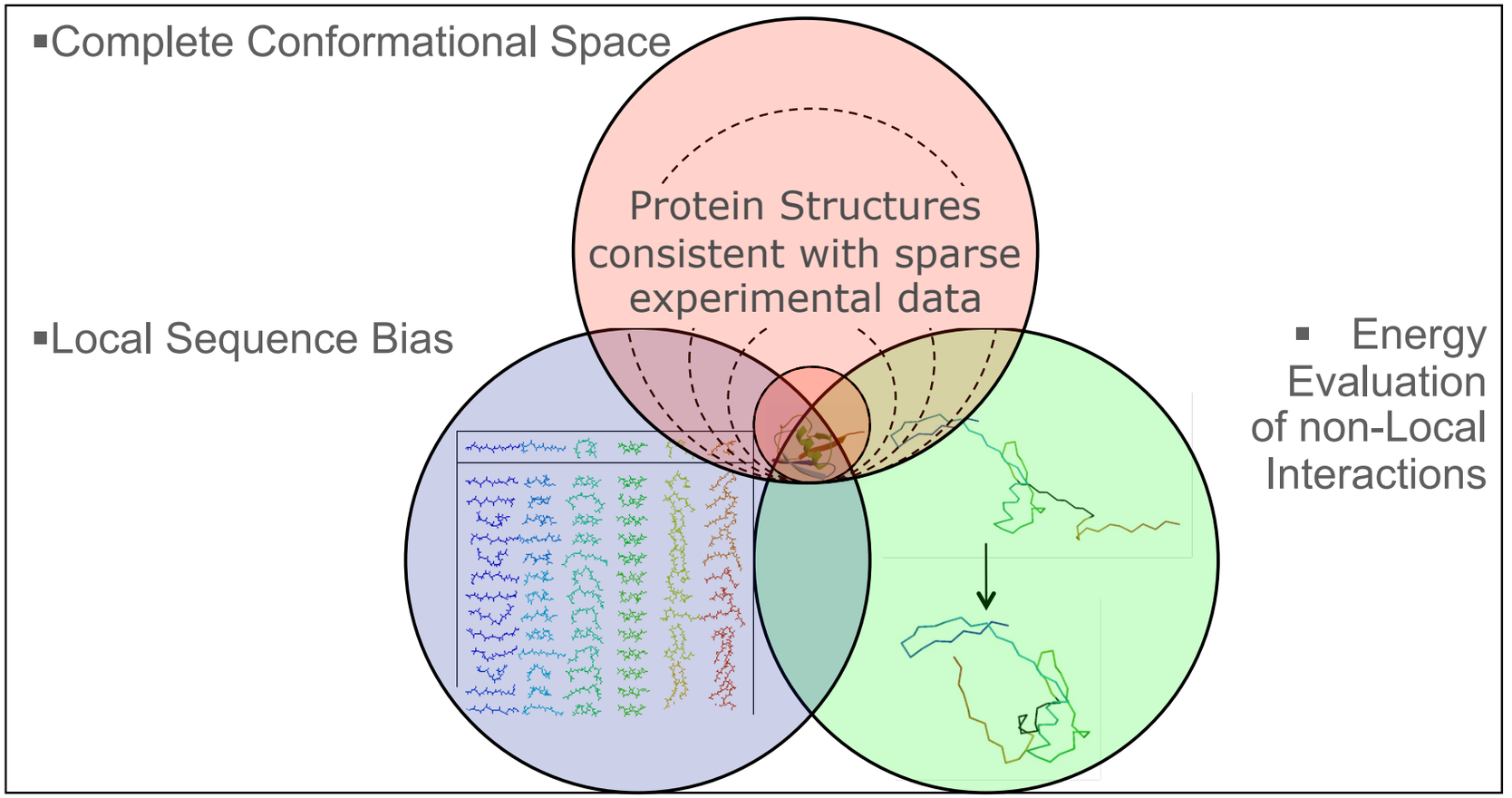
RosettaEPR: An Integrated Tool for Protein Structure Determination from Sparse EPR Data

Stephanie Hirst

RosettaCon

August 4, 2010

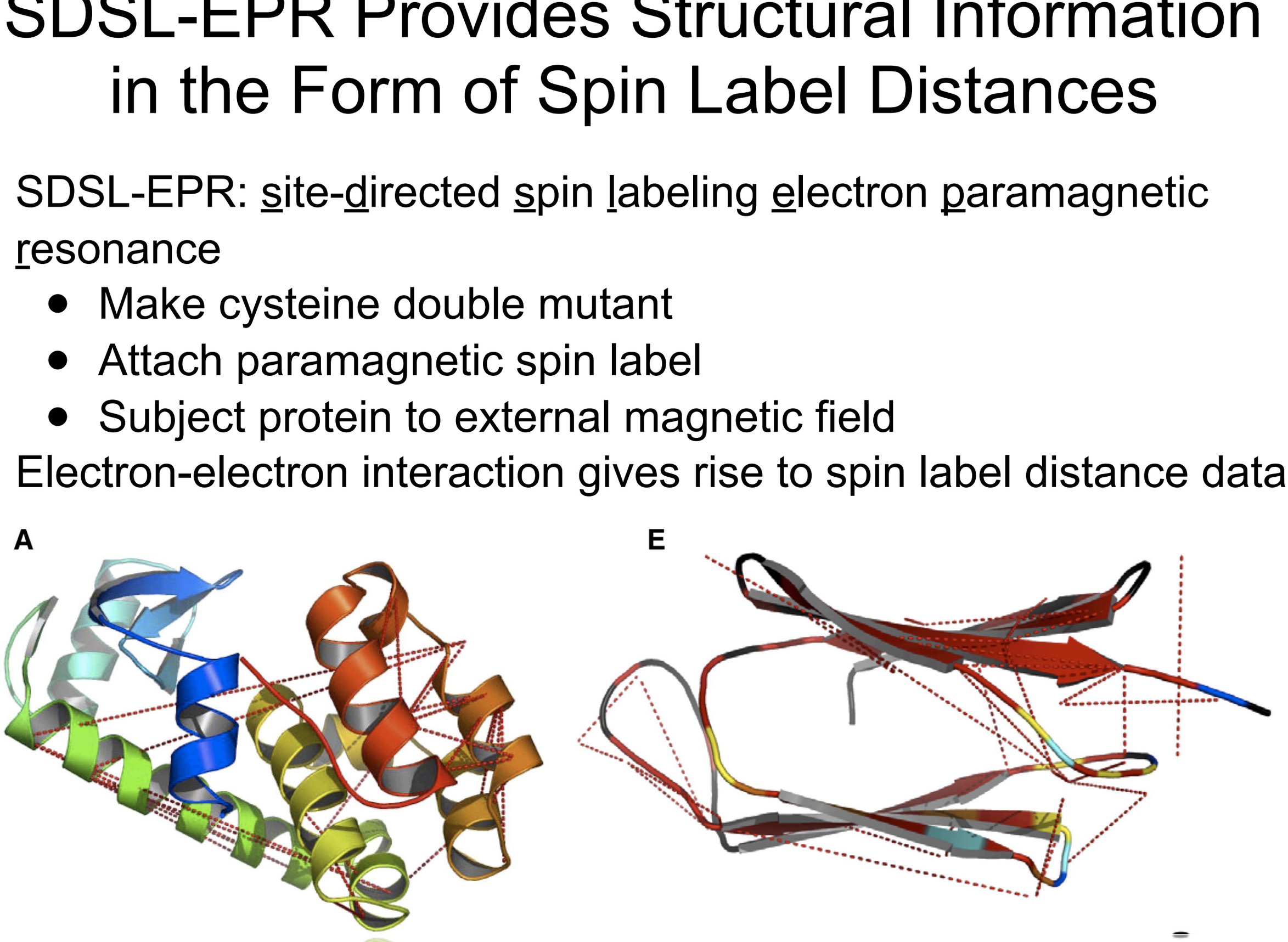
Why Combine Experimental Restraints with Rosetta?



- computationally infeasible to exhaustively sample all of conformational space
- experimental data bias search to more native-like conformations
- allows for validation of predicted structures

SDSL-EPR Provides Structural Information in the Form of Spin Label Distances

- SDSL-EPR: site-directed spin labeling electron paramagnetic resonance
  - Make cysteine double mutant
  - Attach paramagnetic spin label
  - Subject protein to external magnetic field
- Electron-electron interaction gives rise to spin label distance data



Alexander, N., *et al*, *Structure* (2008) *16*, 181-195.

Strengths and Weaknesses of SDSL-EPR

Strengths

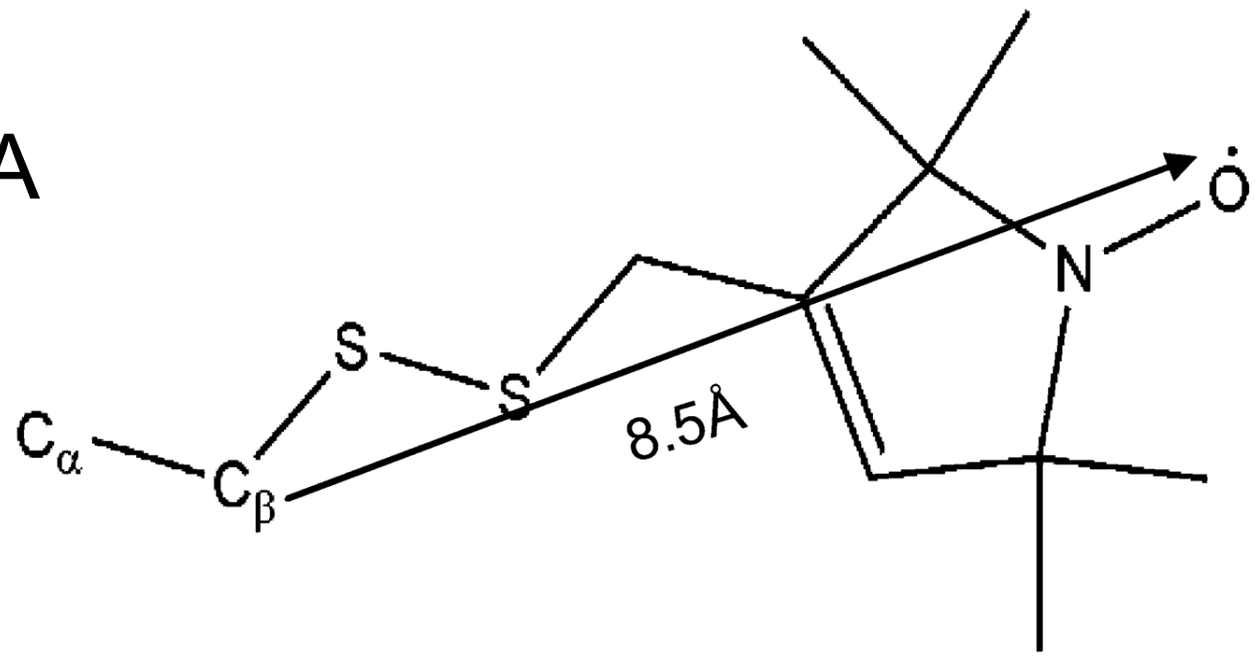
- No crystallization
- No size constraints
- Can study protein in native environment
- Doesn’t require much sample (pmoles)

