Supplementary Information: A simple and efficient statistical potential for scoring ensembles of protein structures

Pilar Cossio,[1] Daniele Granata,[1] Alessandro Laio,[1] Flavio Seno,[2] and Antonio Trovato[2]

[1] *SISSA - Scuola Internazionale Superiore di Studi Avanzati, via Bonomea 256, I-34136 Trieste, Italy*

[2] *Università degli Studi di Padova, and CNISM, Unità di Padova, I-35131 Padova, Italy*

I. SUPPLEMENTARY TEXT

A. Molecular dynamics and bias-exchange metadynamics simulations

Molecular dynamics simulations were performed using the GROMACS 4.5.3 package[1], employing the AMBER99-ILDN[2] force field, using TIP3P water model[3] for the explicit solvent. For protein GB3, the system has been solvated by 6524 water molecules in a 223 nm^3 cubic periodic box. For protein 2VWR, the native conformation has been solvated by 11037 water molecules in a 351 nm^3 cubic periodic box, whereas the two decoy structures by 7339 (8354) water molecules in a 238 (296) nm^3, respectively. Model #1 of protein 2K49 has been solvated by 7780 water molecules in a 254 nm^3 cubic periodic box. The particle-mesh Ewald method was used for long-range electrostatic with a short-range cutoff of 1 nm. All bond lengths were constrained to their equilibrium length with the LINCS algorithm[4]. The time step for the MD simulation has been set to 2.0 fs,the Nose-Hoover thermostat[5] and the Parrinello-Rahman barostat[6] with a relaxation time of 2 ps were used. The atomic coordinates and the energy were saved every ps.

Bias exchange metadynamics together with the OBC [7] implicit solvent model and $T = 400K$ were used to enhance the conformational searching[8]. The PLUMED plugin[9] was used to perform the BE simulation. We have used four replicas of the system, and each replica is biased over one of the following collective variables:

- The *Coordination Number*, is used to count the number of contacts inside the protein[9]. Its Gaussian width is σ=10.
- Alpharmsd (σ=0.2), Parabetarmsd (σ=0.1), Antibetarmsd (σ=0.2) collective variables are used to explore respectively the content in α-helix, parallel and anti-parallel β-sheets [10].

On each walker, we deposited one-dimensional hills with a height of 0.2 kJ/mol every 2 ps, and we attempted exchanges of the bias potentials between the replica every 20 ps.

II. SUPPLEMENTARY FIGURES

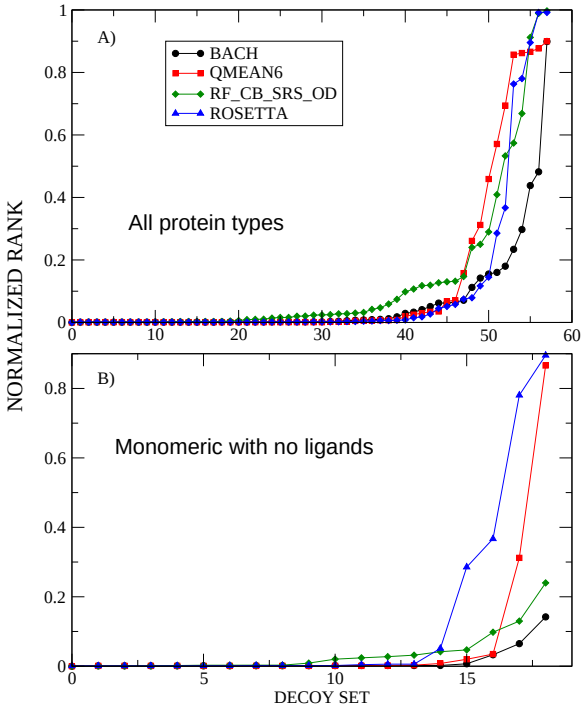

Figure S1: Sorted normalized rank for the native structure, for A) all the sets, including native states with bound partners (e.g. ligands or other protein chains), and (B) only monomeric native conformations with no ligands in the semfold, 4state, fisa, and RosettAll decoy sets.

