# Stony Brook University

# Improved Generalized Born Solvent Model Parameters for Protein and Nucleic Acid Simulations

A Dissertation Presented

by

**Hai Minh Nguyen**

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

**Doctor of Philosophy**

in

**Chemistry**

Stony Brook University

**August 2014**

**Stony Brook University**

The Graduate School

**Hai Minh Nguyen**

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

**Carlos Simmerling – Dissertation Advisor**
**Professor, Department of Chemistry**

**Robert C. Rizzo - Chairperson of Defense**
**Professor, Department of Applied Mathematics and Statistics**

**Peter J. Tonge – Third Member**
**Professor, Department of Chemistry**

**Markus Seeliger – Outside Member**
**Assistant Professor, Department of Pharmacological Sciences**

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

**Improved Generalized Born Solvent Model Parameters**

**for Protein and Nucleic Acid Simulations**

by

**Hai Minh Nguyen**

**Doctor of Philosophy**

in

**Chemistry**

Stony Brook University

**2014**

Generalized Born solvent model offers inexpensive approach for solvation energy calculation as compared to explicit solvent and Poisson-Boltzmann (PB) method. Thanks to its speed, GB models have been widely used in molecular dynamics (MD) simulations. However the speed comes with tradeoffs. Literatures have pointed out the weaknesses of GB models in inaccurately calculating solvation energy that leads to helical bias in protein simulation or unstable DNA/RNA duplex in nucleic acid simulation. Here we introduced the reparameterization of the recently developed GB-Neck model to improve its accuracy. Compared to other pairwise GB models (e.g. GB-OBC and GB-Neck) the new GB models have better agreement to very accurate (but slow) PB method in terms of reproducing solvation energies for a variety of systems from protein to nucleic acid. For the protein, secondary structure preferences are in much better agreement with explicit solvent simulations. We also obtain near-quantitative reproduction of experimental structure and thermal stability profiles for several model peptides. Moreover the model is able to reproduce the folding of microsecond to millisecond time scale folding of a series of larger proteins. For the nucleic acid, simulations maintain stable trajectories for various DNA and RNA duplexes through MD simulations and also correctly fold DNA/RNA hairpin from extended conformation.

**Table of Contents**

**List of Figures**

**List of Tables**

**List of Abbreviations**

| | |
|---|---|
| abs_e | Absolute Solvation Energy Root-Mean-Square-Deviation |
| AR6 | AnaticaR6 Generalized Born solvent model |
| BB-RMSD | Backbone Root-Mean-Square Deviation |
| CASP | Critical Assessment of Techniques for Protein Structure Prediction |
| CFA | Coulomb Field Approximation |
| DAP | 6-amidine-2-(4-amidino-phenyl) indole |
| DD | Dickerson-Drew dodecamer |
| eff_rad_rmsd | Effective radii root-mean-square-deviation |
| FF | Force Field |
| GAFF | General Amber Force Field |
| GA | Genetic Algorithm |
| GB | Generalized Born |
| GBMV | Generalized Born using Molecular Volume |
| GB-Neck | Generalized Born with Neck correction |
| GPU | Graphics Processing Unit |
| MD | Molecular Dynamics |
| MS | Molecular Surface |
| PB | Poisson-Boltzmann |
| PME | Particle mesh Ewald |
| PMF | Potential Mean Force |
| QM | Quantum Mechanics |
| rel_e | Relative Solvation Energy Root-Mean-Square-Deviation |
| REMD | Replica Exchange Molecular Dynamics |
| RMSD | Root-Mean-Square Deviation |

| | |
|---|---|
| SASA | Solvent-Accessible Surface Area |
| TI | Thermodynamic Integration |
| VDW | Van der Waals |

**Acknowledgments**

Doing PhD is a long and challenging process and I would not fulfill it without support from my family. They always stand by my side; encourage me in every step of my life. I would like to thank them for unconditional love, for teaching and guiding me since the day I was born so I can become the grown man I am now.