Weaknesses

- Perturbation of structure upon spin labeling
- Sparseness of EPR data
- Modeling the spin label
- Doesn’t yield atomic-detail models directly

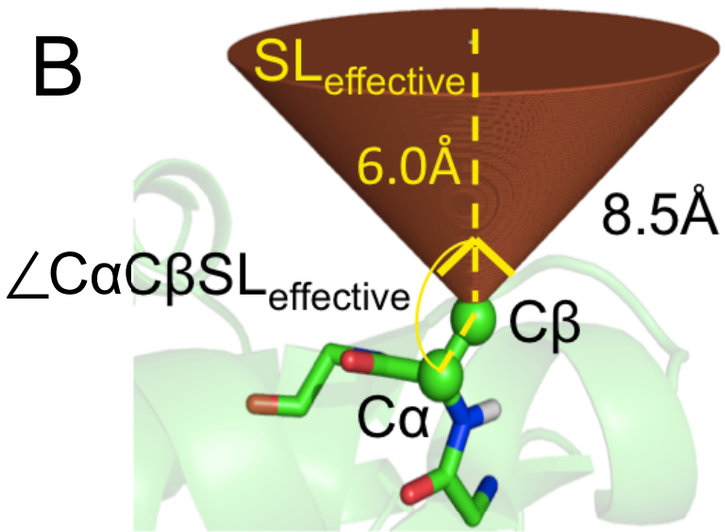
The Cone Model Translates EPR Experimental Data to Structural Restraints

A

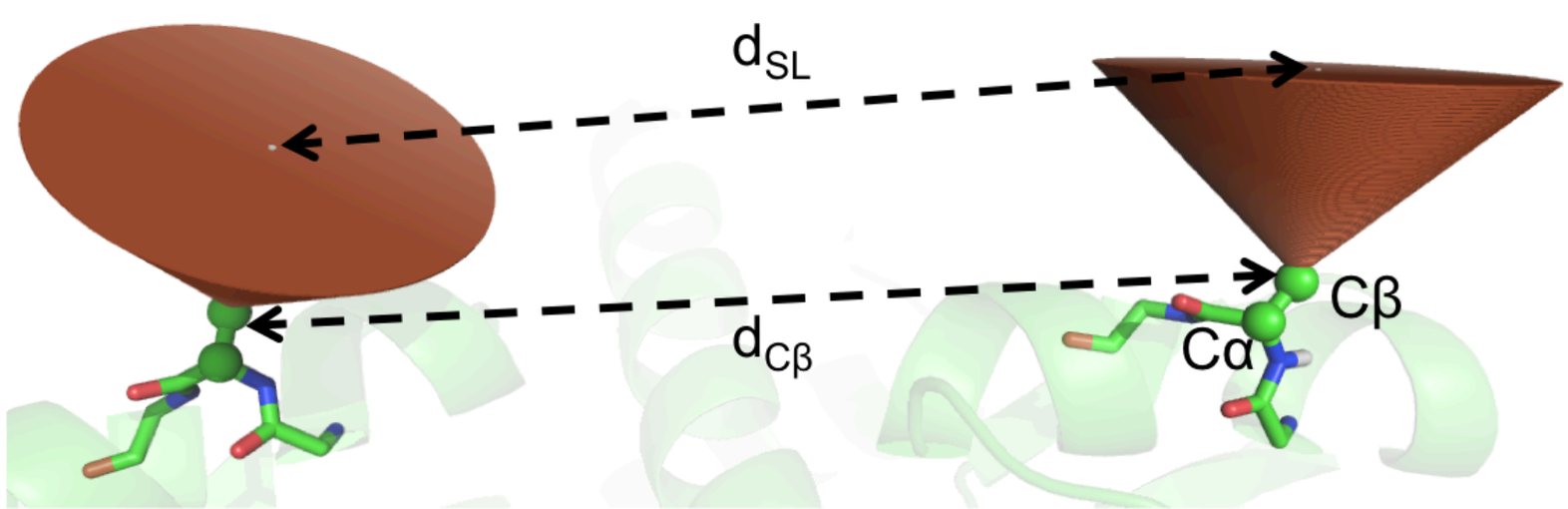


nethanethiosulfonate (MTS)
spin label

B

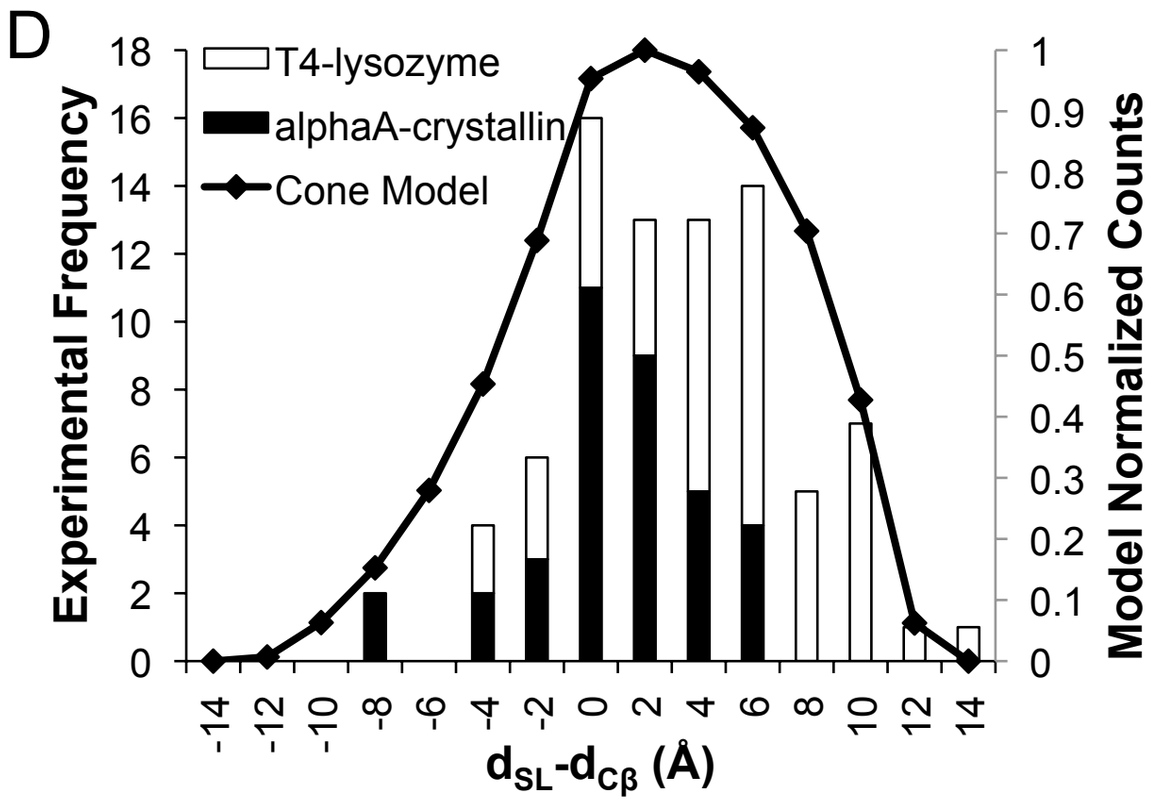


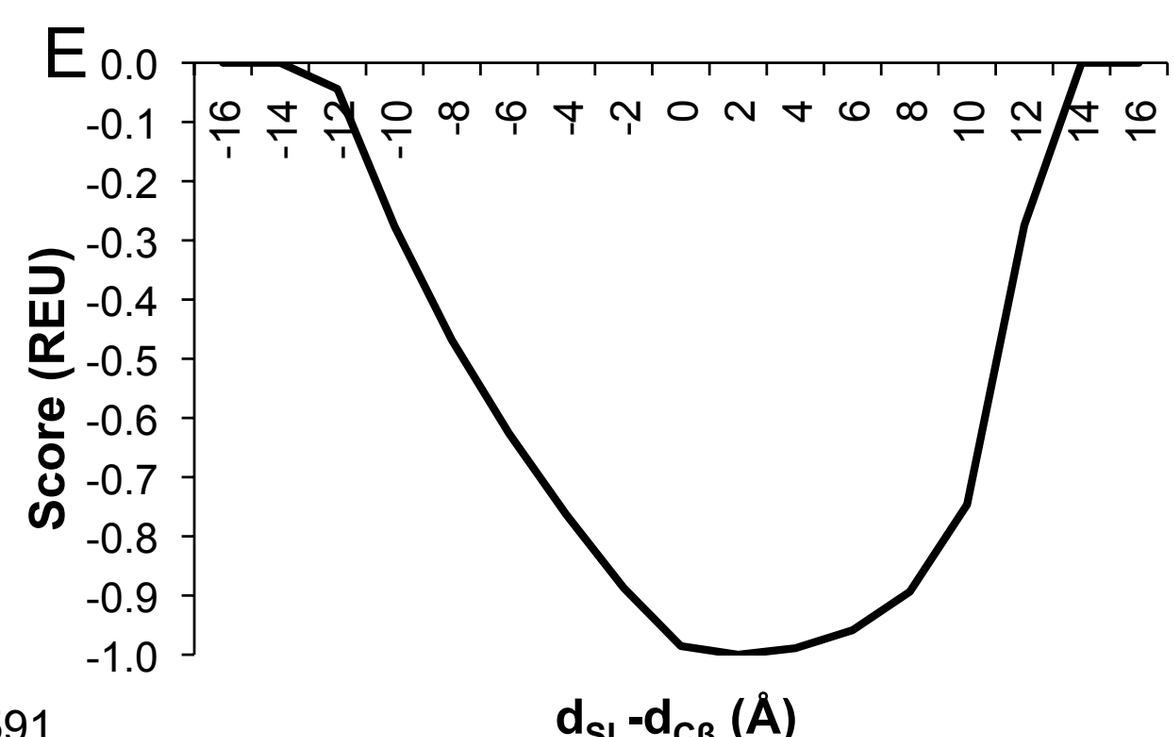
C



Knowledge-Based Cone Model Statistics Reflect Experimental Observations

- Simulated spin label (cone model) was placed at every pair of exposed residues in a protein
- Computed $d_{SL} - d_{C\beta}$
- Repeated for 3,584 soluble proteins in Dunbrack database
- Converted to potential via the Boltzmann relation





Wang, G. and Dunbrack, R. L. *Bioinformatics* (2003) *19*, 1589-1591.

