Supporting information

Explicit solvent hydration benchmark for proteins with application to PBSA method

Piotr Setny* and Anita Dudek

Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland

E-mail: p.setny@cent.uw.edu.pl

1 Regularisation of free energy changes

If we consider three independent free energy changes f_i^0, $i \in \{1,2,3\}$, each with normally distributed error of variance σ_i^2, than their weighted mean square error (MSE) with respect to any three arbitrary assumed free energy changes, f_i, would be MSE$= \sum_{i=1}^{3} w_i(f_i - f_i^0)^2$, where $w_i \sim \sigma_i^{-2}$. Now, if we impose that the three assumed free energy changes must form a closed cycle, that is fulfil a condition: $f_1 = f_2 + f_3$ (Fig. 1, left), we can seek for such their values that minimise the above MSE, subject to this condition. This leads to the following minimisation problem,

$$\mathrm{MSE}_{ABC} = \sum_{i=1}^{3} w_i(f_i - f_i^0)^2 + \lambda_0(f_1 - f_2 - f_3) \tag{1}$$

whose solution, using Lagrange multiplier (λ_0), gives a set of three regularised free energy changes, f_i^{opt} that form a closed cycle, and provide for a minimum error with respect to the reference values.

The problem is formally analogous to finding the minimum potential energy of three connected springs, with spring constants $k_i = \sigma_i^{-2}$, allowed to relax in one dimension. Extending this analogy and finding the resultant "spring constant", $\tilde{k}_{XY} = \tilde{\sigma}_{XY}^{-2}$ between any two connection points X, Y, gives an easy way for obtaining an error estimate of now regularised free energy changes. For the case depicted in Fig. 1, left, we get:

$$\tilde{k}_{AB} = \frac{k_{12} + k_{13} + k_{23}}{k_2 + k_3}, \tag{2}$$

where $k_{ij} = k_i k_j$.

This methodology can be extended to a set of six possible free energy changes between four states (Fig. 1, right). In such case one can identify four conditions, whose fulfilment is required for proper regularisation

*To whom correspondence should be addressed

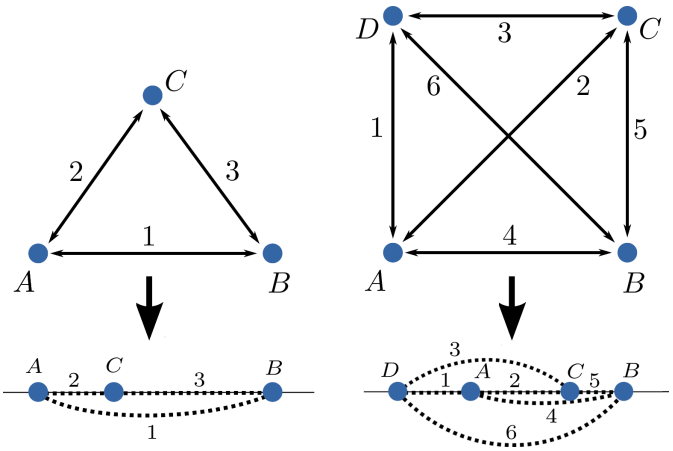

Figure 1: The equivalence between thermodynamic cycles and a system of points and springs for three (left) and four (right) structures.

(i.e. closure of any occurring thermodynamic cycle), which leads to the following expression for MSE:

$$\mathrm{MSE}_{ABCD} = \frac{1}{2}\sum_{i=1}^{6} k_i(f_i - f_i^0)^2 + \lambda_1(f_3 - f_1 - f_2) + \lambda_2(f_6 - f_4 - f_1) + \lambda_3(f_4 - f_2 - f_5) + \lambda_4(f_6 - f_3 - f_5) \quad (3)$$

It can be solved analytically through simple algebra. As above, the resultant spring constant between any two connection points can serve to obtain the uncertainty of now regularised free energy changes. For the case considered in Fig. 1, right, we get:

$$\tilde{k}_{BD} = \frac{k_{123} + k_{135} + k_{125} + k_{234} + k_{345} + k_{245} + k_{134} + k_{145}}{k_{12} + k_{13} + k_{14} + k_{23} + k_{24} + k_{25} + k_{35} + k_{45}} + k_6 \quad (4)$$