I also would like to send many thanks to my advisor, Dr. Carlos Simmerling, for his patience in guiding me in over those years. Since the day I have met him, I love his enthusiasm, his kindness, his knowledge and his professional. Thanks for your belief that we could make better GB model, and yeah, we made it.

I want to thank my committee members: Dr. Robert C. Rizzo, Dr. Peter J. Tonge and Dr. Markus Seeliger. I really appreciate their time, their knowledge and their critical comment and suggestion.

I also want to thank all members in Dr. Simmerling's lab their help during my stay in the lab. Thanks Maier for his knowledge about force field development, for his eagerness and willingness helping my research as well. Thanks for the "overnight-gmail-chatting" about solvent model, about protein folding, about force field and about all other interesting things we came up with. Thanks Carmenza, He Huang, and Yi Shang for always being my labmates and my close friends. Thanks for always being help whenever I need. My life is much easier when having you.

Many thanks to Dr. Daniel Roe, Dr. Christina Bergonzo, Dr. Lauren Wickstrom, Dr. Amber Carr, Dr. Alberto Perez for their help.

Wish you all the best in the future.

**Publications**

1. **Nguyen, H**.; Roe, D. R.; Simmerling, C., Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J. Chem. Theory Comput.* **2013,** *9* (4), 2020-2034.

2. **Nguyen, H.**;* Maier, J.;* Huang, H; Simmerling, C., Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *(submitted)*. (*co-first authors)

3. **Nguyen, H**.; Perez, A.; Simmerling, C., Refinement of Generalized-Born Neck Parameters for Nucleic Acid and Their Complex with Protein. *(in preparation).*

4. Shang, Y.; **Nguyen, H.**; Wickstrom, L.; Okur, A.; Simmerling, C., Improving the description of salt bridge strength and geometry in a Generalized Born model. *J. Mol. Graphics Model.* **2011,** *29* (5), 676-684.

# Chapter 1. Introduction

## 1.1 Implicit solvent

To successfully mimic experiment in Molecular Dynamics (MD) simulations, one must accurately include the solvent effect. While an explicit solvent, which represents discreet solvent molecules, arguably is considered an adequately accurate model, this approach is accompanied by expensive computational costs due to the significantly high degree of freedom. In a case where more conformational sampling is needed, implicit solvents would provide an excellent alternation.

Implicit solvents (especially Generalized Born solvent models) offer several advantages when directly calculating solvation energy, such as: (1) a fast rate of speed in calculating solvation energy for small to medium-sized systems; (2) a low viscosity which makes sampling more efficient;[1] (3) a friendliness to enhanced sampling methods such as the replica exchange molecular dynamics (REMD) method;[2] (4) an excellent increase in speed when implemented in GPU-based Molecular Dynamics code;[3] and (5) a high level of efficiency when implemented in parallel MD.[4] Implicit solvent models are used extensively in various applications, such as when studying protein folding[1a, 5]and the large-scale motions of protein,[6] designing protein,[7] and developing new force fields.[8]

In an implicit solvent model, the solvation energy is directly calculated. This energy is required to insert a solute from a vacuum into a solvent environment. Solvation energy is commonly decomposed by the sum of polar and nonpolar energy: $\Delta G_{solvation} = \Delta G_{polar} + \Delta G_{np}$. Nonpolar energy is the energy required to insert a non-charge system from a vacuum to a solvent, while polar energy is the energy required to turn on the charge in the solvent (figure 1.1).

