Insights into Molecular Recognition: Multi-Resolution Simulations of Proteins and their Hydration Shells to Capture both Global Fluctuations and Chemical Detail

A. C. Fogarty, R. Fiorentini, K. Kremer, R. Potestio

published in

NIC Symposium 2018
K. Binder, M. Müller, A. Trautmann (Editors)

Forschungszentrum Jülich GmbH,
John von Neumann Institute for Computing (NIC),
Schriften des Forschungszentrums Jülich, NIC Series, Vol. 49,
ISBN 978-3-95806-285-6, pp. 127.
http://hdl.handle.net/2128/17544

Insights into Molecular Recognition: Multi-Resolution Simulations of Proteins and their Hydration Shells to Capture both Global Fluctuations and Chemical Detail

Aoife C. Fogarty, Raffaele Fiorentini, Kurt Kremer, and Raffaello Potestio

Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany
E-mail: {fogarty, fiorentini, kremer, potestio}@mpip-mainz.mpg.de

Multi-resolution representations of soft matter systems are modelling strategies that allow the concurrent usage of different levels of model accuracy in the same simulation setup. These approaches unite the advantages of computationally less expensive representations and quantitatively accurate all-atom models. Here we report on the usage of such methods for the calculation of solvation and binding free energies for small biologically significant solutes and the enzyme lysozyme. The method is shown to correctly reproduce the same quantities as computed in reference, all-atom simulations, thus paving the way not only for computationally efficient free energy calculations, rather also for the investigation of the physico-chemical properties of relevant biomolecules.

1 Introduction

Computer simulations of biomolecules in solution must cover a broad range of length and timescales, ranging from femtosecond bond vibrations to protein folding or diffusion on the order of seconds, and from Angstrom-long bonds to enzymes with a scale of nanometres and beyond. Simulating such systems requires the use of models which are sufficiently detailed to capture the biophysical or biochemical processes of interest; however, it also implies to deal with large systems and long simulation times. This makes biomolecular systems a most appropriate field of application for concurrent multiple resolution methodologies, in which computationally expensive high resolution and cheaper low resolution models are simultaneously employed in one simulation setup[1]. Typically, one part of the system is modelled using more expensive molecular mechanical forcefields with a level of resolution in which each atom is represented by an individual particle. The remainder of the system is then modelled using a coarse-graining approach, in which multiple atoms are grouped together in one particle for computational efficiency.

One specific application expected to benefit from such a multi-resolution approach is the study of ligand binding, or the recognition of a small molecule by a protein, a process which plays a central role in biology[2]. Signal transduction, drug action mechanism and enzymatic reactions all rely on specific interactions between ligands and proteins. Elucidating the coupling between the global properties of a protein and the process of ligand binding is fundamental for furthering our understanding of protein function, and finds application in numerous contexts, including drug discovery and the use of enzymes in industrial synthesis, to name just a few[3].

For multi-resolution studies of ligand binding to enzymes, reducing to a minimum the part of the system modelled using atomistic level detail means focusing only on the binding site, the ligand and the neighbouring water molecules. In such a set-up, the global properties of the protein are captured on a coarse-grained level. These in turn modulate

the local properties of the binding site, modelled on an atomistic level. The atomistic and coarse-grained solvent are similarly coupled. Until now, this approach remained relatively undeveloped, with some notable exceptions[4,5]. Some of us have recently developed and validated a methodology[6] allowing the simulation of an atomistic ligand binding site solvated in atomistic water and embedded in a coarse-grained protein. It combines the Adaptive Resolution simulation scheme (AdResS) in which solvent particles change their resolution on the fly during the simulation[7,8], coarse-grained Elastic Network Models (ENMs) for proteins[9], and standard atomistic forcefields. This setup was shown to allow the simulation of stable ligand binding, capturing both global structure and fluctuations of the enzyme, and local physics and chemistry of the ligand binding site and its hydration water. Furthermore, it explicitly includes binding site solvation in a computationally efficient way, which is crucial for correctly modelling ligand binding[10]. The inherently multiscale nature of the model allows a unique insight into the coupling between global and local processes, and its implications for protein function.