2 Finite size corrections (FSC)

According to numerical scheme proposed by Rocklin at al.,[1] a correction to hydration free energy, ΔF, due to finite size effects and related artificial periodicity in explicit solvent simulations (e) is estimated as a difference in hydration free energy of the system, evaluated with Poisson-based implicit electrostatics (i), in periodic, (P), and non-periodic (N) conditions:

$$\Delta\Delta F^{P\to N} = \Delta F^{N,i} - \Delta F^{P,i} \quad (5)$$

Given a change in hydration free energy when going from conformation A to B, evaluated in explicit solvent under periodic conditions, the following correction is thus necessary to obtain the result for nonperiodic case:

$$\Delta F^{N,e}_{A\to B} = \Delta F^{P,e}_{A\to B} - \Delta\Delta F^{P\to N,i}_{A} + \Delta\Delta F^{P\to N,i}_{B} = \Delta F^{P,e}_{A\to B} + \mathrm{FSC} \quad (6)$$

In order to test this methodology, we evaluated separation-dependent hydration free energy changes for two pairs of atomic ions of the size of a united-atom methane molecule[2] with $q_1 = -q_2 = 1e$, and $q_1 = q_2 = 1e$, solvated in cubic boxes with different sizes, considered under periodic boundary conditions (simulation methodology was the same as described for protein systems in the main text). As evidenced in Fig. 2, box size dependent differences in hydration free energy profiles reach ~ 5 kcal/mol, but become indistinguishable after applying finite size corrections, as described in Eq. 6. We note, that since the net system charge remained constant in both cases, the correction worked equally well for neutral and charged system, without the need for explicitly accounting for background neutralising charge.[1]

In order to further test the correction scheme on more complex solute, we recalculated all results for protein G, originally obtained using cubic box of side $L = 59$ Å, with extremely small simulation box ($L = 49$ Å), and two conformational changes with a larger box ($L = 69$ Å). As can be seen in Table 1, in most cases the agreement between hydration free energy differences obtained for different box sizes is poor. All discrepancies vanish after inclusion of finite size correction, which reaches almost 10 kcal/mol for the

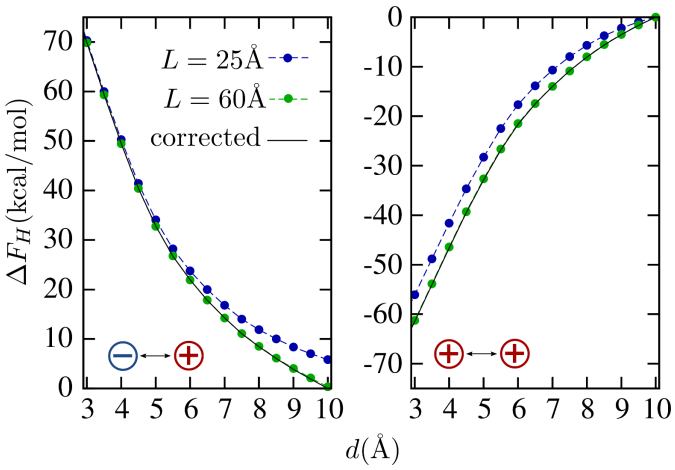

Figure 2: Hydration free energy as a function of charge separation d, obtained with explicit solvent simulations using two different cubic box sizes, L, and periodic boundary conditions. Solid line denotes results with FSC, which become indistinguishable.

case with the largest change in protein dipole moment (Table 2).

Table 1: Hydration free energy changes for protein G and finite size corrections, obtained for three simulation boxes with different sizes. All results in kcal/mol.

$A \to B$	L [Å]	$\Delta\tilde{F}^{P,e}$	FSC	$\Delta\tilde{F}^{N,e}$
b, u	49	-201.5	5.9	-195.6
	59	-197.1	2.8	-194.3
	69	-196.4	1.7	-194.7
h, u	49	-78.1	9.7	-68.4
	59	-72.9	4.1	-68.8
	69	-71.1	2.6	-68.5
b, m	49	-122.2	-3.0	-125.2
	59	-124.3	-1.0	-125.3
b, h	49	-119.4	-3.8	-123.2
	59	-124.1	-1.4	-125.5
h, m	49	-0.7	0.8	0.1
	59	-0.4	0.4	0.0
m, u	49	-77.5	8.9	-68.6
	59	-71.8	3.8	-68.0