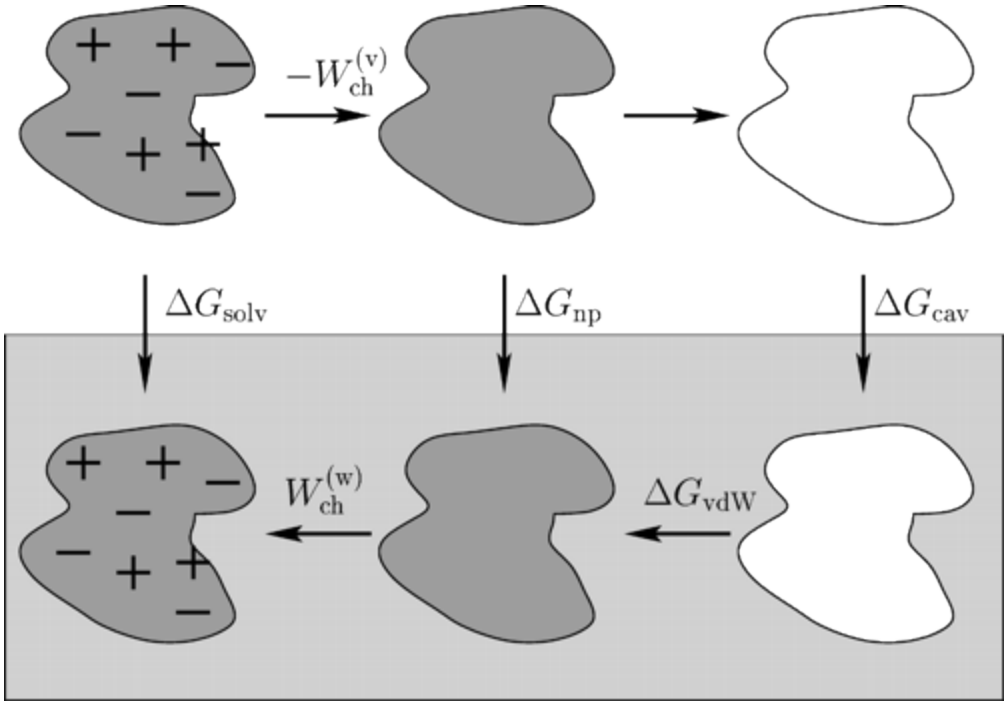


**Figure 1.1** Decomposition of solvation energy to polar and non-polar terms. Non-polar energy is further decomposed to VDW and cavity terms. Figure was reproduced from Levy et al.[9]

### 1.1.1 Nonpolar model

Nonpolar energy is commonly approximated by $\Delta G_{np} = \sigma * A$, where $\sigma$ is the surface tension coefficient and the A term is the solute surface area. The surface tension coefficient is generally given a range of 0.005 to 0.138 kcal/(mol*Å$^2$), based on different definitions of the solute/solvent boundary (VDW, MS, or solvent assessable surface area) or different experimental conditions.[10] Since this coefficient is small, the resulting nonpolar force calculated from the nonpolar energy is usually much smaller than that which is calculated from polar energy.[11] This nonpolar term, as a result, tends to be neglected in MD simulations because it slows down the energy calculation.

The simple nonpolar energy approximation outlined above has been shown to overestimate the pairwise nonpolar interaction.[12] To further decompose the cavity and VDW term[9] (eq. 1.1) (figure 1.1):

$$\Delta G_{np} = \Delta G_{cavity} + \Delta G_{vdw} \qquad (1.1)$$

Here, $\Delta G_{cavity}$ ($=\sigma * A$) is free energy that can be used to create the solvent cavity when inserting the solute from the vacuum into the solvent, while $\Delta G_{vdw}$ is the free energy measuring the solute–solvent VDW dispersion interaction. The limitation of this current simple

approximation of the nonpolar process is likely a lack of $\Delta G_{vdw}$ leading to a more favorable nonpolar interaction in the solution over the nonpolar interaction between the solute and the solvent.[12]

Nonpolar energy can also be added to another correction term, pV.[13]: $\Delta G_{np} = pV + \sigma A + \Delta G_{vdw}$ (1.2), where V and p are solute volume and volume parameter, respectively. Other terms have already been described. However, this new equation for the calculation of nonpolar energy has not been broadly tested in MD simulations.