2 Multi-Resolution Modelling of a Protein

The multi-resolution modelling of a protein involves different methodological and technical aspects, which can be divided into two broad domains: first, the development of the AdResS methodology, and second the development of a dual-resolution model of the molecule itself. The AdResS method[7,8,11] allows the smooth coupling of atomistic and coarse-grained models for diffusive particles within the same simulation box. In AdResS, one defines two regions: the atomistic (AT) region where non-bonded interactions are modelled using the atomistic forcefield, and the coarse-grained (CG) region, where the coarse-grained forcefield is used. Between them is a hybrid (HY) region where particles smoothly change their resolution from atomistic to coarse-grained. For the specific systems here under examination, the AT region is a sphere of radius d_{at} centred on a point $\mathbf{r}_{centre}$ in the ligand-binding site. The HY region is then a spherical shell of width d_{hy}, and the remainder of the system is the CG region. Solvent particles diffuse freely between regions, changing resolution as a function of their instantaneous position. Non-bonded interactions are modelled using a force-interpolation scheme, in which the intermolecular force between the centres of mass of molecules α and β is given by

$$\mathbf{F}_{\alpha\beta} = w(\mathbf{r}_\alpha)w(\mathbf{r}_\beta)\mathbf{F}^{AT}_{\alpha\beta} + [1 - w(\mathbf{r}_\alpha)w(\mathbf{r}_\beta)]\mathbf{F}^{CG}_{\alpha\beta} \,,$$

where

$$\mathbf{F}^{AT}_{\alpha\beta} = \sum_{i\in\alpha}\sum_{j\in\beta}\mathbf{F}^{AA}_{ij} \,,$$

$\mathbf{F}^{AA}_{ij}$ is the atomistic non-bonded interaction between atoms i and j, $\mathbf{F}^{CG}_{\alpha\beta}$ is the coarse-grained non-bonded interaction between molecules α and β, and w is a weighting function that varies smoothly and monotonically across the HY region, from a value of 1 in the AA region to 0 in the CG region.

This methodology allows a significant reduction in the number of degrees of freedom simulated at the high-resolution level, while still reproducing the properties of the binding site and hydration water as they would be in a fully atomistic simulation. In a previous

study of a fully atomistic protein, ubiquitin, in adaptive resolution water, some of us validated the extension of AdResS to biomolecular systems via a detailed exploration of the structural and dynamical properties of the protein and water[13]. This included the conformational fluctuations of the protein as measured by the root mean square fluctuations and NMR S^2 order parameters of the backbone, and the water density, local structure and reorientational dynamics, in the hydration shell and in the bulk atomistic region.