3 Protein structures

As a source of benchmark structures we selected 5 proteins representing diverse folds, with polypeptyde chain lengths ranging from 10 to 238 amino acids. Four of them served to generate 4 sets, each containing 4 distinct conformations, and one, λ-repressor, a single set with 10 conformations. In each case, protein structure was extracted from respective file from the Protein Databank (PDB) [3] and parametrised with Amber-ff-99 force field,[4] with default protonation states assigned by `pdb2gmx` GROMACS[5] tool. Particular conformations were obtained as follows:

chignolin (CH): *folded, (f)*, state was taken as an NMR structure (PDB id. 1uao), with potential energy minimised *in vacuo*, using 1000 steepest descent steps, with harmonic restraints of 2.4 kcal/mol/Å^2 put on heavy atoms; *misfolded, (m)*, *prefolded, (p)*, and *unfolded, (u)*, states were extracted from explicit solvent folding simulations (see below) as centroids of clusters representing wrongly folded state, kinetic trap just before the final folding event, and an arbitrary selected conformation of unfolded polypeptyde chain.

protein G (PG): *beta* state was extracted from PDB structure (PDB id. 1qkz, chain A, residues 45-61, sequence GEWTYDDATKTFTVTE) and energy minimised as above; *misfolded, (m)*, and *unfolded, (u)*, states were generated during explicit solvent, unfolding simulation at temperature of 400 K, as centroids of clusters representing: a meta-stable state, and an arbitrary conformation of unfolded polypeptyde chain, respectively; *helix, (h)*, state was generated as a canonical α-helix and energy minimised as above;

λ-repressor (LR): in this case we considered 10 conformations which were divided into 3 sets, sharing one common member that represented energy minimised crystal (*native, n)* structure (PDB id. 1lmb); *small rms* set (LR_S) contained 3 additional structures (*s1 - s3*) that were extracted from explicit solvent simulation at temperature of 300 K, such that pairwise C_α root mean square distance (RMSD) between them and the native structure was below 1.2 Å; *intercluster* set (LR_I), contained additional 3 structures (*i1 - i3*) representing centroids of 3 most populated clusters (obtained with C_α RMSD threshold of 2.5 Å) from the same simulation; *unfolding* set (LR_U) contained 3 additional structures (*u1 - u3*) gathered from simulation at temperature of 400 K: two representing centroids of clusters obtained with C_α RMSD threshold of 4.0 Å, and one taken from a simulation frame with minimal content of secondary structure, as evaluated by DSSP method[6] implemented in `do_dssp` GROMACS tool.

adenylate kinase (AK): *open, (o)*, and *closed, (c)*, conformations, representing an apo and a ligand-bound state (ligand structure was removed), were obtained from energy-minimised crystal structures (PDB id: 4ake, 1ake, respectively); *semi, (s)*, and *semi2, (s2)*, conformations were taken from MD simulations (100 ns of production run; see below) that were started from the open and closed states, respectively, as frames with the lowest RMSD with respect to their opposite – closed and open – states, respectively.

lao binding protein (LP): *open, (o)*, *closed, (c)*, *semi, (s)*, and *semi2, (s2)*, states were derived in the same manner as in the case of adenylate kinase, with crystal structures 2lao, and 1lst representing the apo and ligand bound conformations, respectively.

4 Protein descriptors

Table 2: Characteristics of considered protein conformations. S: surface area (MD-based), V: solvent excluded volume (MD-based), b.at: fraction of buried atoms, $|d|$: dipole moment (as calculated by `gmx dipole` GROMACS tool), $\langle \mathrm{HB}_{pw} \rangle$ - an average number of protein-water hydrogen bonds (as calculated by `gmx hbond` GROMACS tool), helix, beta, turn: proportions of respective secondary structure motifs, HSn, the number of surface hydration sites with peak water density $\rho \geq n\rho_0$.