### 1.1.2 Generalized Born theory

Since polar energy dominates nonpolar energy in total solvation energy, much effort has been dedicated to developing a more accurate polar model. Arguably, one can achieve highly accurate polar energy by solving the Poison Boltzmann (PB) Equation:

$$\Delta[\varepsilon(\boldsymbol{r})\nabla\varphi(\vec{r})] = -4\pi\rho(\boldsymbol{r}) \quad (1.3)$$

where $\varepsilon(\boldsymbol{r}), \varphi(\vec{r}), \rho(\boldsymbol{r})$ are the position-dependent dielectric constant, electrostatic potential, and charge distribution, respectively. The accuracy of the PB method comes with a trade-off, however, which is slow speed.[14]

An alternative and popular approach is the Generalized Born (GB) model. The GB model approximates polar solvation energy by summing all of the pairwise interaction energies for all of the atoms (equation 1.4 introduced by Still et al.[15]):

$$\Delta G_{GB} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\sum_{i,j}\frac{q_i q_j}{f_{ij}^{GB}(r_{ji})} \quad (1.4)$$

where ($q_i$, $q_j$) and $r_{ij}$ are partial charges and the distance between atoms i and j, respectively. The value for $f_{ij}^{GB}$ is commonly given by equation 1.5, although there is also an alternative form.[16]

$$f_{ij}^{GB}(r_{ij}) = \sqrt{r_{ij}^2 + R_i R_j \exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)} \quad (1.5)$$

Here, $R_i$ and $R_j$ are the so-called effective radii of atoms *i* and *j*. These effective radii represent the degree of burial of the atoms inside the solute. The key to success with the GB model is to get the effective radii to be close to 'perfect' radii,[16] which are obtained from the PB method by calculating the self-energy when only turning on a partial charge in the interested atom:

$$\Delta G_{self(i)} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\frac{q_i^2}{R_i} \qquad (1.6)$$

#### 1.1.2.1 Functional form of effective radii

Different GB versions have different ways of calculating effective radii; however, they can be grouped into three categories: Coulomb Field Approximation (CFA)-based models (or R4 model),[17] CFA-correction models,[18] and R6 models.[17, 19] The CFA-based model makes a rough assumption that the electric field of a point charge is not affected by the electric field of the solute. From this assumption, the inverse of the effective radii is given by the following equation:

$$\alpha_4 = R_i^{-1} = \rho_i^{-1} - \frac{1}{4\pi}\int_{r>\rho_i} |\boldsymbol{r}|^{-4} dV \qquad (1.7)$$

where $\rho_i$ is the intrinsic radius of atom *i*. The variable $\boldsymbol{r}$ is a vector originating at the center of atom i. The 3D integral is calculated in the region inside the solute but excluding the volume of atom *i*. The CFA is exact for the charge at the center of a perfect sphere. However, when the charge is off-center or the molecule is not spherical, this approximation overestimates the effective radii.[17, 19]

The effective radii equation was later generalized by an RN model (where N is an integer): [17, 19]

$$\alpha_N = R_i^{-1} = \left(\rho_i^{3-N} - \frac{N-3}{4\pi}\int_{r>\rho_i} |\boldsymbol{r}|^{-N} dV\right)^{1/(N-3)} \qquad (1.8)$$

Since CFA-based model has limitations, Lee et al.[18] introduced a correction term, $\alpha_5$ (N=5) to $\alpha_4$ (named GBMV model),[18] and later linearly combined $\alpha_4$ (N=4) and $\alpha_7$ (n=7) to form a more accurate GBMV2 model: [20]

$$R_i^{-1} = \left(1 - \frac{\sqrt{2}}{2}\right)\alpha_4 + \frac{\sqrt{2}}{2}\alpha_7 \qquad (1.9)$$

An alternative equation for the effective radii can be achieved by using N=6 (the R6 model):

$$\alpha_6 = R_i^{-1} = \left(\rho_i^{-3} - \frac{3}{4\pi}\int_{r>\rho_i} |\boldsymbol{r}|^{-6} dV\right)^{1/3} \quad (1.10)$$

Mongan et al.[17] has shown that the CFA-based correction model (GBMV2) is actually a special case of the R6 model. The performances of R6 and GBMV2 relative to the PB calculation are similar when the effective radii are numerically derived.[17] Summary of selected GB models is shown in table 1.1.