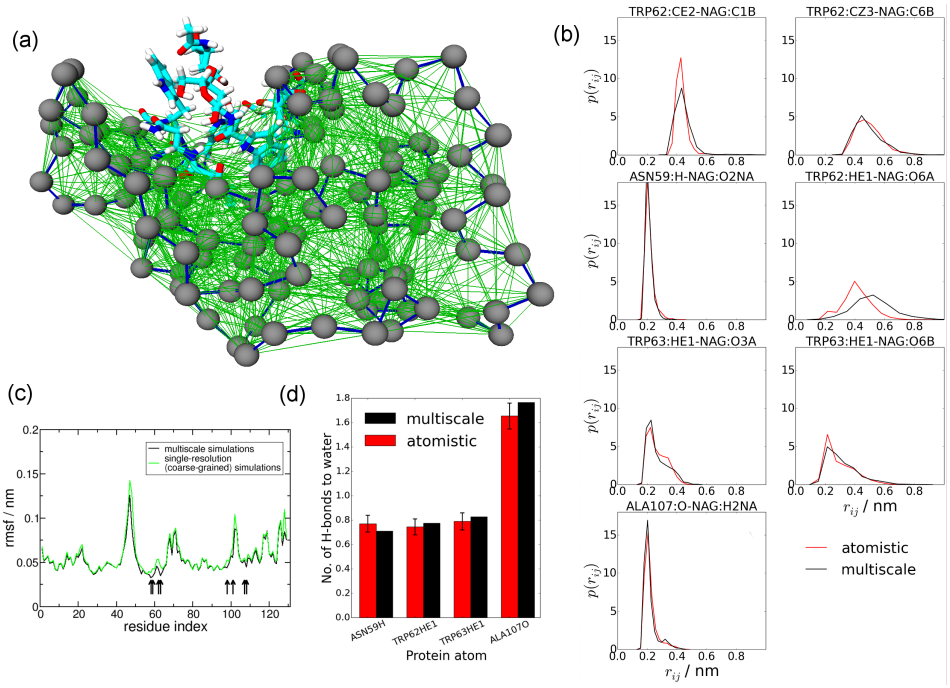

Figure 1. (a) Multi-resolution protein model showing the atomistic-resolution binding site in red, blue, cyan and white (O, N, C and H atoms). The coarse-grained protein is modelled by an Elastic Network model, shown as grey beads connected by harmonic springs (dark blue and green lines). The system also contains atomistic and coarse-grained water, not shown. (b) Stable ligand binding in the fully atomistic and multi-resolution models, as measured by the probability density functions for key contact distances between ligand and protein. (c) Global protein fluctuations in the purely coarse-grained and multi-resolution models, as measured by root mean square fluctuations. (d) Hydration of the binding site in the fully atomistic and multi-resolution models, as measured by the number of protein-water hydrogen bonds at key sites. Adapted from Ref. 6.

Subsequently, we constructed a multi-resolution protein model as illustrated in Fig. 1. In this methodology, atomistic detail is inserted into a coarse-grained protein model to capture the chemical detail of the ligand binding site. The coarse-grained protein, in fact, supplies the structure and fluctuations of the remainder of the macromolecule that in turn modulate the properties of the binding site. For the low-resolution part of the molecule we made use of an Elastic Network Model (ENM), a widely used coarse-grained protein model[14,15] in which coarse-grained beads represent the protein C_α and all pairs of beads within a cutoff are connected by harmonic springs. The elastic parameters and reference

distances in the ENM are parametrised so as to reproduce the fluctuations determined experimentally or through atomistic simulations. The binding site is solvated in a sphere of atomistic resolution water molecules, coupled to a reservoir of coarse-grained water molecules via the AdResS methodology.

This methodology was the first to couple atomistic and coarse-grained detail in both the protein and aqueous solution, and was shown to correctly simulate their structure and conformational fluctuations. As well as reducing the number of interactions to be calculated, it also facilitates the simulation of systems containing components for which atomic-resolution structures are not available, but which can nevertheless be represented and parametrised on a coarse-grained level.

The models and approaches discussed have been employed to compute solvation and binding free energies of biologically significant molecules, as it is detailed hereafter.

3 Calculation of Solvation and Binding Free Energies

Proteins are essential components of living organisms, playing a critical role in innumerable biophysical and biochemical processes. The vast majority of drug targets are proteins[16], and proteins also have a wide range of industrial applications. However, many details of their function remain mysterious. Controversies abound, for example regarding the process by which enzymes catalyse reactions[17], or the mechanisms of ligand binding[18]. Many of these open questions are related to the coupling of global protein properties and the local chemical details of protein function. This is where the above described multiresolution protein model can play a role. Our approach, which couples local atomistic and global coarse-grained detail, allows us a unique amount of control in precisely which degrees of freedom are represented at a given level of chemistry or physics in our simulations. By taking advantage of this to systematically vary the coupling between local and global properties of the protein, we expect to be able to shed new light on its function.

A question we addressed before attacking the problem of binding free energy calculation, though, is that of solvation free energies of small molecules solvated in water. The usage of the AdResS scheme in this context is nontrivial, due to the *a priori* incompatibility between the force-based method, which relies on a non-conservative force field, and free energy calculation methods such as thermodynamic integration based on a Hamiltonian formulation. This issue is shown to be soluble, and the results of a solvation free energy computation for two small side chain analogues are reported hereafter.

On the other hand, regarding the process of inhibitor binding, which is of crucial importance in drug design, we focus on the fundamental interplay between the binding process and global system fluctuations. We calculate free energies of ligand binding for an enzyme in aqueous solution with an inhibitor chosen for its experimental relevance, and explore the coupling with global system properties.

3.1 Using Force-Based Adaptive Resolution Simulations to Calculate Solvation Free Energies of Amino Acid Sidechain Analogues

The AdResS method is based on the interpolation of atomistic and coarse-grained forces, and it is known that no global Hamiltonian is defined[8]. The simulations thus obey Newton's third law but are not energy-conserving. On the other hand, calculating free energies

via the Thermodynamic Integration[19] (TI) approach involves taking the derivative of the potential energy with respect to an order parameter λ which defines a path between two thermodynamic states. Recently[20], we showed how AdResS can be combined with the TI approach, relying on the fact that the relevant solute-solvent interaction potential is locally well defined, and that AdResS samples the same configurations in the atomistic region as would be sampled in the corresponding sub-region of a fully atomistic simulation.