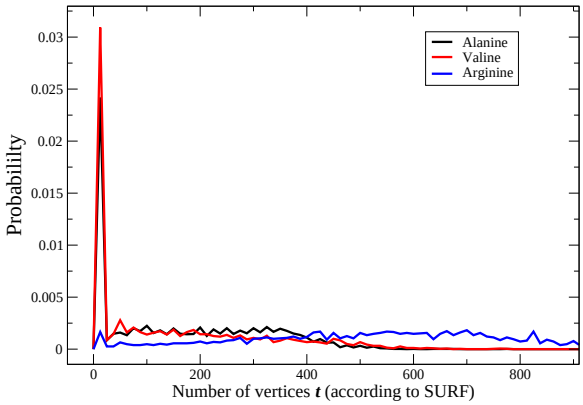

Figure S2: Distribution of the values of the number of triangle vertices observed (t), according to SURF[11], in the data set TOP500[12] for alanine, valine and arginine residues.

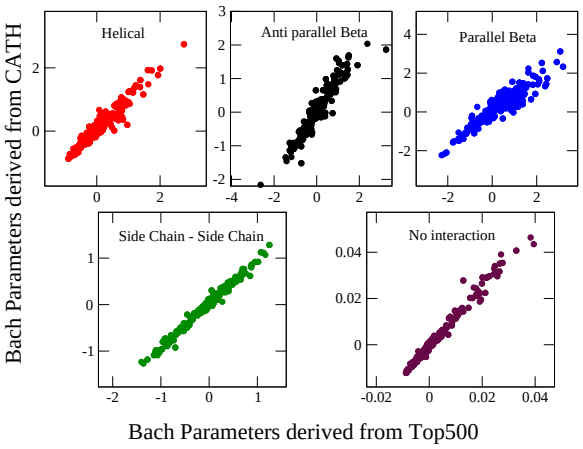

Figure S3: Correlation between the BACH pairwise contact parameters derived from 8000 CATH structures and from the Top500 dataset.

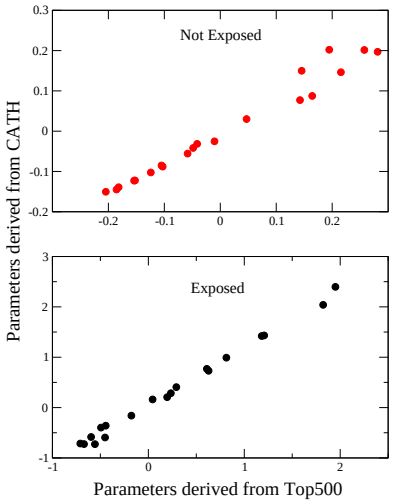

Figure S4: Correlation between the BACH solvation parameters derived from 8000 CATH structures and from the Top500 dataset.

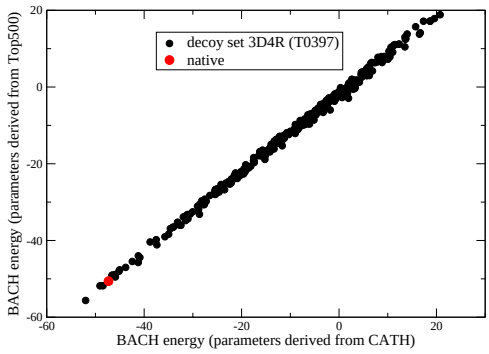

Figure S5: Correlation between the BACH energies computed for the structures in the CASP8 (T0397) decoy set, using parameters derived from two different data sets (Top500 and CATH).