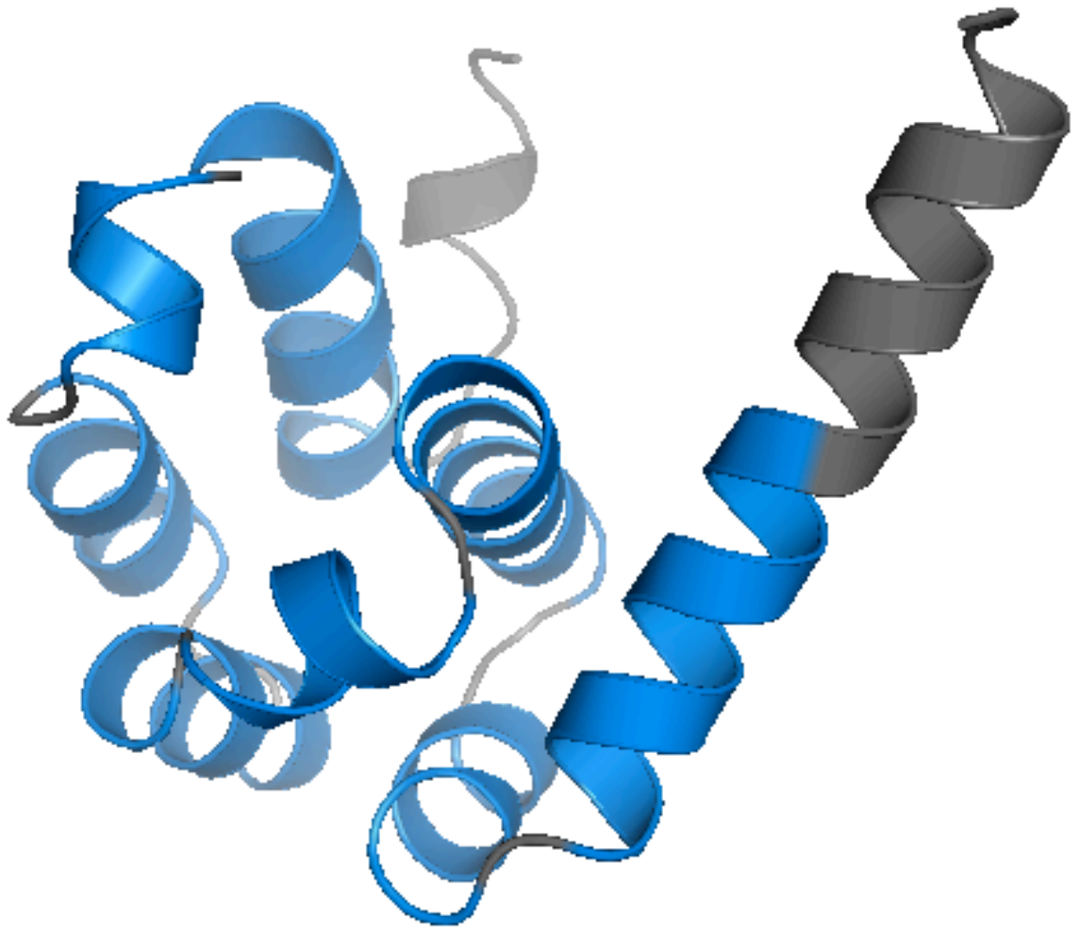
New EPR Knowledge-Based Potential Requires a New Constraint Function

- SplineFunc added to constraint function types
- Reads in any histogram and create a cubic spline over it
- Given a value of *x* (e.g., $d_{C\beta}$), return the corresponding energy
- When specifying EPR_DISTANCE, reads in EPR distance histogram from the database and returns the appropriate energy for $d_{SL} - d_{C\beta}$
- Can call by -constraints::epr_distance

cst type	atm1	res1	atm2	res2	function	RosettaEPR	d_{SL}	wt	bin size
AtomPair	CB	31	CB	43	SPLINE	EPR_DISTANCE	6.0	4.0	0.5

RosettaEPR Benchmarking Protocol

- Benchmark: T4-lysozyme (PDB ID: 2LZM) residues 58-164
- *De novo* folded 10,000 models with 25 distance restraints
  - scored with bounded restraint potential and knowledge-based potential over a range of weights
- Computed $RMSD_{C\alpha}$ over core and exclude all loops



De novo Folding with EPR Restraints Improves Sampling Overall

Weight	% Models with $RMSD_{C\alpha} < 3.5Å$	% Models with $RMSD_{C\alpha} < 7.5Å$	% Models with $RMSD_{C\alpha} < 3.5Å$	% Models with $RMSD_{C\alpha} < 7.5Å$
0	0.03	7.17		
	RosettaEPR		**Bounded**	
1	0.73	21.98	0.89	37.56
2	1.41	31.07	1.18	40.95
3	2.01	37.20	1.58	41.84
4	2.05	42.08	1.62	41.09
5	1.83	45.65	1.43	40.44
6	1.60	47.29	1.40	39.50
7	1.35	49.60	1.40	38.42
8	1.31	51.21	1.62	38.01
9	0.87	50.89	1.59	37.42
10	1.02	52.70	1.57	37.22
20	0.51	54.89	1.44	34.02
30	0.46	53.28	1.22	32.77
40	0.25	49.74	1.27	32.16
50	0.17	47.43	1.12	32.27
60	0.07	43.86	1.01	31.07
70	0.03	43.95	1.29	31.67
80	0.02	43.07	1.34	31.05
90	0.01	40.92	1.39	31.22
100	0.01	41.11	1.12	30.62

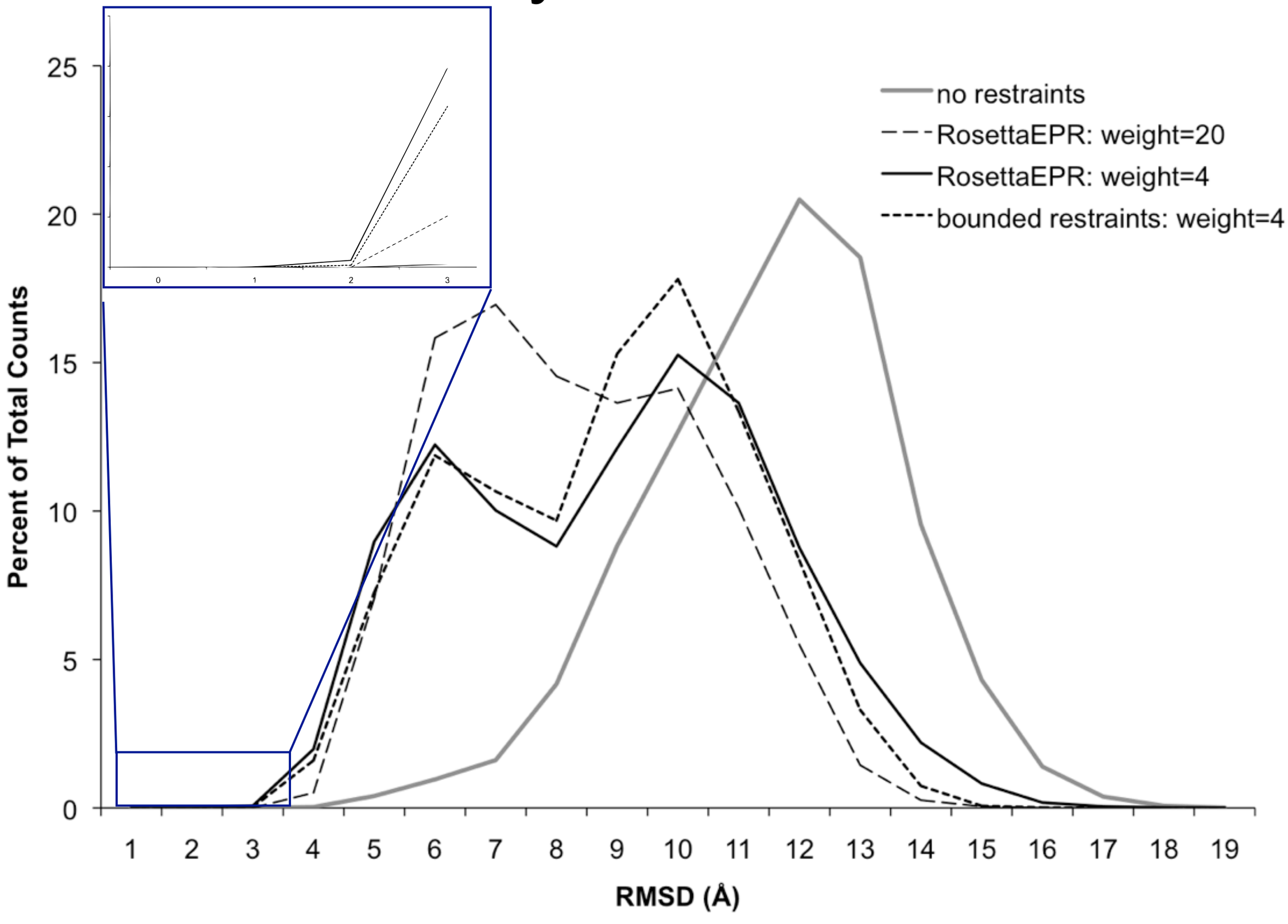
Knowledge-Based Potential Consistently Recovers the Correct Topology Better

Weight	% Models with $RMSD_{C\alpha} < 3.5Å$	% Models with $RMSD_{C\alpha} < 7.5Å$	% Models with $RMSD_{C\alpha} < 3.5Å$	% Models with $RMSD_{C\alpha} < 7.5Å$
0	0.03	7.17		
	RosettaEPR		**Bounded**	
1	0.73	21.98	0.89	37.56
2	1.41	31.07	1.18	40.95
3	2.01	37.20	1.58	41.84
4	2.05	42.08	1.62	41.09
5	1.83	45.65	1.43	40.44
6	1.60	47.29	1.40	39.50
7	1.35	49.60	1.40	38.42
8	1.31	51.21	1.62	38.01
9	0.87	50.89	1.59	37.42
10	1.02	52.70	1.57	37.22
20	0.51	54.89	1.44	34.02
30	0.46	53.28	1.22	32.77
40	0.25	49.74	1.27	32.16
50	0.17	47.43	1.12	32.27
60	0.07	43.86	1.01	31.07
70	0.03	43.95	1.29	31.67
80	0.02	43.07	1.34	31.05
90	0.01	40.92	1.39	31.22
100	0.01	41.11	1.12	30.62