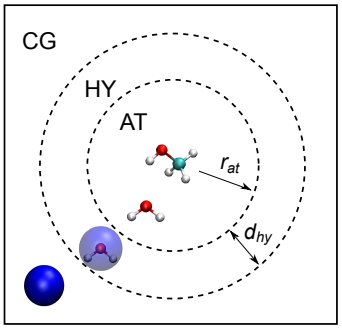

Figure 2. Illustration of the AdResS approach, showing the solute and solvent, and the atomistic, hybrid and coarse-grained regions (not to scale). Adapted from Ref. 20.

We demonstrated this by calculating solvation free energies of amino acid sidechain analogues methanol and 3-methylindole. We compared fully atomistic systems with different simulation box sizes to AdResS systems with different atomistic region sizes and different CG potentials. One of the CG potentials used was obtained via systematic coarse-graining to reproduce the atomistic structure (Iterative Boltzmann Inversion or IBI[21]), while the other CG potential was a structureless fluid of non-interacting particles, an ideal gas[12].

The free energies (Coulomb and Lennard-Jones contributions) are shown in Fig. 3(a) and (b) for the example of methanol, one of the systems we studied. The AdResS values agree with those calculated from fully atomistic simulations to within a fraction of a kJ mol^{-1}. This is the case even for small atomistic regions whose size is on the order of the correlation length, or when the properties of the coarse-grained region are extremely different from those of the atomistic region. These accurate free energy calculations are possible because AdResS allows the sampling of atomistic configurations which are equivalent to those of fully atomistic simulations, as we confirmed by comparing the structure and molecular fluctuations of the AdResS and fully atomistic systems (a selection of examples are shown in Fig. 3(c) and (d)). This thorough exploration of the AdResS+TI combination justifies the application of this methodology in more complex biomolecular systems.

3.2 Insights into Molecular Recognition: Using Multi-Resolution Simulations to Calculate Free Energies of Protein-Ligand Binding

Subsequently, we employed the multi-resolution protein model solvated in fully atomistic water for best control over the density of water in the ligand binding site, in order to cal-

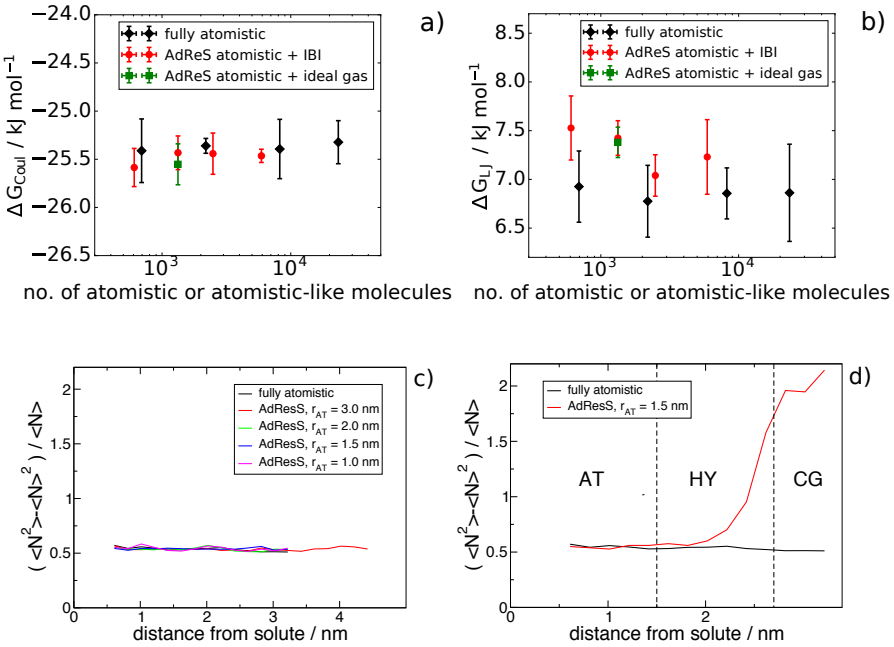

Figure 3. (a) Coulomb and (b) LJ contributions to the free energies of solvation for methanol, fully atomistic versus AdResS with an IBI CG potential and with an ideal gas CG region. Both subplots are plotted such that the y-axis covers a range of 2.5 kJ mol^{-1}. (c) and (d) Molecular fluctuations in spherical concentric bins of equal surface-to-volume ratio, as a function of distance from the solute atom defining the centre of the atomistic region, (c) AdResS with IBI coarse-grained potential, (d) AdResS with ideal gas coarse-grained region. Adapted from Ref. 20.