protein	S [Å^2]	V [Å^3]	b. at.	$\lvert d \rvert$ [D]	$\langle \mathrm{HB}_{pw} \rangle$	helix	beta	turn	HS5	HS6
CH										
folded	859	1503	0.4	64	32	0	0.4	0.2	0	0
misfolded	851	1486	0.3	60	35	0	0.4	0.3	1	0
prefolded	924	1520	0.3	64	40	0	0	0	1	0
unfolded	1056	1540	0.2	84	43	0	0	0.2	1	0
PG										
beta	1438	2655	0.4	103	55	0	0.5	0.3	1	0
helix	1517	2602	0.3	162	52	0.7	0	0.1	1	0
misfolded	1513	2499	0.3	97	65	0.2	0	0	6	1
unfolded	1908	2716	0.2	43	76	0	0	0	2	0
LR										
n	4714	12622	0.6	128	185	0.6	0	0.2	10	0
u1	4614	12589	0.6	57	159	0.6	0	0.2	5	0
u2	4422	12625	0.6	121	167	0.7	0	0.1	7	0
u3	5423	13209	0.5	274	182	0.4	0	0.1	8	2
i1	4805	12729	0.5	162	175	0.6	0	0.2	13	0
i2	4679	12619	0.5	166	181	0.6	0	0.2	11	0
i3	4599	12674	0.6	146	179	0.6	0	0.1	11	0
s1	4523	12501	0.6	105	178	0.7	0	0	13	0
s2	4666	12621	0.6	111	179	0.6	0	0.2	14	0
s3	4617	12550	0.6	119	182	0.7	0	0	13	0
AK										
open	11185	32509	0.6	276	468	0.1	0.5	0.1	41	5
closed	10974	32959	0.6	182	457	0.1	0.4	0.2	59	3
semi	11612	32411	0.5	361	498	0.1	0.4	0.2	53	8
semi2	11834	33193	0.5	199	500	0.1	0.5	0.1	40	5
LP										
open	11646	34965	0.6	378	484	0.3	0.2	0.1	38	2
closed	11066	35194	0.6	192	472	0.4	0.2	0.1	43	2
semi	12116	36250	0.6	412	497	0.4	0.2	0.2	46	5
semi2	12108	36409	0.6	305	508	0.3	0.2	0.2	52	8

5 Explicit solvent results

Hydration free energy changes and their components, obtained with TIP3P and SPC/E water models (Table 2 in the main text) are presented in In Fig.3 for visual analysis. Note likely systematic discrepancies in the case of nonpolar component.

