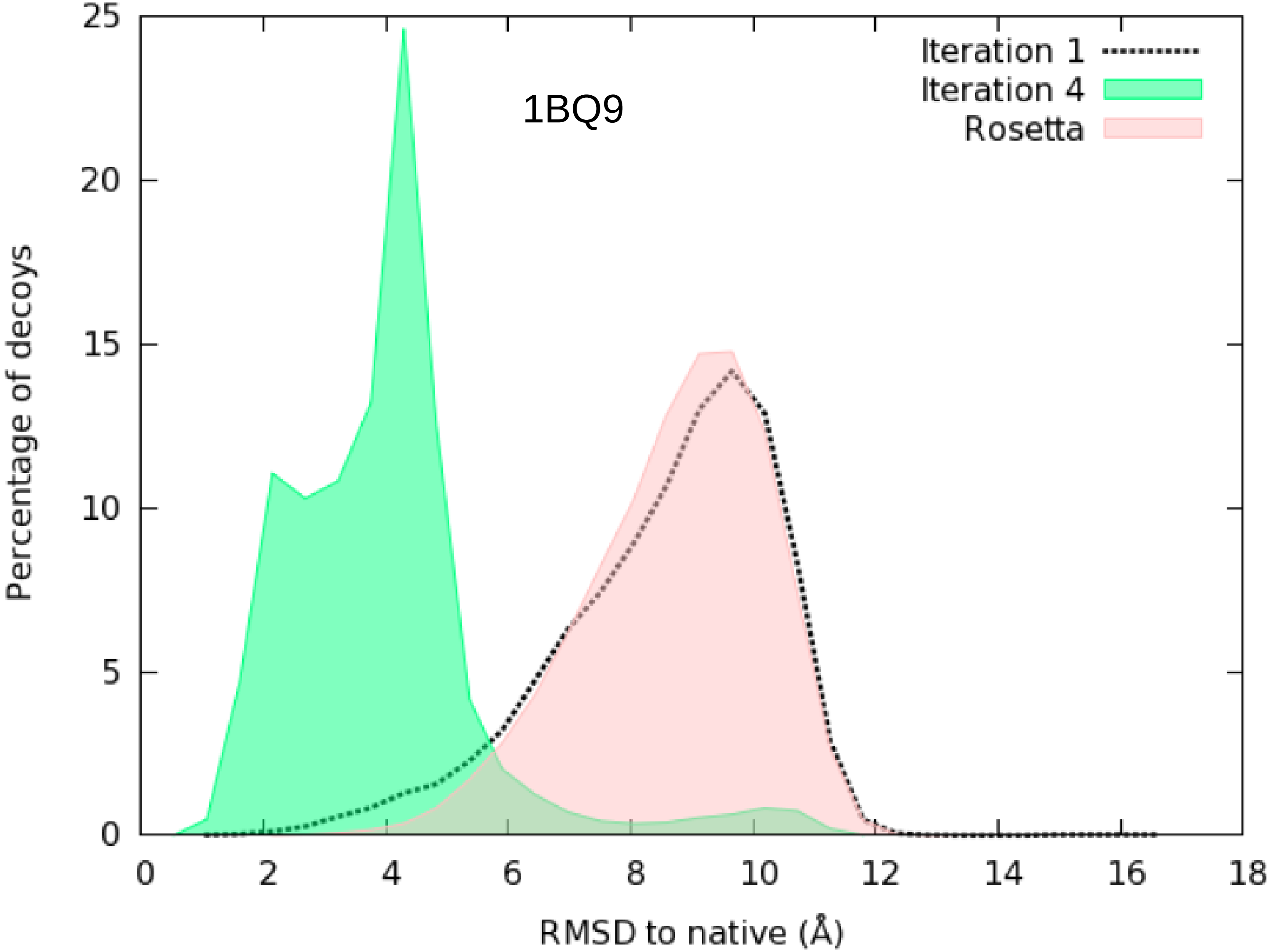


Estimation on all-atom models

Energy distribution of all-atom models

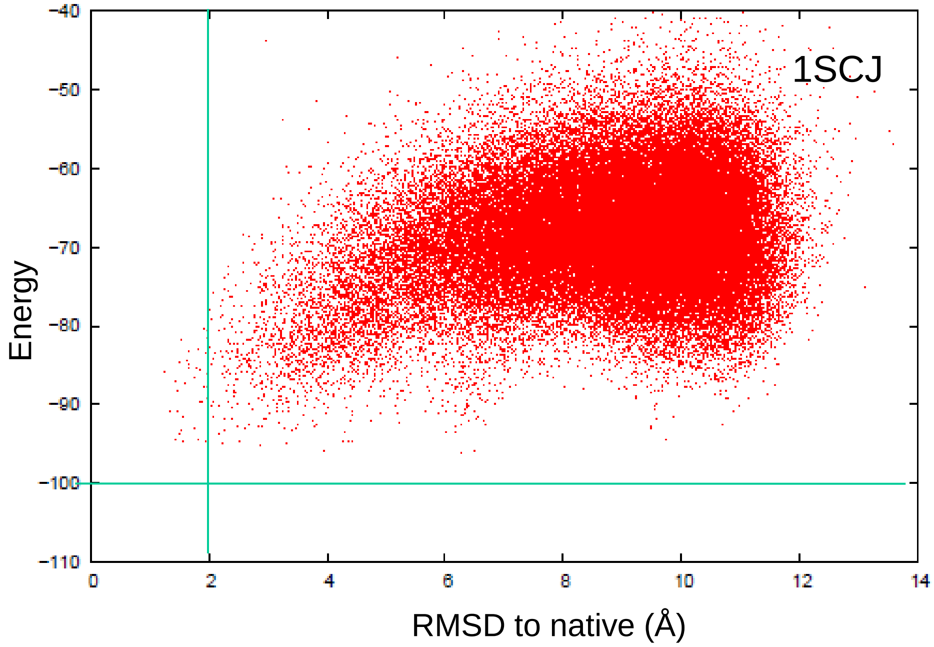


Quality distribution of all-atom models

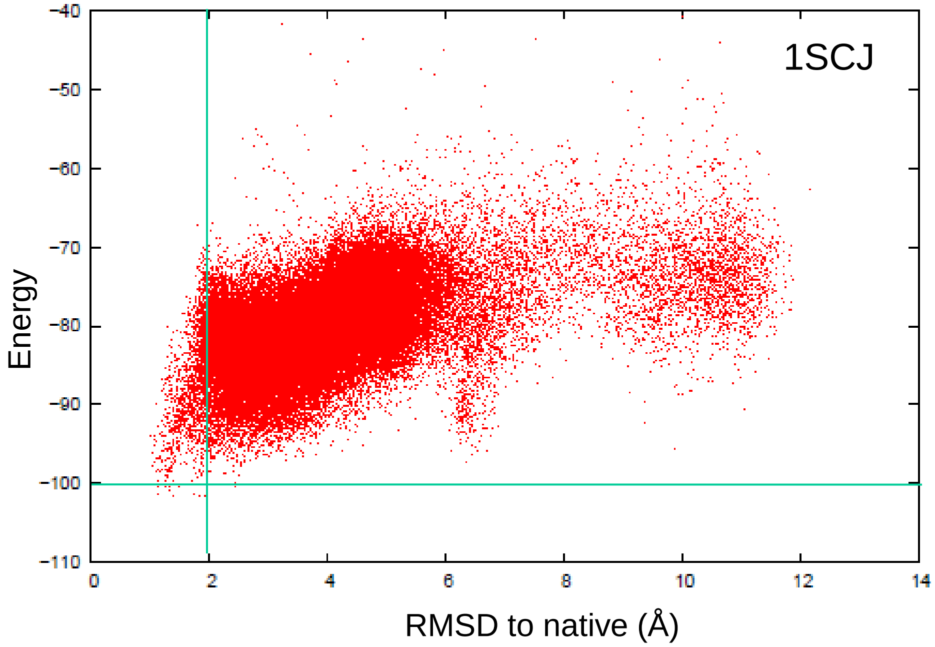


Energy landscape of all atom models

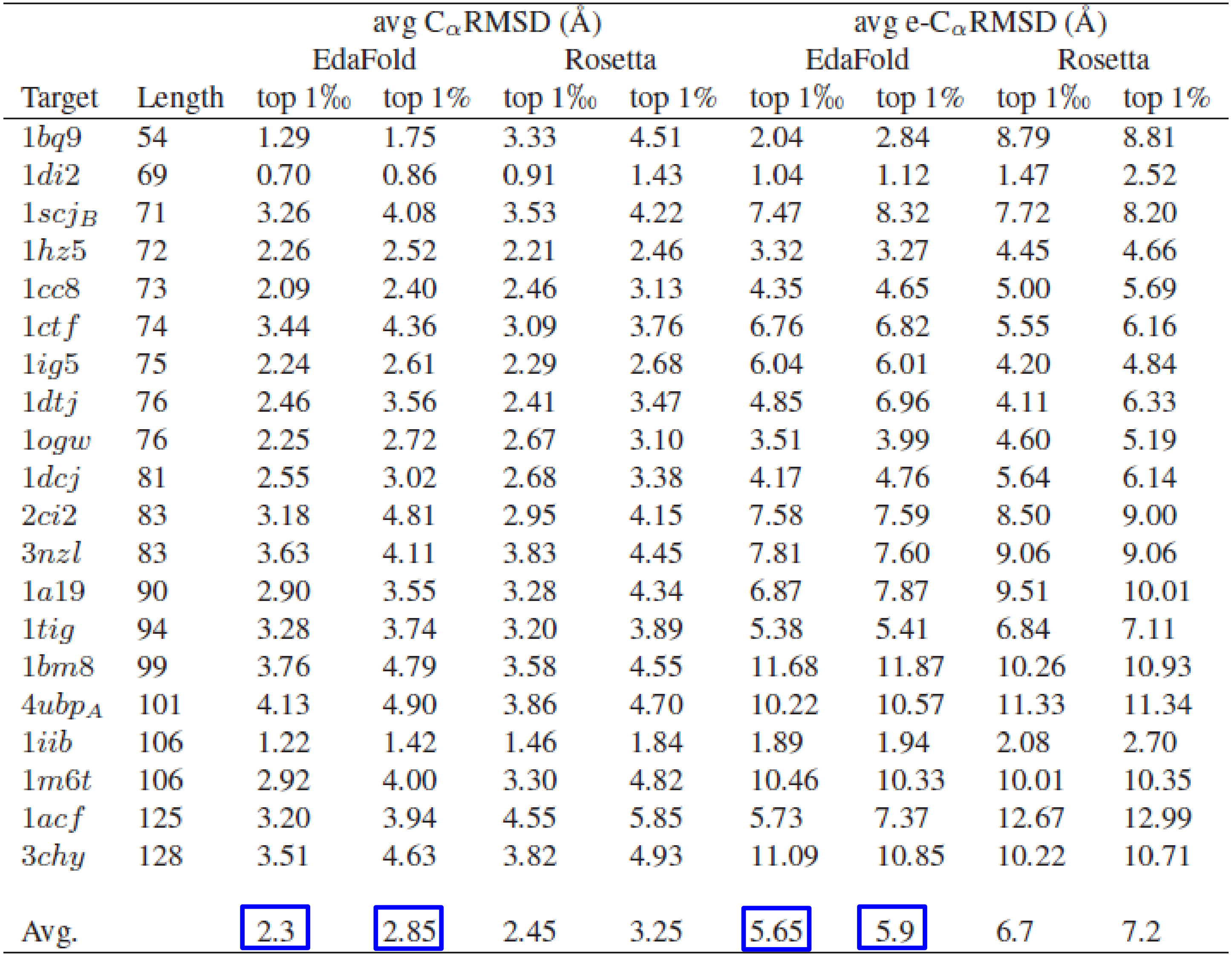
EdaFold iteration 1



EdaFold iteration 4

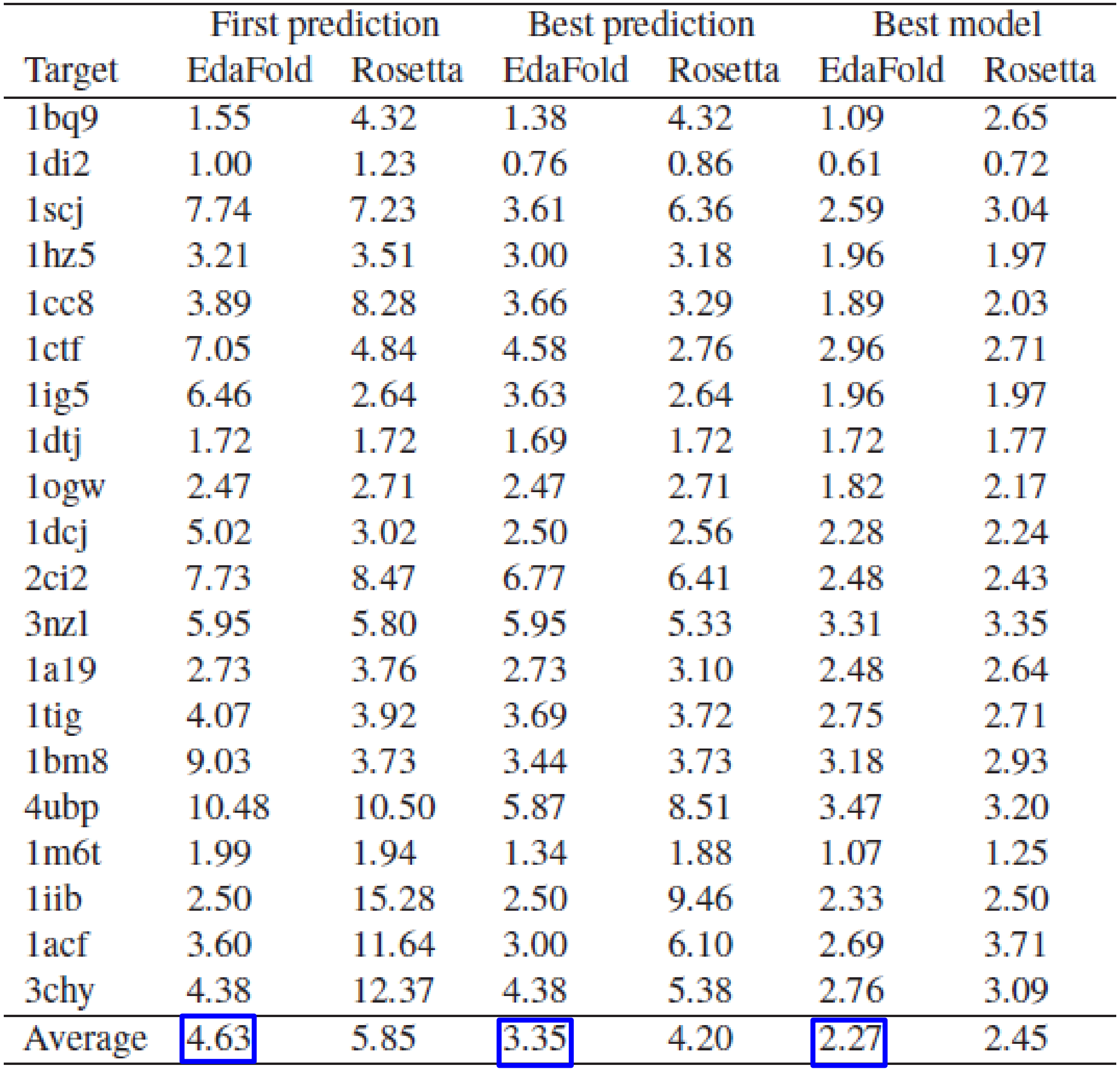


Performance of EdaFold and Rosetta



		avg C$_\alpha$RMSD (Å)				avg e-C$_\alpha$RMSD (Å)			
		EdaFold		Rosetta		EdaFold		Rosetta	