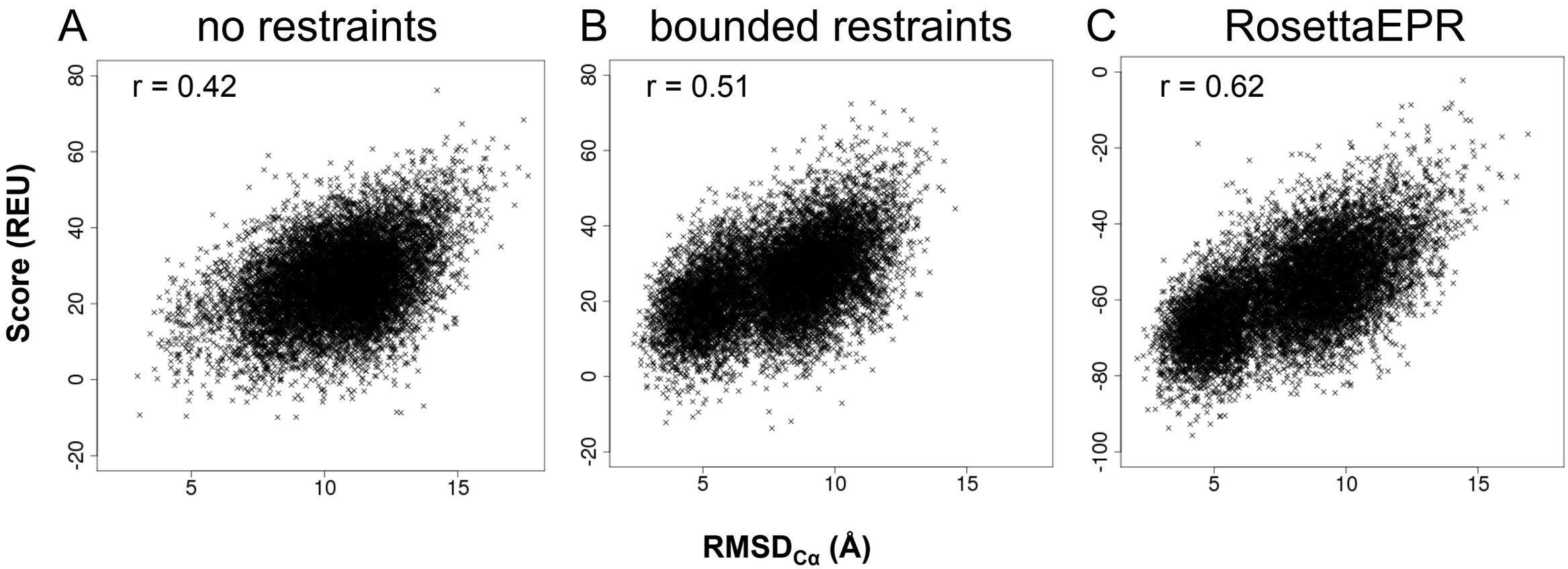
At Optimal Weight, Knowledge-Based Potential Recovers more Native-Like Models

Weight	% Models with $RMSD_{C\alpha} < 3.5Å$	% Models with $RMSD_{C\alpha} < 7.5Å$	% Models with $RMSD_{C\alpha} < 3.5Å$	% Models with $RMSD_{C\alpha} < 7.5Å$
0	0.03	7.17		
	RosettaEPR		**Bounded**	
1	0.73	21.98	0.89	37.56
2	1.41	31.07	1.18	40.95
3	2.01	37.20	1.58	41.84
4	2.05	42.08	1.62	41.09
5	1.83	45.65	1.43	40.44
6	1.60	47.29	1.40	39.50
7	1.35	49.60	1.40	38.42
8	1.31	51.21	1.62	38.01
9	0.87	50.89	1.59	37.42
10	1.02	52.70	1.57	37.22
20	0.51	54.89	1.44	34.02
30	0.46	53.28	1.22	32.77
40	0.25	49.74	1.27	32.16
50	0.17	47.43	1.12	32.27
60	0.07	43.86	1.01	31.07
70	0.03	43.95	1.29	31.67
80	0.02	43.07	1.34	31.05
90	0.01	40.92	1.39	31.22
100	0.01	41.11	1.12	30.62