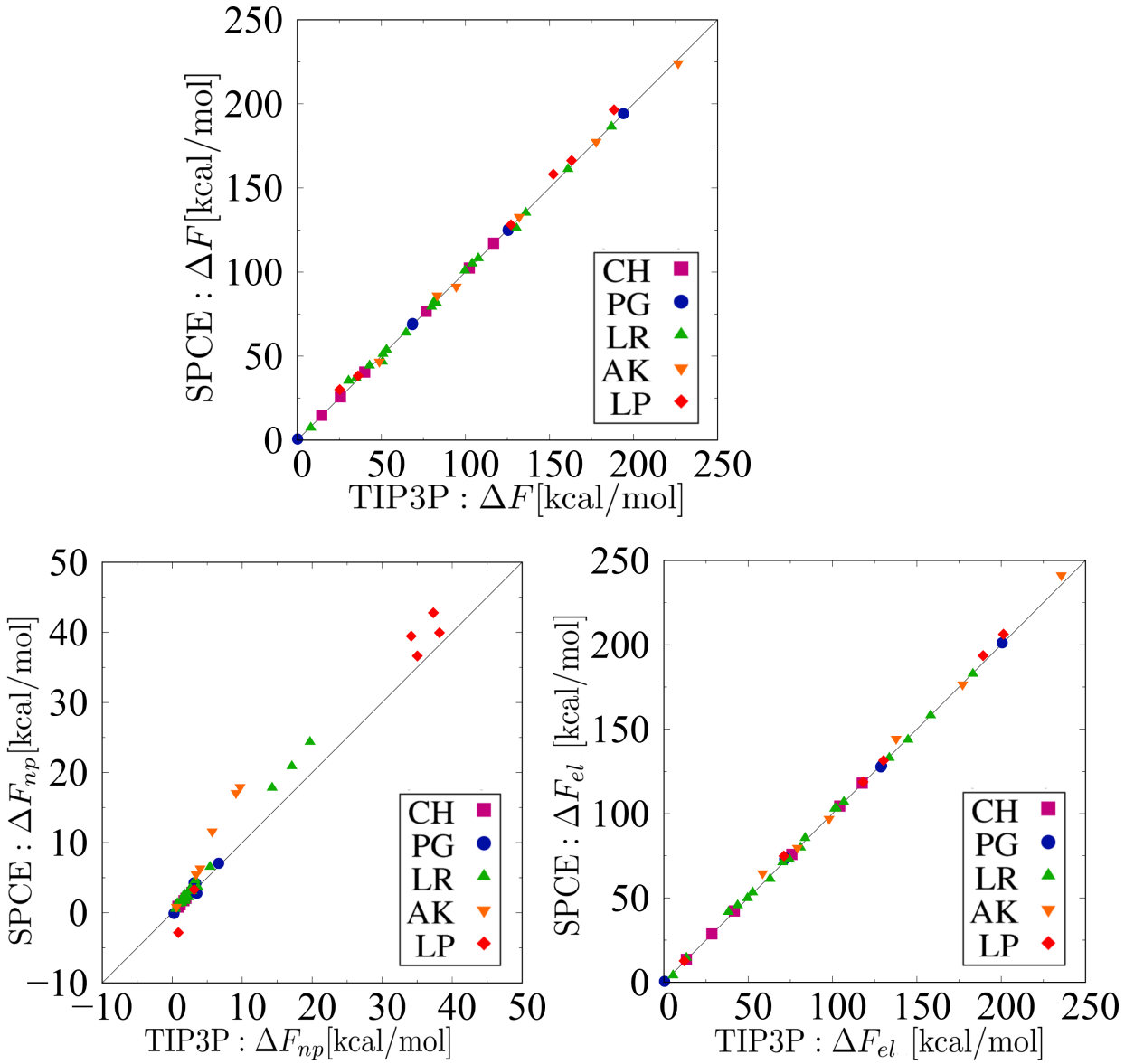

Figure 3: Explicit solvent results for hydration free energy changes, and their nonpolar and electrostatic components, ΔF, ΔF_{np}, ΔF_{el}, respectively, as listed in Table 2, in the main text.

6 Implicit solvent results

Table 3: Numerical data for hydration free energies in explicit (TIP3P) and implicit (Poisson + surface area) solvent, with optimal parameters as given in the main text. All hydration free energy values in kcal/mol. RMSD in Å.