**Table 1.1** Selected GB models in the popular simulation packages AMBER and CHARMM

| GB model | Year | Programs | Calculation method | Solute/solvent boundary | Category |
|---|---|---|---|---|---|
| GB-HCT[21] | 1995 | AMBER[22] | Analytical pairwise | VDW | CFA |
| GB-OBC[23] | 2004 | AMBER[22] | Analytical pairwise | VDW | CFA |
| GB-Neck[24] | 2007 | AMBER[22] | Analytical pairwise | VDW + "Neck" correction | CFA |
| GBMV2[20] | 2003 | CHARMM[25] | Numerical integration | MS | CFA correction |
| GBSW[26] | 2003 | CHARMM[25] | Numerical integration | VDW + smooth boundary | CFA correction |
| R6[27] | 2013 | n/a | Numerical integration<br>Analytical approximation | MS<br>VDW | R6 |

Although the CFA-based model overestimates the effective radii, it is much easier to derive the analytical approximation for equation 1.7, either by the pairwise approximation approach introduced by Hawkins et al.[21] or Gallicchio et al.[28] The rigorous parameter fitting introduced fortuitous error cancellation that made GB effective radii and GB solvation energies match more closely to the PB calculation.[28-29] This approach also made it easier to calculate the derivative of the energy. Thus, its models have been widely used in MD simulations. The analytical form of the R6 model is still in developing-process. Its performance (as compared to the PB model) was only tested for small systems,[27] and there are no other reports of its implementation in MD simulations.

### 1.1.2.2 Generalized Born dielectric boundary.

An exact definition of the solute/solvent boundary has yet to be developed. Traditionally the PB model uses molecular surface (MS) to define this boundary. This MS is generated by rolling the solvent molecule with a given radius over the solute molecule (figure 1.2). One approach used by the GB model is to use this MS in a way similar in physical meaning to that of the PB method.[18, 20] However, there is no mathematical form for MS; the calculation for effective radii is expensive.[14] Alternatively, a less expensive (and less accurate) approach is to use the VDW surface (figure 1.2). The VDW-based approach is commonly used for analytical solutions for an effective radius.[21, 23, 28]

Once either the MD or VDW boundary is defined, there is still an arbitrary choice regarding atomic radii to be made. This choice also controls the accuracy of a given GB model as compared to an explicit solvation calculation.[30]

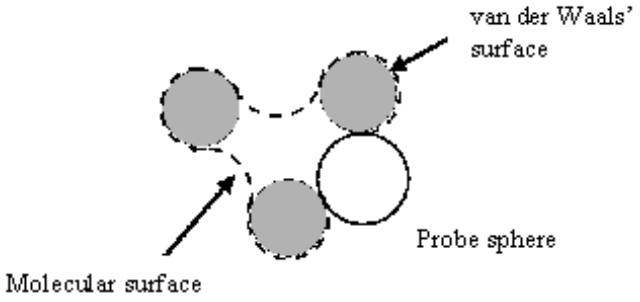


**Figure 1.2.** Molecular surface (MS) and Van der Waals (VDW) surface. MS is generated by rolling the solvent probe along VDW surface. The figure is reproduced from http://www.ccp4.ac.uk/newsletters/newsletter38/03_surfarea.html

### 1.1.2.3 Analytical model for MD simulation.

Due to the advantages listed above, GB models are of interest for use in MD simulations. Below is an introduction to several popular approaches to approximating effective radii.