RosettaEPR Enhances Sampling of Correctly Folded Models



The Knowledge-Based Potential Improves Correlation of Score and Model Quality



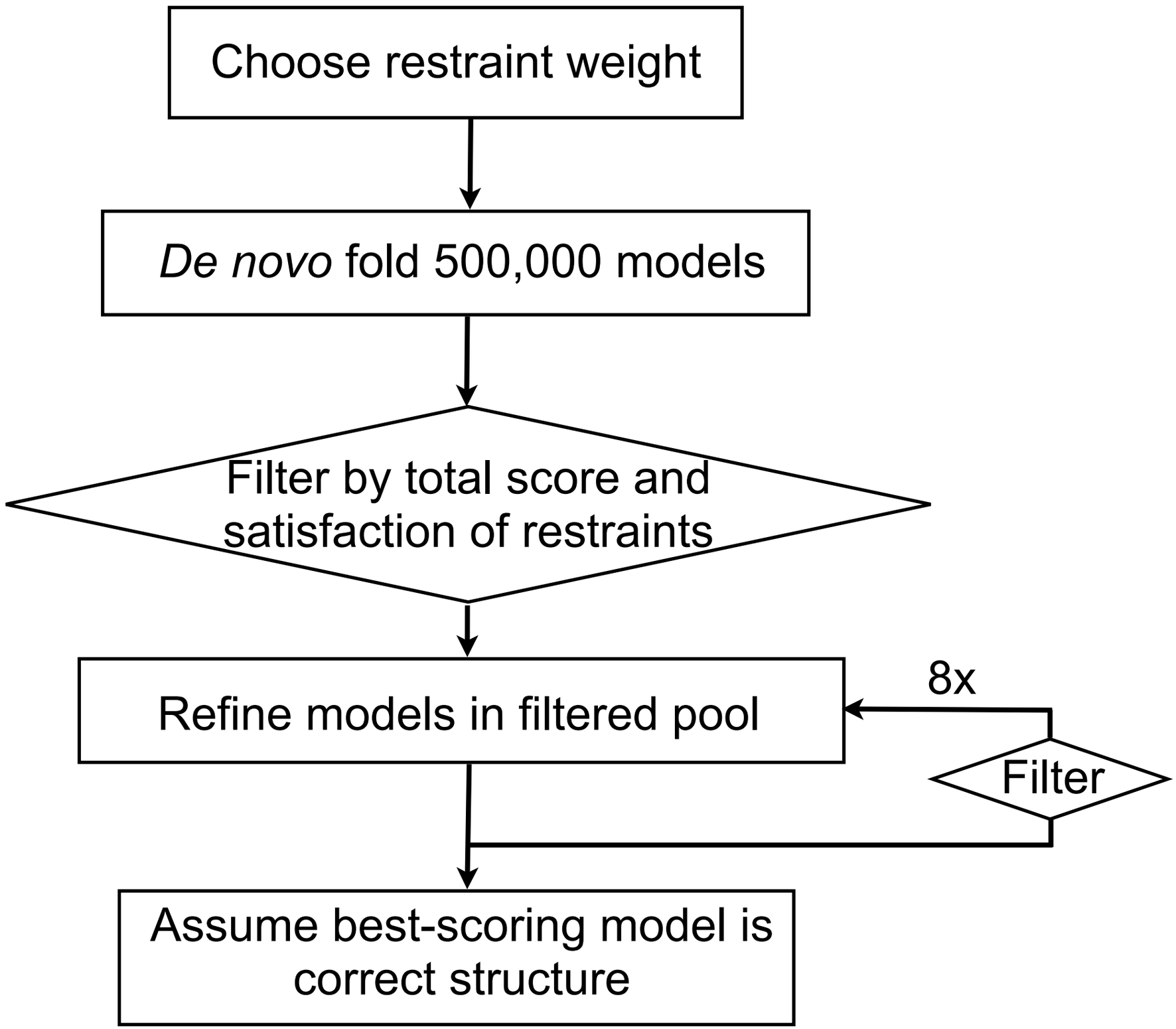
Best-Scoring Mode $RMSD_{C\alpha}$

8.24Å

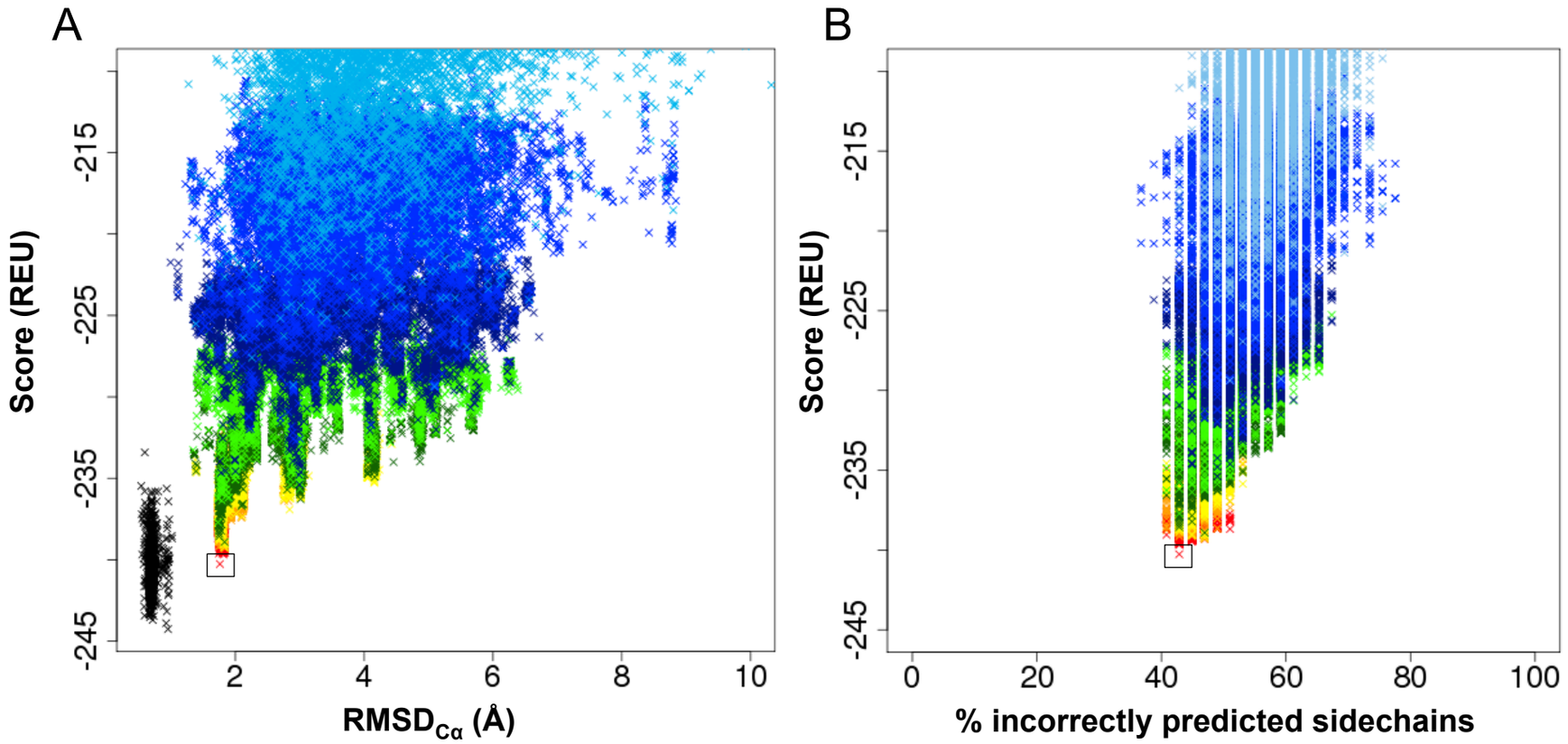
7.61Å

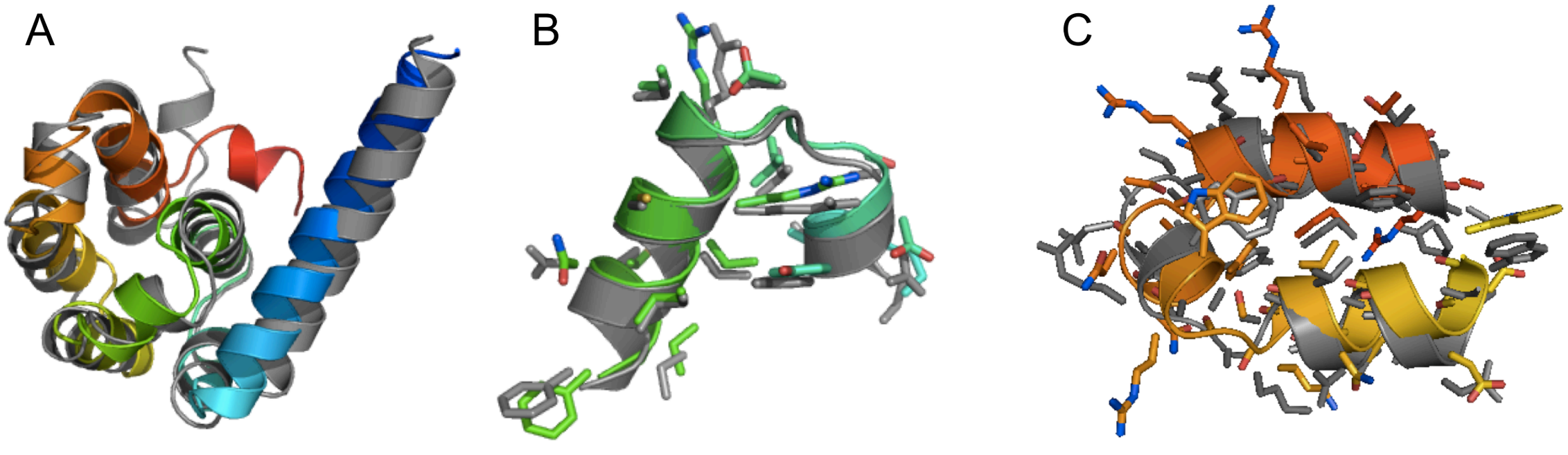
4.17Å

De novo Folded Models Must be Refined to Atomic Detail



Predicted Models of T4-lysozyme can be Refined to 1.7Å Accuracy



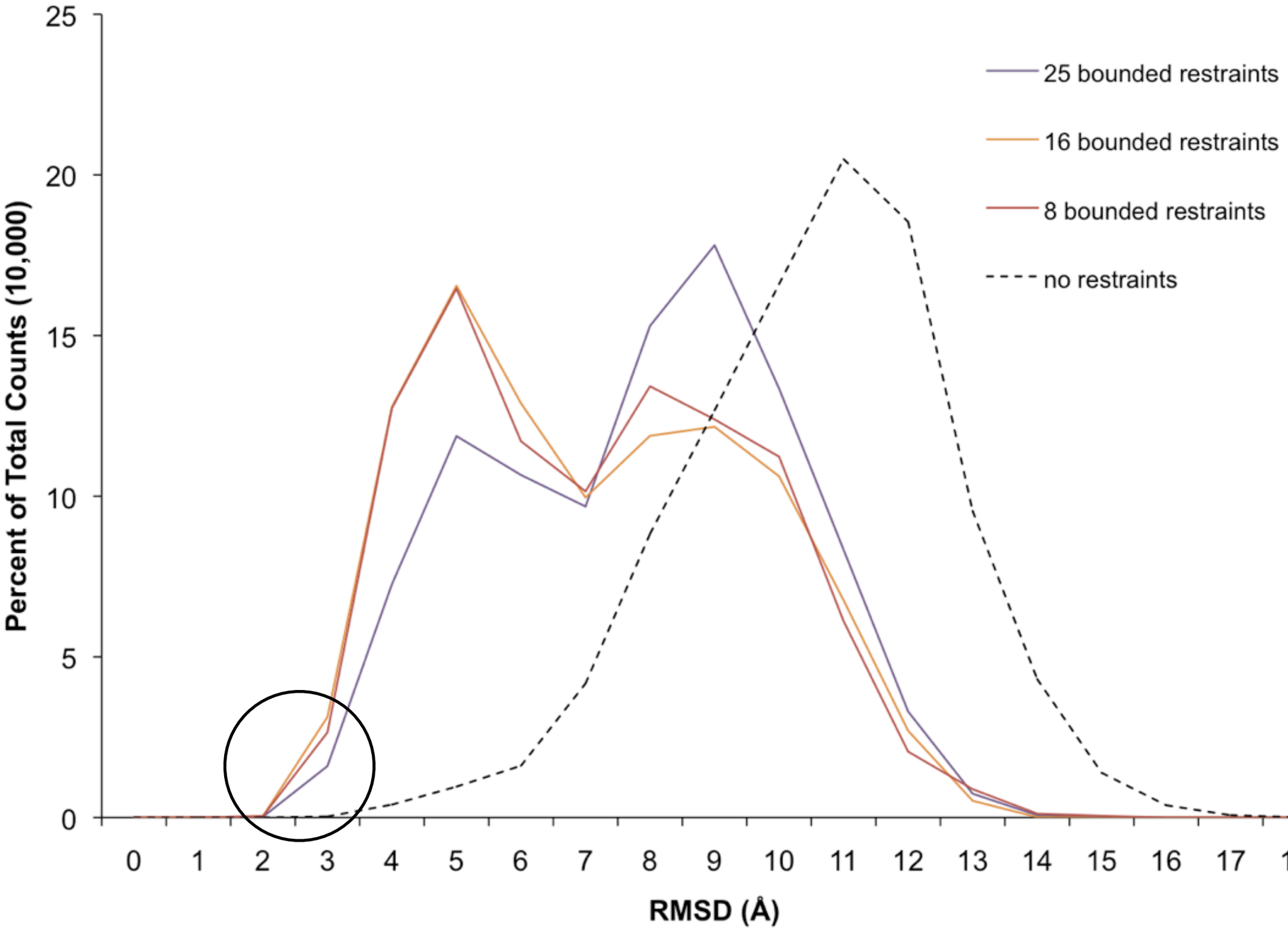


Conclusions

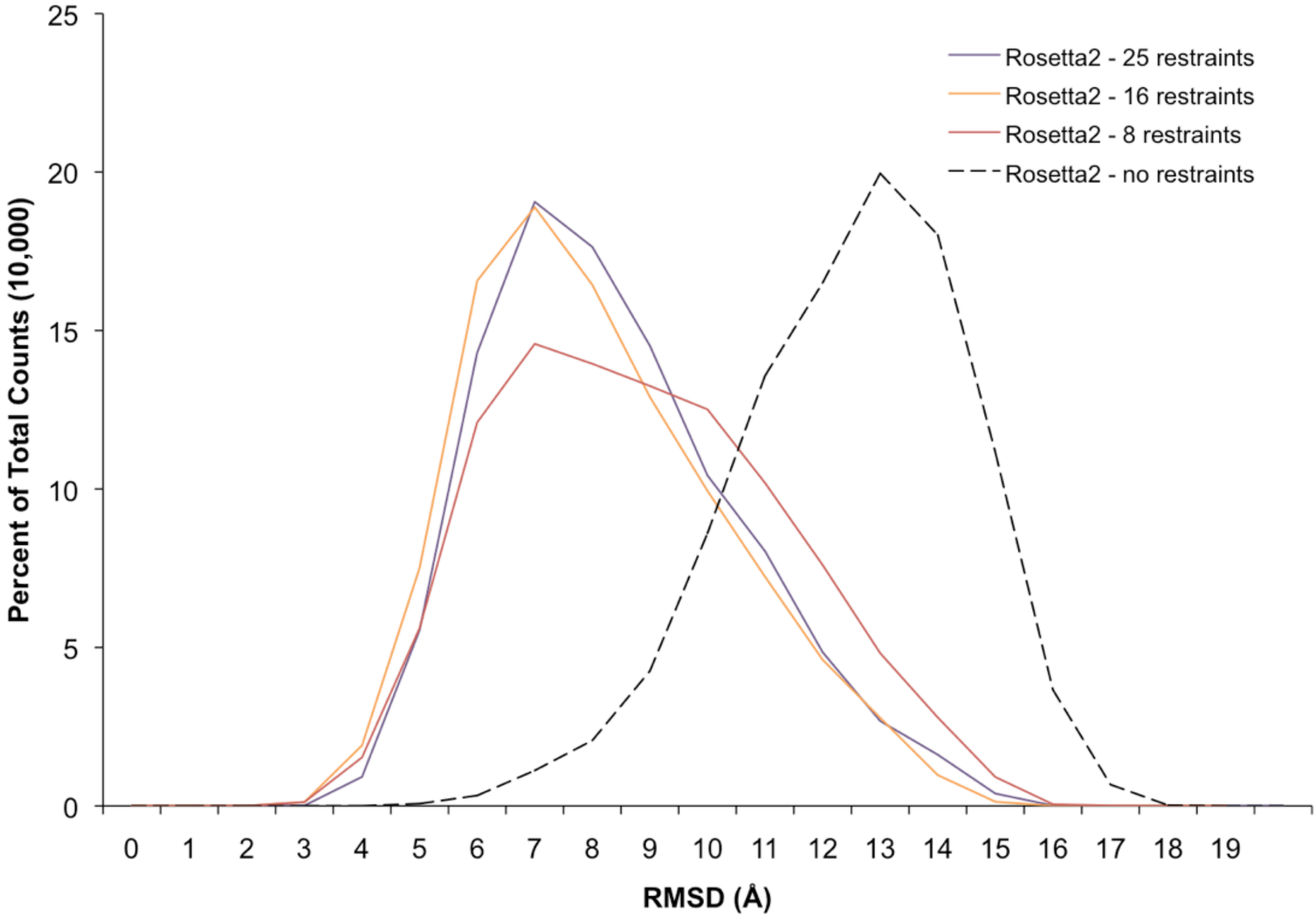
- RosettaEPR knowledge-based potential allows for better sampling of native-like folds
- The knowledge-based potential is more robust than bounded restraint potential
- Can choose atomic-detail model based solely on the Rosetta all-atom scoring function

But Wait! There's More...