culate protein-ligand binding free energies using the multi-resolution set-up. This model gives us unique control over the representation of system components at an atomistic or coarse-grained level, allowing us to explore the interplay between the binding process, local chemical detail, and global system structure and fluctuations. We varied the number of protein degrees of freedom included atomistically in the multi-resolution model, comparing the resulting free energy values to a fully atomistic reference.

Fig. 4 shows the values of $\Delta G_{Coul,c}$, $\Delta G_{LJ,c}$ and $\Delta G_{restr,c}$ as a function of the percentage of protein degrees of freedom which are included atomistically in the dual-resolution protein model. These data suggest that including 8 atomistic residues (8.5%) is already enough to simulate stable ligand binding and reproduce the structural properties of the binding site. The individual components of the ligand binding free energy are an even more sensitive measure of the ability of a given model to fully reproduce the ligand binding process. In Fig. 4, we see that the free energies oscillate about the fully atomistic reference (horizontal dashed line) as the number of atomistic residues increases. With only 10.6% atomistic detail (10 atomistic residues) included in the protein, our multi-resolution model yields the same free energy values as a fully atomistic simulation.

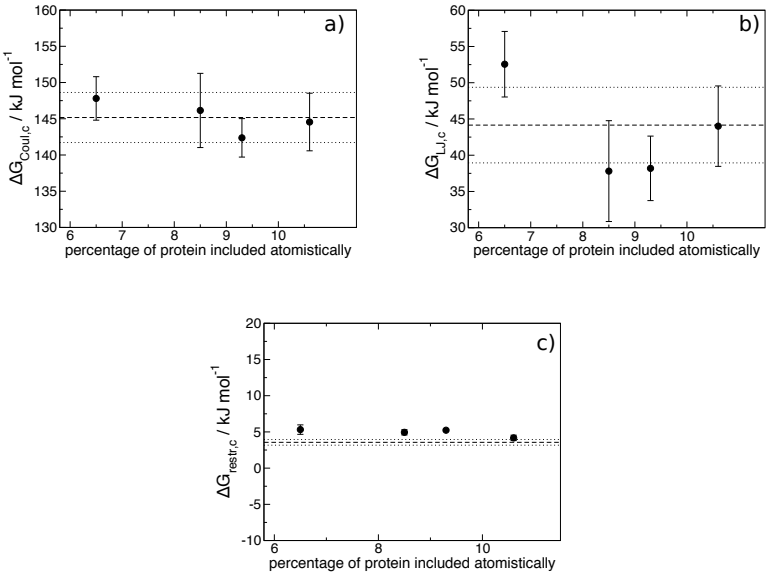

Figure 4. (a) Coulomb, (b) Lennard-Jones and (c) restraint free energies in the protein-ligand complex, as a function of the percentage of protein included in atomistic detail in the multi-resolution set-up. The points correspond to 6, 8, 9 and 10 atomistic residues respectively. The heavy dashed black horizontal lines are the reference values from fully atomistic simulations, and the lighter dotted black horizontal lines are the error bars for those values. All y-axes cover the same energy range.

4 Concluding Remarks

This multi-resolution models here discussed can be employed to explore how changes in the coarse-grained part of the protein influence the free energy values. The simplicity of the Elastic Network-based coarse-grained model allows one to easily control different aspects of its properties, for example retaining the same equilibrium structure but distorting the global fluctuations. This would enable one to gain insight into the coupling between global and local processes, and its implications for protein function.

Acknowledgements

K. K. and A. C. F. acknowledge research funding through the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant Agreement No. 340906-MOLPROCOMP. We thank the John von Neumann Institute for Computing at the Jülich Supercomputing Centre for computer time on JURECA - project ID hmz35.