#### CFA-based GB-HCT model approach.[21]

The integral for each atom i in eq.1.7 is approximated by the pairwise sum of all integrals over the spherical volumes of each atom $j \neq i$

$$\alpha_4 = R_i^{-1} \approx \rho_i^{-1} - \frac{1}{4\pi}\sum_j \int_{|r_{ij}-r|<\rho_i} |\boldsymbol{r}|^{-4} d^3r \qquad (1.11)$$

where $\rho_i$ is the intrinsic radius of atom *i*. The value for $\boldsymbol{r}$ is a vector having its origin at the center of atom i. The individual integral can easily be approximated if using the VDW volume for each atom.[21] To avoid an overlap of the volumes of all atoms j that lead to an overestimation of the integral sum, a variable called scaling factor $S_i$ was introduced to rescale the intrinsic radius: $\rho_i \rightarrow S_i * \rho_i$. The *Si* value depends upon atom type, and its value is traditionally smaller than 1.0. However, it could be treated as an adjustable parameter (>1.0) to compensate for the non-perfection of a given GB model.[24]

The VDW-based approach helps to achieve a faster, more effective radius calculation; however, this approach introduces severe errors. It does not account for the interstitial space between atomic spheres. In the MS-based approach, which is more physically correct, this region is treated as a low dielectric area; conversely, the VDW-based approach considers it a high dielectric area. This error leads to the underestimation of the effective radii of deeply buried atoms in the GB-HCT model, which in turn results in biasing the more compacted structures in the MD simulation.[31] There have been several efforts to correct this limitation. Firstly, an adjustable parameter set [α,ß,γ] was introduced to rescale up the effective radii of buried atoms while retaining similar effective radius values for the surface atoms (GB-OBC model).[23] The integral over interstitial regions are later (GB-Neck model) added to the VDW volume to make the solute/solvent boundary closer to that of the MS surface.[24]

**CFA correction-based GBSW approach.**

Another example is the effective radius calculation in the GBSW model,[26] which is an analytical form of the GBMV2 model (CFA correction-based approach). Instead of using the VDW surface to define the solute/solvent boundary, GBSW uses a smooth volume exclusion function.[26] This function was introduced to avoid the numerical instability in calculating the solvation force at the solute/solvent boundary.[20] Although GBSW was developed from the more accurate GB model, its performance relative to the PB calculation approach is no better than the GB-OBC model (which is based on the CFA approach).[14] This implies that the performance of a given analytical GB model is strongly affected by its analytical form and parameter-fitting process.

#### 1.1.2.4 GB limitation in MD simulations.

Since the GB model itself approximates the PB solution, it has a generic error (as do other implicit solvent models, including the PB method). The error increases when the analytical forms are used, rather than the numerical form.

- It lacks an explicit solvent molecule. This might lead to the instability of some proteins if localized water plays an important role.
- It lacks an explicit ion such as $Mg^{2+}$ which is required for the stability of several nucleic acid motifs such as ribosomal RNA, riboswitch, and others.[32]

Beside the generic errors, some fast analytical forms such as GB-HCT and GB-OBC strongly bias the helical structure, due to their limitations in term of effective radii calculation.[31, 33] In these models, the effective radii of deeply buried atoms are underestimated.[23] The GB MD simulations, therefore, tend to favor a more compacted structure. The GB-Neck model, which introduces interstitial region correction to GB-HCT and GB-OBC, does not like any native structures, whether beta sheet or helix. This is likely due to the systematical overestimation of the effective radii.[24, 29] GB models have also been reported to have too strong a salt bridge interaction, as compared to explicit solvent simulation.[30b, 34]

Although there is a significant number of publications using GB models for protein, this is not the case for nucleic acid simulation.[35] Most GB models tend to destabilize the DNA/RNA duplex.[36] Some GB models, such as GB-HCT[21] or GB-OBC,[23] can maintain a stable duplex in MD simulation[36] but introduce a strong helical bias in protein simulation.[31] This could be problematic if one wants to simulate the protein and nucleic acid complex.

### 1.2 Protein folding

Understanding protein folding has been challenge for 50 years.[37] Detailed knowledge of protein folding process would help better understand the relationship between structure and function; understanding misfolding process[38] which is implied relating to protein misfolding disease such as Alzheimer, Parkinson diseases or help designing protein with new function or new scaffold.