		total			electrostatic			nonpolar		
$A \to B$	RMSD	*TIP3P*	Bondi	Amber	*TIP3P*	Bondi	Amber	*TIP3P*	Bondi	Amber
CH										
f, u	7.4	−116.8	−113.5	−108.5	−117.6	−117.1	−112.3	0.8	3.6	3.8
f, m	3.8	−14.5	−11.8	−13.0	−13.3	−11.6	−12.8	−1.1	−0.2	−0.2
f, p	2.6	−40.1	−39.9	−39.1	−41.8	−40.8	−40.0	1.7	0.8	1.0
u, m	6.8	102.3	101.7	95.5	104.3	105.5	99.5	−1.9	−3.8	−4.0
u, p	6.4	76.7	73.5	69.5	75.8	76.3	72.3	0.8	−2.7	−2.8
m, p	3.3	−25.6	−28.2	−26.1	−28.4	−29.2	−27.2	2.8	1.1	1.2
PG										
b, h	9.9	−125.4	−129.6	−119.9	−128.6	−131.9	−122.3	3.3	2.4	2.4
b, m	6.5	−125.5	−114.6	−109.8	−129.1	−116.1	−111.9	3.5	1.4	2.1
b, u	10.8	−194.0	−197.3	−186.4	−200.6	−205.9	−196.2	6.6	8.7	9.8
h, m	8.0	−0.1	14.9	10.1	−0.4	15.9	10.4	0.3	−0.9	−0.3
h, u	8.0	−68.6	−67.7	−66.5	−72.0	−74.0	−73.9	3.4	6.3	7.4
m, u	8.4	−68.5	−82.6	−76.6	−71.6	−89.9	−84.3	3.1	7.3	7.7
LR_U										
n, u1	6.7	161.0	149.4	149.8	158.2	150.7	152.0	2.8	−1.4	−2.2
n, u2	4.6	81.2	73.2	70.8	83.8	76.5	74.6	−2.6	−3.3	−3.8
n, u3	10.3	30.5	5.9	14.9	13.4	−7.3	−0.8	17.1	13.1	15.7
u1, u2	5.1	−79.8	−76.2	−78.9	−74.4	−74.3	−77.3	−5.4	−2.0	−1.6
u1, u3	11.6	−130.5	−143.5	−134.9	−144.8	−158.0	−152.8	14.3	14.5	17.9
u2, u3	10.3	−50.7	−67.3	−55.9	−70.3	−83.7	−75.5	19.7	16.4	19.6
LR_I										
n, i1	3.7	104.0	109.6	105.8	102.2	109.0	104.8	1.9	0.6	1.1
n, i2	3.9	−82.8	−74.1	−80.1	−81.0	−74.5	−80.2	−1.8	0.3	0.1
n, i3	1.9	53.1	52.9	49.3	52.6	53.6	50.1	0.5	−0.8	−0.7
i1, i2	3.7	−186.9	−183.7	−185.9	−183.2	−183.4	−184.9	−3.7	−0.3	−0.9
i1, i3	3.6	−51.0	−56.7	−56.5	−49.6	−55.3	−54.7	−1.4	−1.4	−1.8
i2, i3	4.0	135.9	127.0	129.4	133.6	128.1	130.3	2.3	−1.1	−0.8
LR_S										
n, s1	2.0	435.0	37.9	42.5	38.3	41.9	46.4	−3.3	−4.0	−3.9
n, s2	2.0	43.0	55.5	47.3	43.7	57.4	49.1	−0.7	−1.8	−1.8
n, s3	2.1	−64.8	−54.9	−54.3	−63.0	−52.7	−52.2	−1.7	−2.3	−2.1
s1, s2	1.8	8.0	17.6	4.8	5.4	15.5	2.7	2.6	2.1	2.1
s1, s3	1.0	−99.7	−92.9	−96.8	−101.3	−94.6	−98.6	1.6	1.7	1.8
s2, s3	1.7	−107.7	−110.5	−101.7	−106.7	−110.0	−101.3	−1.0	−0.5	$4 - 0.4$
AK										
c, s	6.3	−226.6	−201.5	−198.7	−235.7	−215.2	−209.6	9.1	13.7	10.9
c, o	7.2	−48.8	−36.7	−39.3	−58.5	−41.3	−41.6	9.7	4.6	2.3
c, s2	3.9	−132.0	−124.3	−130.2	−137.7	−139.2	−142.9	5.7	14.9	12.7
s, o	3.7	177.7	164.8	159.4	177.2	173.9	168.0	0.6	−9.1	−8.6
s, s2	3.8	94.5	77.2	68.4	97.9	76.0	66.7	−3.4	1.2	1.8
o, s2	5.2	−83.200	−87.7	−91.0	−79.2	−97.9	−101.3	−4.0	10.3	10.4
LP										
c, o	4.9	188.4	175.8	187.7	189.3	171.2	182.5	−0.9	4.6	5.2
c, s	3.2	152.3	116.9	119.0	118.1	102.5	103.4	34.2	14.4	15.5
c, s2	4.0	25.2	6.6	25.3	−12.1	−8.2	7.7	37.3	14.8	17.6
o, s	3.4	−36.1	−58.9	−68.7	−71.2	−68.7	−79.0	35.0	9.8	10.3
o, s2	3.2	−163.2	−169.2	−162.4	−201.4	−179.4	−174.8	38.2	10.2	12.4
s, s2	1.7	−127.1	−110.3	−93.7	−130.2	−110.8	−95.7	3.1	0.5	2.0