Bounded Restraints Perform much Better with Fewer Restraints



But Bounded Restraints Perform About the Same in Rosetta2



So What's Really Going On?

- In Rosetta2, *de novo* folding with fewer long-range restraints scored by a bounded potential seemed to do perform the same.
- In Rosetta3, folding with fewer restraints resulted in *significantly* better recovery of correctly folded models.
- Given:
  - These protocols probably meant to be used with several short-range distances
  - Using more restraints can lead to over-penalization and restrict sampling
- *What are the differences between the Rosetta2 and Rosetta3 FoldConstraints protocols?*

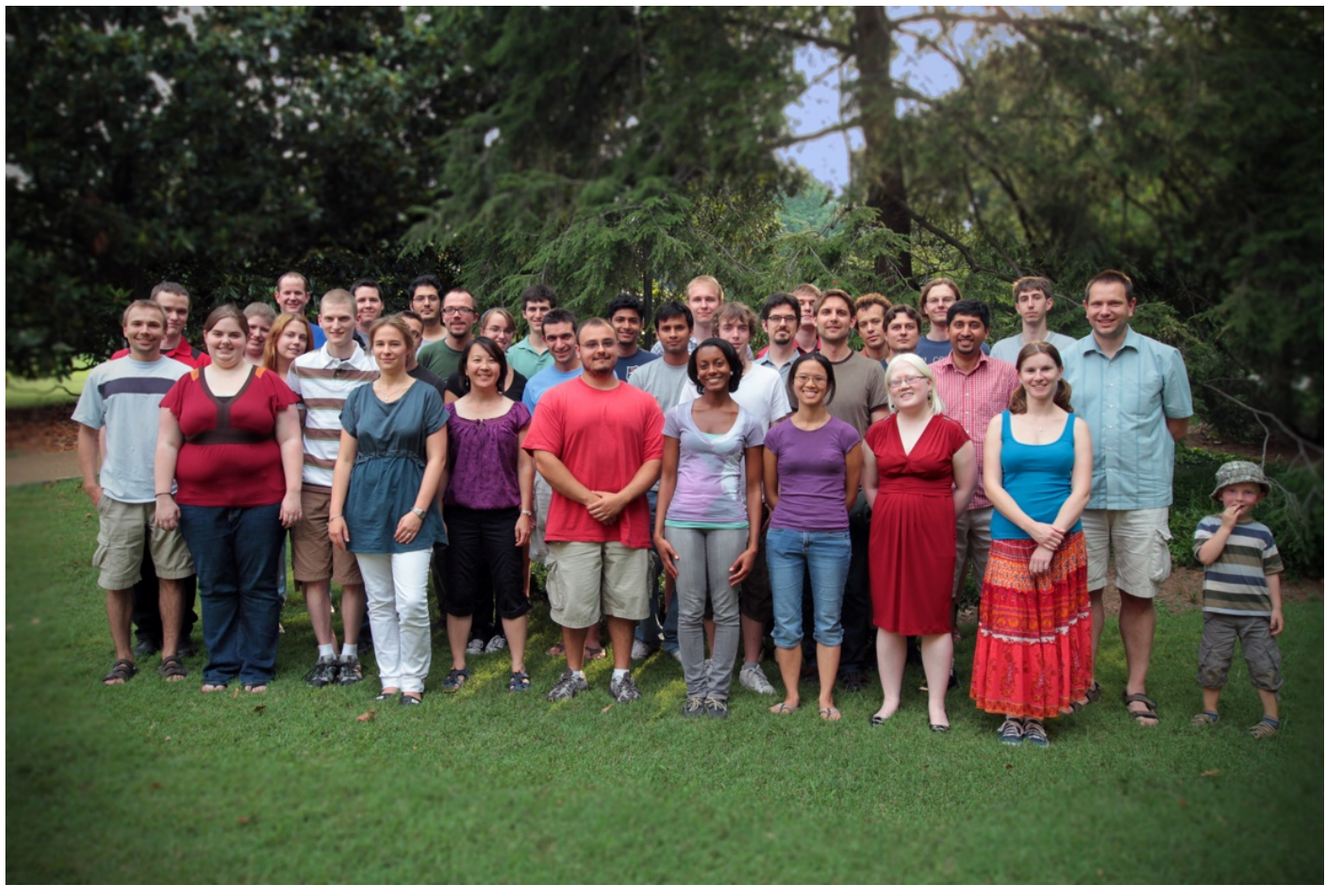
Acknowledgements

Jens Meiler
Nathan Alexander
Kristian Kaufmann
Samuel DeLuca
Gordon Lemmon
The Meiler Lab