Target	Length	top 1‰	top 1%	top 1‰	top 1%	top 1‰	top 1%	top 1‰	top 1%
1bq9	54	1.29	1.75	3.33	4.51	2.04	2.84	8.79	8.81
1di2	69	0.70	0.86	0.91	1.43	1.04	1.12	1.47	2.52
1scj$_B$	71	3.26	4.08	3.53	4.22	7.47	8.32	7.72	8.20
1hz5	72	2.26	2.52	2.21	2.46	3.32	3.27	4.45	4.66
1cc8	73	2.09	2.40	2.46	3.13	4.35	4.65	5.00	5.69
1ctf	74	3.44	4.36	3.09	3.76	6.76	6.82	5.55	6.16
1ig5	75	2.24	2.61	2.29	2.68	6.04	6.01	4.20	4.84
1dtj	76	2.46	3.56	2.41	3.47	4.85	6.96	4.11	6.33
1ogw	76	2.25	2.72	2.67	3.10	3.51	3.99	4.60	5.19
1dcj	81	2.55	3.02	2.68	3.38	4.17	4.76	5.64	6.14
2ci2	83	3.18	4.81	2.95	4.15	7.58	7.59	8.50	9.00
3nzl	83	3.63	4.11	3.83	4.45	7.81	7.60	9.06	9.06
1a19	90	2.90	3.55	3.28	4.34	6.87	7.87	9.51	10.01
1tig	94	3.28	3.74	3.20	3.89	5.38	5.41	6.84	7.11
1bm8	99	3.76	4.79	3.58	4.55	11.68	11.87	10.26	10.93
4ubp$_A$	101	4.13	4.90	3.86	4.70	10.22	10.57	11.33	11.34
1iib	106	1.22	1.42	1.46	1.84	1.89	1.94	2.08	2.70
1m6t	106	2.92	4.00	3.30	4.82	10.46	10.33	10.01	10.35
1acf	125	3.20	3.94	4.55	5.85	5.73	7.37	12.67	12.99
3chy	128	3.51	4.63	3.82	4.93	11.09	10.85	10.22	10.71
Avg.		2.3	2.85	2.45	3.25	5.65	5.9	6.7	7.2

Performance of EdaFold and Rosetta