7 Electrostatic component for Amber-based radii

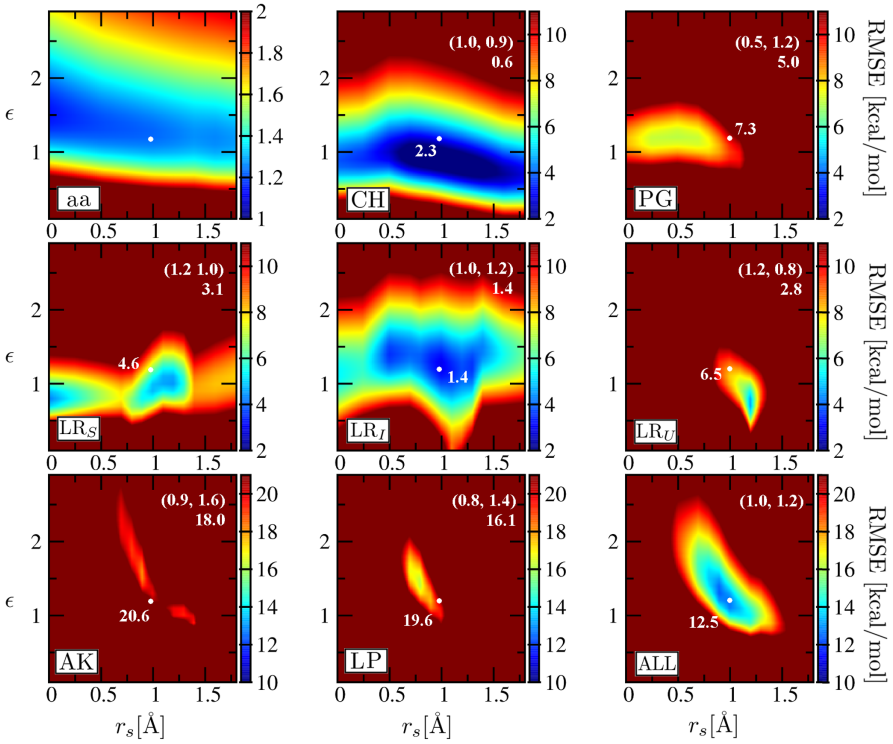

Figure 4: RMSEs for Poisson-based electrostatic hydration component with respect to TIP3P ΔF_{el}, for neutral amino acids (aa), individual protein structures, and all protein structures (ALL). Numbers in upper-right corners denote individually optimal (r_s, ϵ) pairs, and corresponding RMSEs in kcal/mol. White dots and neighbouring numbers denote the position of global RMSE minimum: $(r_s, \epsilon) = (1.0, 1.2)$, and respective RMSEs for each protein structure.

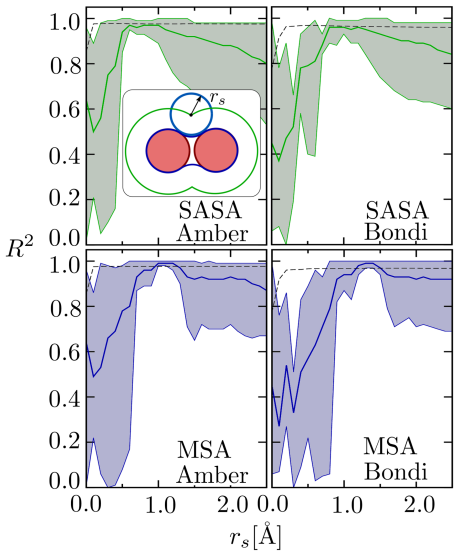

Figure 5: Square of Pearson's correlation coefficients (R^2) for correlation between explicit solvent based solute-solvent boundary areas and MSA/SASA obtained for different solvent probe radii (r_s) and protein atomic radii. Thick lines correspond to R^2 averaged over all protein sets, shaded area extend from minimal to maximal correlation obtained for a given r_s. Dashed lines denote R^2 for amino acid structures. Inset: scheme illustrating the construction of SAS (green), MS (blue), and van der Waals (red) surfaces.

8 Solute-solvent boundary

Most implicit solvent models critically depend on the definition of solute-solvent boundary. For the electrostatic term it specifies the border between low (solute) and high (solvent) dielectric regions. For for the nonpolar term the boundary area determines the amount of work needed to create a cavity within the solvent necessary to accommodate the solute. To date, most popular definitions (Fig. 5, inset) used in the context of electrostatics are based on molecular surface (MS),[7] while the nonpolar contributions are defined based on solvent accessible surface (SAS).[8] The two definitions converge to van der Waals surface, if the radius of a solvent probe, r_s, used for surface construction is 0.