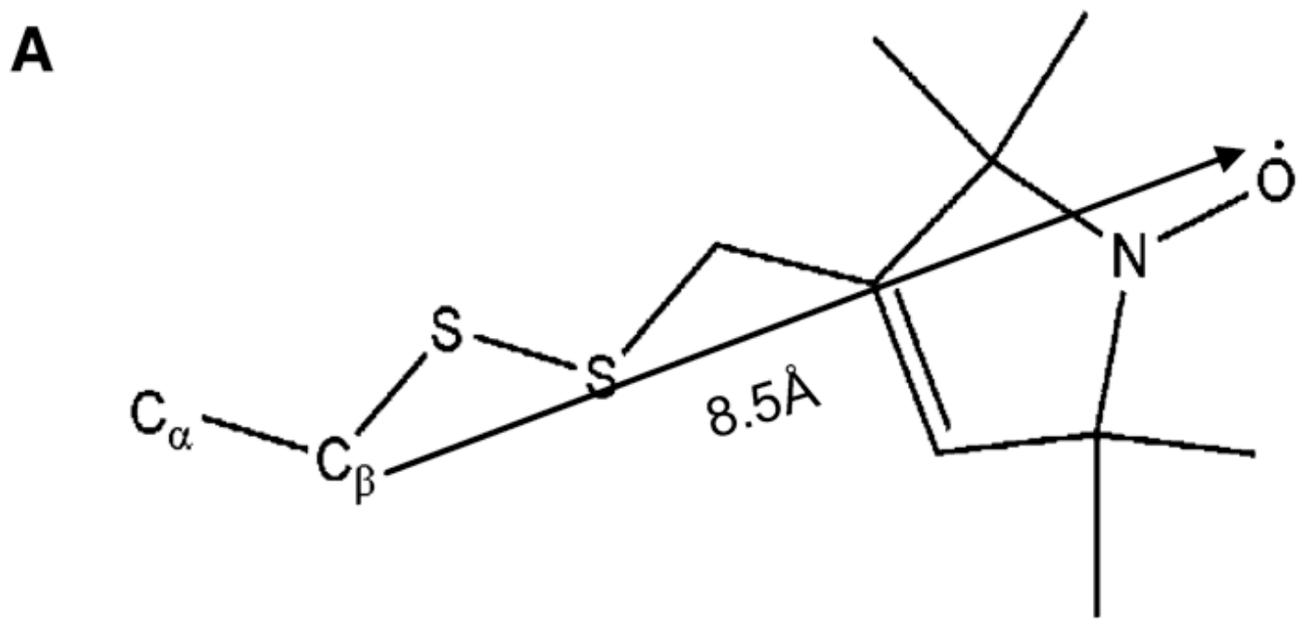
Hassane Mchaourab
Kelli Kazmier

Rosetta Community

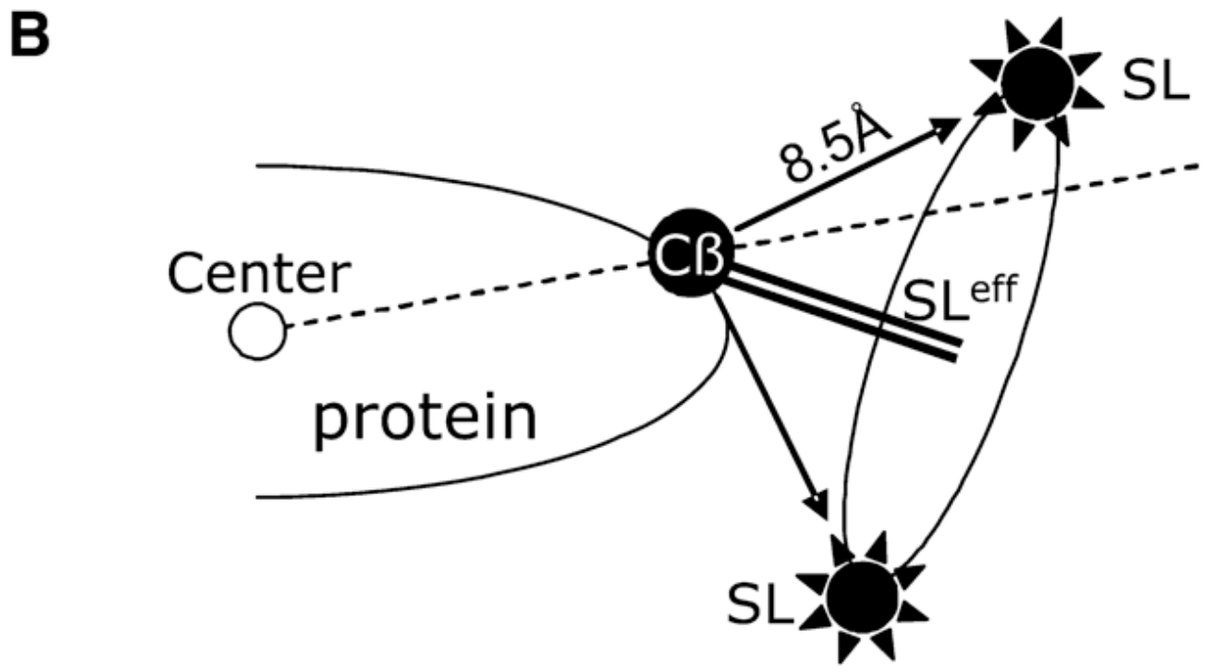
Funding:
NIH-1R01GM080403



The Cone Model Translates EPR Experimental Data to Structural Restraints



methanethiosulfonate (MTS) spin label



$d_{Cß\text{-}SL} = 8.5$ Å

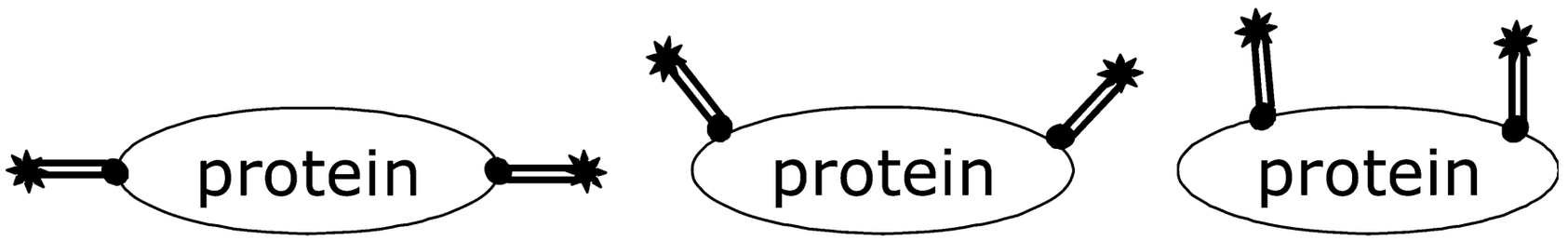
$d_{Cß\text{-}SL}^{eff} = 6$ Å

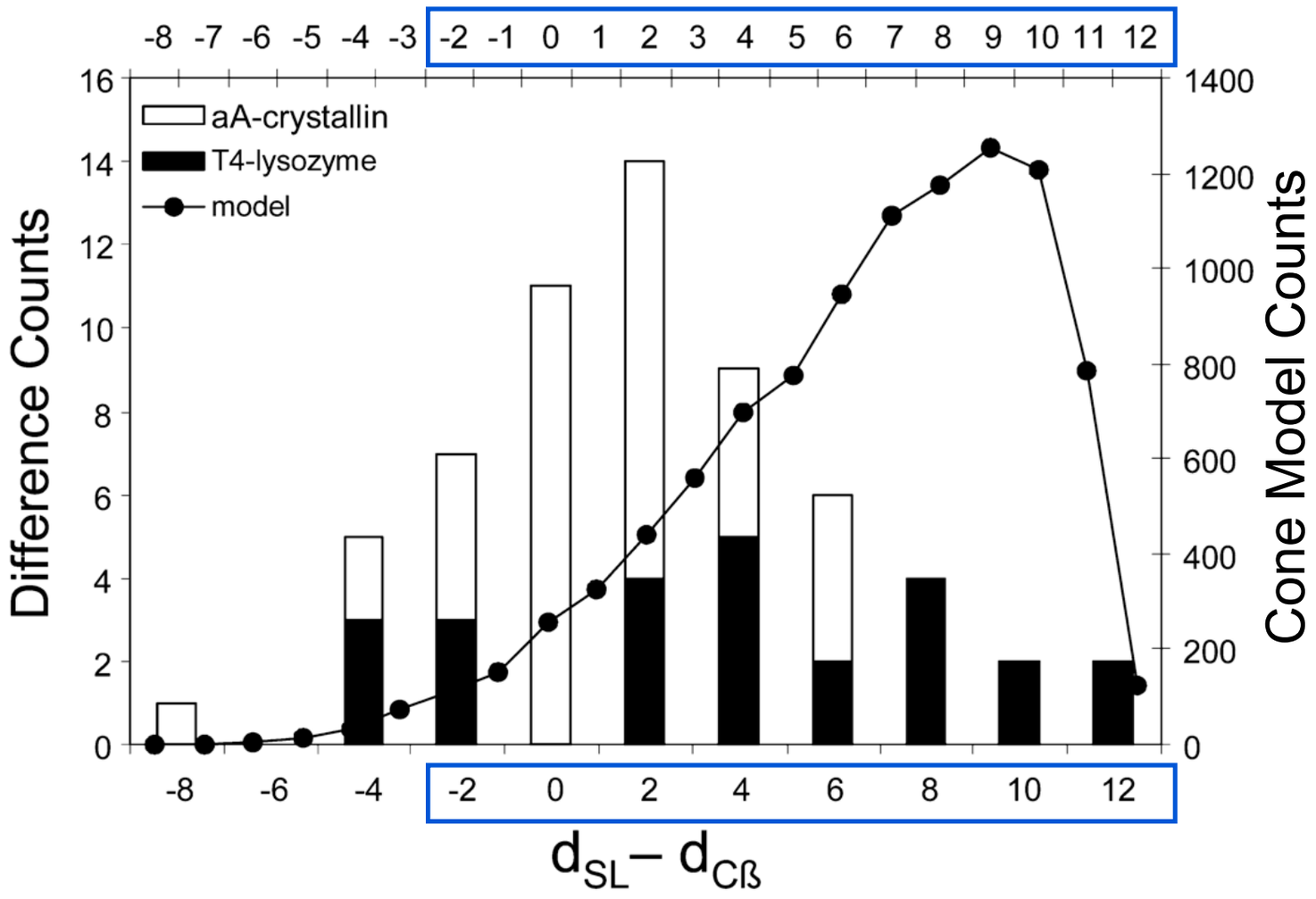
$_{SL\text{-}Cß\text{-}SL}^{eff} = 45°$

$180° = {}_{C\text{-}Cß\text{-}SL}^{eff} = 120°$

Alexander, N., *et al*, *Structure* (2008) *16*, 181-195.

The Cone Model was Used to Convert EPR Distances to a Bounded Restraint Potential

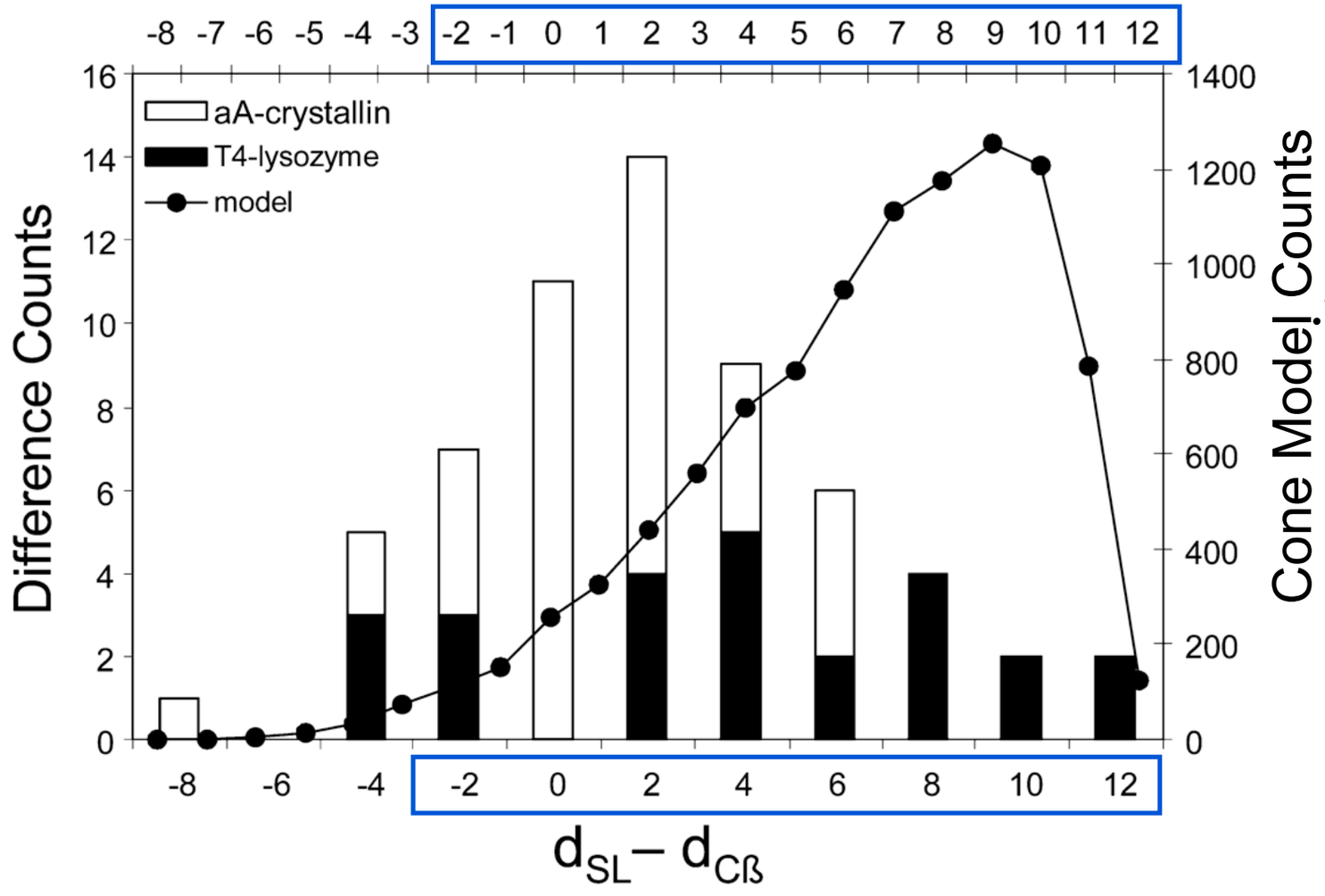




- Simulated spin label (cone model) was placed at a random position on the surface of the ellipsoid
- Computed $d_{SL} - d_{C\beta}$
- Repeated until had 10,000 distances

Alexander, N., *et al*, *Structure* (2008) *16*, 181-195.