III. SUPPLEMENTARY TABLES

Table S1: Decoy sets from CASP 8-9 targets used in this work: PDB code (CASP target code)

3PNX (T0517)	3NRD (T0522)	3NRE (T0526)	2L0B (T0539)
2L0D (T0541)	2L3W (T0544)	2KRX (T0562)	3ON7 (T0563)
2KYW (T0569)	3NRQ (T0575)	2KY9 (T0579)	3NI8 (T0594)
3NKH (T0623)	3O1L (T0626)	3NUW (T0628)	3CYN (T0388)
3D4R (T0397)	3D6W (T0415)	3CZX (T0425)	3D3Y (T0427)
3DAI (T0432)	3D7L (T0433)	2K3I (T0437)	3DCP (T0440)
3DAO (T0445)	3DO6 (T0447)	3DMC (T0451)	2K5W (T0468)
2K49 (T0472)	3DLC (T0485)	2VWR (T0488)	3DLM (T0504)
3E03 (T0511)			

Table S2: List of 50 randomly chosen structures from the Top500 dataset.

1a2zAH	1a8iH	1aieH	1aohBH	1atzAH
1b3aAH	1babBH	1benABH	1bhpH	1bqcH
1bu7AH	1c02AH	1c75H	1chdH	1cpqH
1cxyH	1dbgH	1difAH	1dqsAH	1ek6AH
1fasH	1fncH	1gciH	1h2rSH	1hxnH
1jhgABH	1lbuH	1mfmH	1mrjH	1nddBH
1nulBH	1pdoH	1ptfH	1qfmH	1qhvH
1qnjH	1qtsH	1rhsH	1svyH	1thvH
1uaeH	1vccH	1whiH	1zinH	2bbkHH
2ctcH	2garH	2mhrH	2rn2H	3chbDH

[1] E. Lindahl, B. Hess and D. van der Spoel, *GROMACS 3.0: a package for molecular simulation and trajectory analysis.* J. Mol. Mod. 7: 306-317 (2001)

[2] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, S. L. Klepeis, R. Dror and D. E. Shaw, *Improved side-chain torsion potentials for the Amber ff99SB protein force field* Proteins: Struct., Funct., Bioinf. 78: 1950-1958 (2010)

[3] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *Comparison of simple potential functions for simulating liquid water* J. Chem. Phys. 79: 926-935 (1983)

[4] B. Hess, H. Bekker, H. J. C. Berendsen, G. E. M. J. Fraaije, *LINCS: A linear constraint solver for molecular simulations* J. Comput. Chem. 18: 1463 (1997)

[5] W. G. Hoover, *Canonical dynamics - Equilibrium phase-space distributions* Phys. Rev. A 31: 1695, (1985)

[6] M. Parrinello, and A. Rahman. *Polymorphic transitions in single crystals: A new molecular dynamics method.* J. App. Phy. 52: 7182-7190(1981).

[7] A. Onufriev, D. Bashford, D. A. Case, *Exploring protein native states and large-scale conformational changes with a modified generalized born model* Proteins: Struct., Funct., Bioinf. 55: 383-394 (2004)

[8] P. Cossio , A. Trovato , F. Pietrucci. F. Seno , A. Maritan and A. Laio, *Exploring the universe of protein strucutres beyond the protein data bannk*, Plos Compt. Biol 6: 11 (2010)

[9] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia and M. Parrinello, *PLUMED: a portable plugin for free-energy calculations with molecular dynamics* Comp. Phys. Comm. 180: 1961 (2009)

[10] F. Pietrucci and A. Laio, *A collective variable for the efficient exploration of protein beta-sheet structures: application to SH3 and GB1* J. Chem. Theory Comput. 5, 2197 (2009)

[11] A. Varshney and F. P. Brooks and W. V. Wright. *Computing smooth molecular-surfaces.* IEEE Computer Graphycs and applications. 14, 19-25 (1994).

[12] S.C. Lovell , I.W. Davis, W.B. Arendall, P.I. de Bakker , J.M. Word , M.G. Prisant , J.S. Richardson, D.C. Richardson DC., *Structure validation by Calpha geometry: phi,psi and Cbeta deviation*, Proteins, 50: 437-50 (2003)