The problem of selecting the most appropriate r_s value is still under debate.[9] A reasonable assumption would be that the constructed surface should properly reflect explicit solvent distribution around the solute. A crude measure of such correspondence may be the scaling of surface areas for different solute geometries. Having constructed solvent surfaces for all considered protein conformations based on explicit solvent boundaries derived from MD simulations (see Methods), we investigated the correlation of their areas with MS and SAS areas (MSA, and SASA, respectively) obtained for the same protein conformations.

As can be seen in Fig. 5, van der Waals surface (the limit of $r_s \to 0$) results in rather poor correlation. It is due to the fact that it involves contributions from tiny interatomic spaces that are not accessible to solvent, and persist regardless of the actual protein conformation. An average correlation rapidly increases with growing r_s for both MS and SAS, reaches maximum, and then slowly decreases, as the solvent probe becomes larger than water molecules ($r_s \sim 1.4$ Å) and can't penetrate solvent accessible crevasses at protein surface. The optimal r_s values, resulting in universally good correlations for all considered protein structures, are consistently larger for MS than for SAS, and for Bondi compared to Amber-based radii (Bondi radii are generally slightly smaller, except for hydrogen atom types).

In contrast to dependencies observed for protein structures, the sensitivity of surface area correlations to r_s value is practically nonexistent for simple amino acid compounds (dashed lines in Fig. 5). It indicates that conclusions regarding the performance of implicit solvent models based on particular surface definition for small molecules are not easily transferable to larger, complex structures.

9 Solvent excluded volume (SEV) measurements

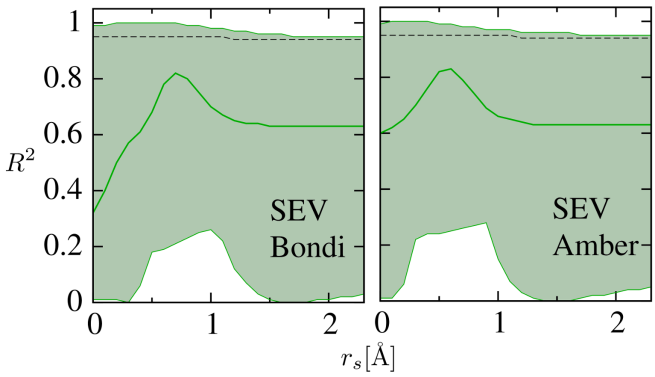

Figure 6: Square of Pearson's correlation coefficients (R^2) for correlation between explicit solvent based solute volumes and SEV based on solvent accessible surface obtained for different solvent probe radii (r_s) and protein atomic radii. Thick lines correspond to R^2 averaged over all protein sets, shaded areas extend from minimal to maximal correlation obtained for a given r_s. Dashed lines denote R^2 for amino acid structures.

10 Benchmark files

For each protein conformation considered for the calculations of hydration free energy changes we provide the following files:

- `data.csv`
  a text file with numerical data from Table 2 of the manuscript,
- `*gro`
  structure file (protein + solvent) in GROMACS format,
- `topology.top`
  GROMACS topology file (protein + solvent); the same file is valid for all conformations of a given protein,
- `*pqr`
  protein geometry with partial charges and atomic radii based on Amber force field, in the format used by APBS,
- `*_b.pqr`
  as above, but with atomic radii taken from the Bondi set,
- `*dx`
  volumetric data for solvent density distribution around protein structures (number density of water oxygen atoms) obtained based on simulations with TIP3P water model.

All files are available for download **at the following addresses**:

- `https://github.com/SetnyLab/Protein-Benchmark`
- `http://www3.cent.uw.edu.pl/~piosto`

More data is available upon request (*email:* p.setny@cent.uw.edu.pl).

References

(1) Rocklin, G. J.; Mobley, D. L.; Dill, K. A.; Hünenberger, P. H. *J. Chem. Phys.* **2013**, *139*, 184103.

(2) Jorgensen, W. L. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.

(3) Berman, H. M. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(4) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(5) Abraham, M. J.; Murtola, T.; Schulz, R.; Páall, S.; Smith, J. C.; Hess, B.; Lindah, E. *SoftwareX* **2015**, *1-2*, 19–25.

(6) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

(7) Connolly, M. L. *Science* **1983**, *221*, 709–713.

(8) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379–400.

(9) Pang, X.; Zhou, H.-X. *Commun. Comput. Phys.* **2013**, *13*, 1–12.