Simplified Cone Model Can be Used to Convert Dsl-Dcb to Bounded Restraints

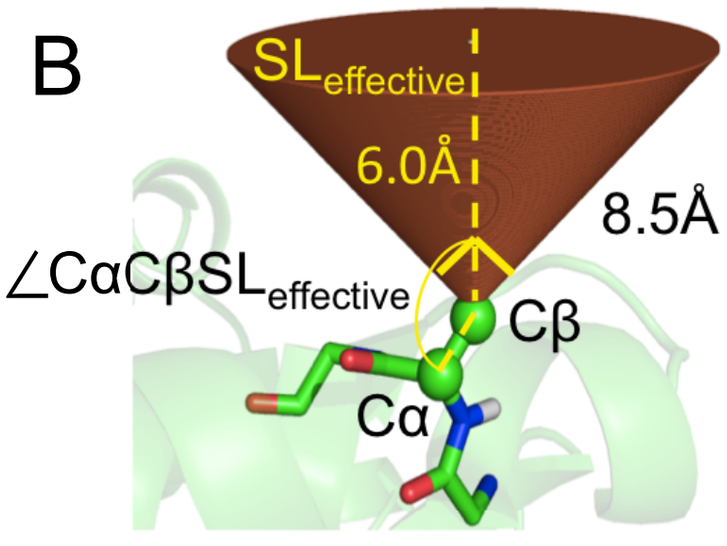


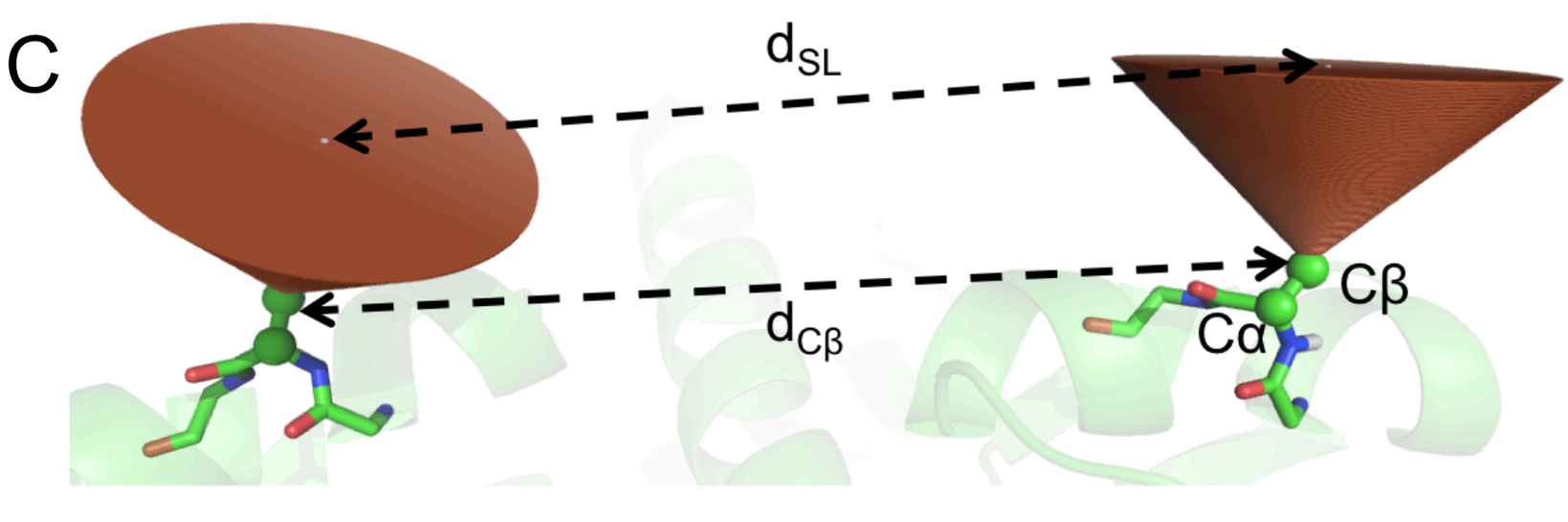
$$-2.5 < (d_{SL} - d_{C\beta}) < 12.5$$

$$\therefore (d_{SL} - \sigma_{SL} - 12.5) < d_{C\beta} < (d_{SL} + \sigma_{SL} + 2.5)$$

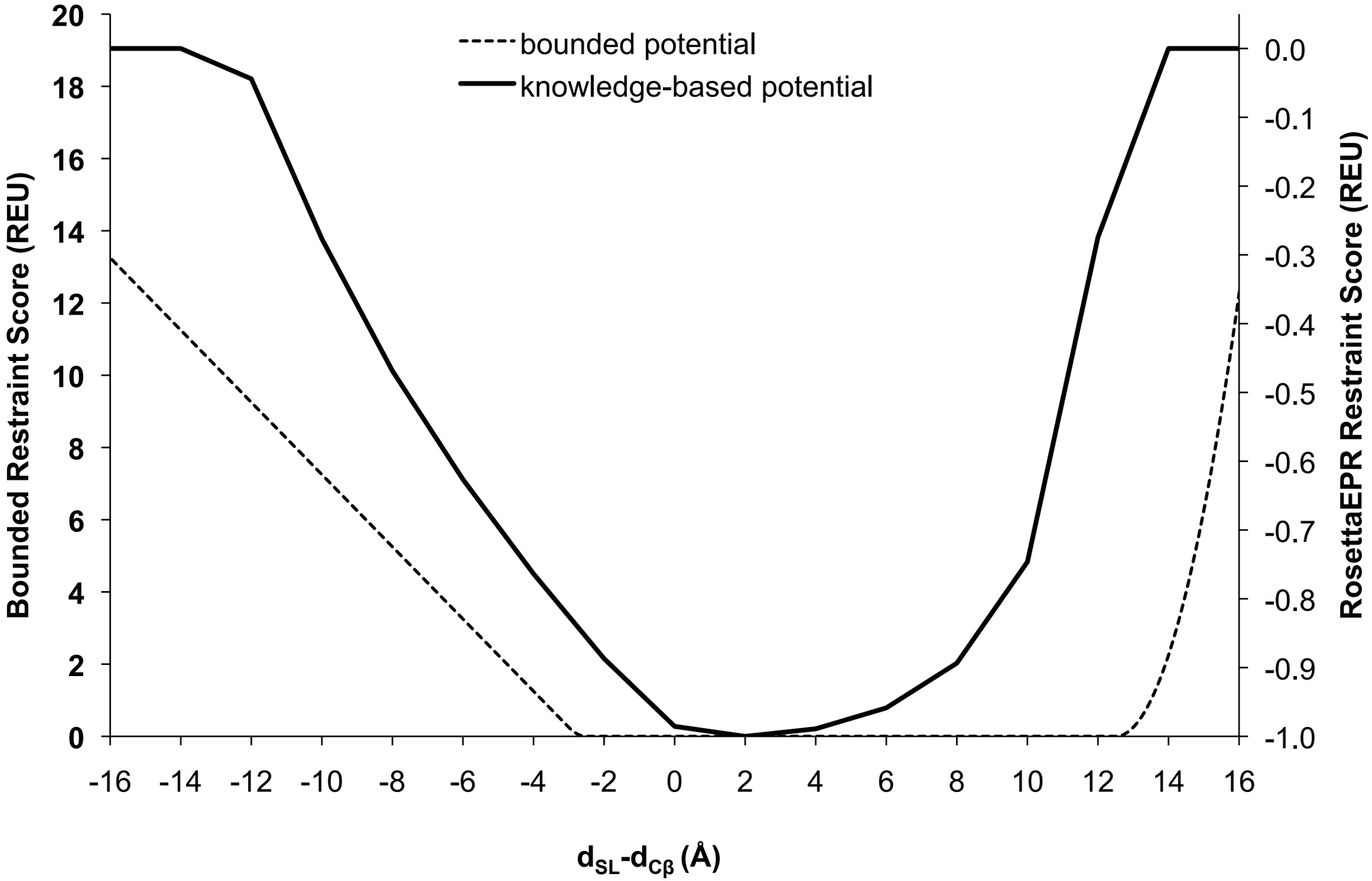
Bounded Restraint Potential Could be Improved to Contain More Information

- The previously reported cone model statistics did not reflect nuances in experimental data.
- Probably due to low-resolution of the ellipsoid model and low number of $d_{SL} - d_{C\beta}$ values computed
- Restraints used were therefore broad and lacked information content
- Expect to gain more information by using real proteins and collecting more data

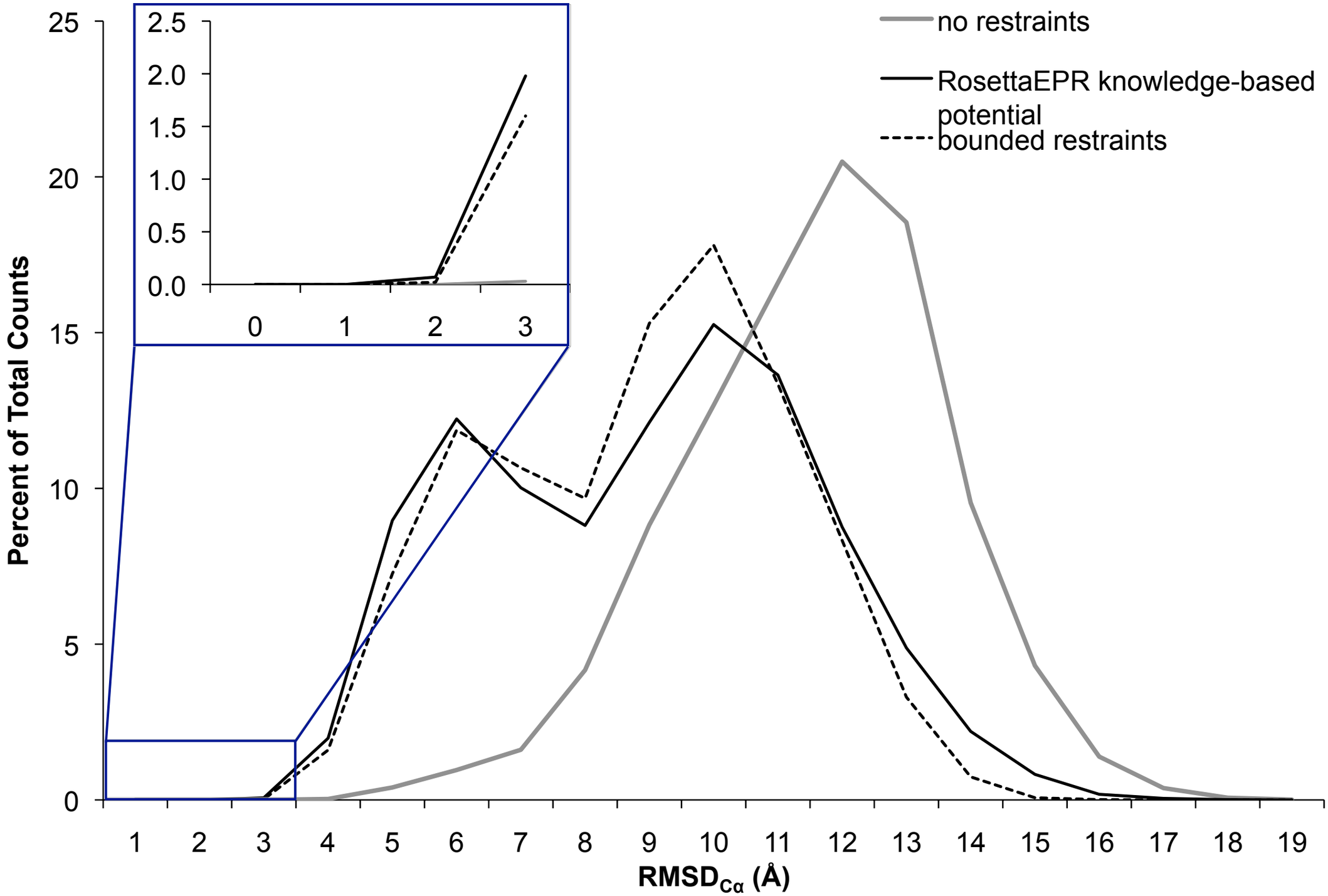




A Knowledge-Based Potential is More Informative than Simple Bounded Restraints



RosettaEPR Recovers More Native-Like Models than Bounded Restraints



Repeat Using Fewer Restraints with Highest Information Content...

Restraint Type	# Restraints	% Models RMSD < 3.5Å	% Models RMSD < 7.5Å	correlation coefficient
none	0	0.03	7.17	0.42
KB potential	25	2.05	42.08	0.62
bounded	25	1.62	41.09	0.51
KB potential	16	1.47	38.03	0.62
bounded	16	3.17	53.35	0.60
KB potential	8	1.52	34.44	0.59
bounded	8	2.69	53.75	0.59

Rosetta2 testing information content...

Restraint Type	# Restraints	% Models RMSD < 5Å
none	0	0.00
bounded	25	0.94
bounded	16	2.03
bounded	8	1.65