	First prediction		Best prediction		Best model	
Target	EdaFold	Rosetta	EdaFold	Rosetta	EdaFold	Rosetta
1bq9	1.55	4.32	1.38	4.32	1.09	2.65
1di2	1.00	1.23	0.76	0.86	0.61	0.72
1scj	7.74	7.23	3.61	6.36	2.59	3.04
1hz5	3.21	3.51	3.00	3.18	1.96	1.97
1cc8	3.89	8.28	3.66	3.29	1.89	2.03
1ctf	7.05	4.84	4.58	2.76	2.96	2.71
1ig5	6.46	2.64	3.63	2.64	1.96	1.97
1dtj	1.72	1.72	1.69	1.72	1.72	1.77
1ogw	2.47	2.71	2.47	2.71	1.82	2.17
1dcj	5.02	3.02	2.50	2.56	2.28	2.24
2ci2	7.73	8.47	6.77	6.41	2.48	2.43
3nzl	5.95	5.80	5.95	5.33	3.31	3.35
1a19	2.73	3.76	2.73	3.10	2.48	2.64
1tig	4.07	3.92	3.69	3.72	2.75	2.71
1bm8	9.03	3.73	3.44	3.73	3.18	2.93
4ubp	10.48	10.50	5.87	8.51	3.47	3.20
1m6t	1.99	1.94	1.34	1.88	1.07	1.25
1iib	2.50	15.28	2.50	9.46	2.33	2.50
1acf	3.60	11.64	3.00	6.10	2.69	3.71
3chy	4.38	12.37	4.38	5.38	2.76	3.09
Average	4.63	5.85	3.35	4.20	2.27	2.45

Acknowledgment

- David Simoncini
- Francois Berenger
- Rojan Shrestha

- Dr. David Baker & members of Rosetta commons for making Rosetta source code available
- Dr. Ryutaro Himeno and his staff at HPC for help
- RICC for computing time
- RIKEN Initiative Research Unit Program for funding

Thank you!