References

1. K. Meier, A. Choutko, J. Dolenc, A. P. Eichenberger, S. Riniker, and W. F. van Gunsteren, *Multi-Resolution Simulation of Biomolecular Systems: A Review of Methodological Issues*, Angew. Chem. Int. Ed., **52**, no. 10, 2820–2834, 2013.
2. M. K. Gilson and H.-X. Zhou, *Calculation of Protein-Ligand Binding Affinities*, Annu. Rev. Biophys. Biomol. Struct., **36**, no. 1, 21–42, 2007.
3. E. Persch, O. Dumele, and F. Diederich, *Molecular Recognition in Chemical and Biological Systems*, Angew. Chem. Int. Ed., **54**, no. 11, 3290–3327, 2015.
4. M. Neri, C. Anselmi, M. Cascella, A. Maritan, and P. Carloni, *Coarse-Grained Model of Proteins Incorporating Atomistic Detail of the Active Site*, Phys. Rev. Lett., **95**, 218102, Nov. 2005.
5. M. R. Machado, P. D. Dans, and S. Pantano, *A hybrid all-atom/coarse grain model for multiscale simulations of DNA*, Phys. Chem. Chem. Phys., **13**, 18134–18144, 2011.
6. A. C. Fogarty, R. Potestio, and K. Kremer, *A multi-resolution model to capture both global fluctuations of an enzyme and molecular recognition in the ligand binding site*, Proteins: Structure, Function, and Bioinformatics, 2016, doi:10.1002/prot.25173.
7. M. Praprotnik, L. Delle Site, and K. Kremer, *Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly*, J. Chem. Phys., **123**, no. 22, 224106, 2005.
8. R. Potestio, C. Peter, and K. Kremer, *Computer Simulations of Soft Matter: Linking the Scales*, Entropy, **16**, no. 8, 4199–4245, 2014.
9. M. M. Tirion, *Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis*, Phys. Rev. Lett., **77**, 1905–1908, Aug. 1996.
10. N. Hansen and W. F. van Gunsteren, *Practical Aspects of Free-Energy Calculations: A Review*, J. Chem. Theory Comput., **10**, no. 7, 2632–2647, 2014.
11. R. Potestio, S. Fritsch, P. Español, R. Delgado-Buscalioni, K. Kremer, R. Everaers, and D. Donadio, *Hamiltonian Adaptive Resolution Simulation for Molecular Liquids*, Phys. Rev. Lett., **110**, 108301, Mar 2013.
12. K. Kreis, A. C. Fogarty, K. Kremer, and R. Potestio, *Advantages and challenges in coupling an ideal gas to atomistic models in adaptive resolution simulations*, Eur. Phys. J. Spec. Top., **224**, 2289–2304, 2015.
13. A. C. Fogarty, R. Potestio, and K. Kremer, *Adaptive resolution simulation of a biomolecule and its hydration shell: Structural and dynamical properties*, J. Chem. Phys., **142**, no. 19, 195101, 2015.
14. W. Zheng and P. Glenn, *Probing the folded state and mechanical unfolding pathways of T4 lysozyme using all-atom and coarse-grained molecular simulation*, J. Chem. Phys., **142**, no. 3, 035101, 2015.
15. L. Orellana, M. Rueda, C. Ferrer-Costa, J. Lopez-Blanco, P. Chacón, and M. Orozco, *Approaching Elastic Network Models to Molecular Dynamics Flexibility*, J. Chem. Theory Comput., **6**, no. 9, 2910–2923, 2010.
16. J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, *How many drug targets are there?*, Nat. Rev. Drug Discov., **5**, 993–996, 2006.
17. S. C. L. Kamerlin and A. Warshel, *At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?*, Proteins: Struct., Funct., Bioinf., **78**, no. 6, 1339–1375, 2010.

18. K. Okazaki and S. Takada, *Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms*, Proc. Natl. Acad. Sci. USA, **105**, no. 32, 11182–11187, 2008.
19. J. G. Kirkwood, *Statistical Mechanics of Fluid Mixtures*, J. Chem. Phys., **3**, no. 5, 300–313, 1935.
20. R. Fiorentini, K. Kremer, R. Potestio, and A. C. Fogarty, *Using force-based adaptive resolution simulations to calculate solvation free energies of amino acid sidechain analogues*, J Chem. Phys., **146**, no. 24, 244113, 2017.
21. D. Reith, M. Pütz, and F. Müller-Plathe, *Deriving effective mesoscale potentials from atomistic simulations*, J Comput. Chem., **24**, no. 13, 1624–1636, 2003.