Molecular Dynamics (MD) simulations have been a powerful tool to explore atomic motion of protein. However, their strengths are limited by the time scale they can reach. For instance, the longest MD simulation before 2008 is 10 μs, utilizing ~90 days of supercomputer

time,[39] while protein folding time is in the range of μs[40] to millisecond[38, 41] or even second time scale.[37a, 42] There have been great efforts to improve the speed or the scaling of MD simulation. First, a special-purpose supercomputer (Anton)[43] was designed for MD simulation, from which few tens of μs could be generated per day. Another approach is to perform thousands of short MD simulations by using either GPU cluster or distributed computers (Folding@Home)[44] and then using Markov State Model (MSM) to build the kinetics and thermodynamics for the protein systems.[41, 45] Either approaches are successful charactering the folding of millisecond protein folding such as NTL9,[41, 46] Ubiquitin[47] or 12 different protein-fold motifs studied by Shaw et al.[46]

Two approaches above are powerful for studying protein folding, however they are not reachable to most research group. Is there an alternative way to study with much less expensive cost? Implicit solvent model, especially the pairwise Generalized Born solvent (e.g. GB-OBC[23]), in combination with GPU cards offer excellent speed.[48] Thanks to the low viscosity, GB models can significantly accelerate large scale movement.[1a]

However the speed of sampling in GB simulation comes with trade-off accuracy. As discussed above, some fast models such as GB-HCT or GB-OBC favors alpha helix[31] or overestimate ion interaction[30b] compared to explicit solvent simulations. Improving accuracy of those models is then needed for broader application in protein folding.

### 1.3 Overall goal of this dissertation

GB models' extensive number of applications (as listed in previous section) and their sampling advantages motivated me to improve their accuracy. The target GB model is the fast analytical form GB-Neck model used in MD simulation.[24] This model is the later version of the GB-HCT[21] and GB-OBC[23] (CFA-approach category) models which have been used for protein and nucleic acid simulation for 20 years, due to their rapid speed in calculating solvation energies and atomic forces. GB-Neck was shown to be more theoretically accurate than its ancestors;[17, 24] however, it tends to destabilize the native protein[49] or nucleic acid[24] in MD simulation.

The overall goal in this dissertation is to fix/reduce the current limitations of fast GB models. Specifically, a newly developed GB model should have better alpha/beta balance (compared to the older models) and have better salt bridge profile in protein simulation if using

popular explicit water simulation as benchmark. Additionally, the model should be applied to nucleic acid simulation.

I hypothesize that the original GB-Neck parameters have not yet been properly fitted to higher theory level PB method, resulting in its poor performance. This research is then organized as follows: first, I introduce the refitting procedure for protein simulation (Chapter 2), I then introduce the application of this model in protein folding (Chapter 3) and I introduce the parameter development for nucleic simulation (Chapter 4). The final chapter (Chapter 5) discusses some potential directions.

## Chapter 2. Improved Generalized Born Solvent Model Parameters for Protein Simulations

**Acknowledgments.** This chapter is direct excerpt with minor change from "Nguyen, H.; Roe, D. R.; Simmerling, C., Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J. Chem. Theory Comput.* **2013,** *9* (4), 2020-2034.". Roe and Simmerling revised the paper. Nguyen thanks Yi Shang and Christina Bergonzo for critical reading of the manuscript.

**Abstract**: The generalized Born (GB) model is one of the fastest implicit solvent models and it has become widely adopted for Molecular Dynamics (MD) simulations. This speed comes with tradeoffs, and many reports in the literature have pointed out weaknesses with GB models. Because the quality of a GB model is heavily affected by empirical parameters used in calculating solvation energy, in this work we have refit these parameters for GB-Neck, a recently developed GB model, in order to improve the accuracy of both the solvation energy and effective radii calculations. The data sets used for fitting are significantly larger than those used in the past. Comparing to other pairwise GB models like GB-OBC and the original GB-Neck, the new GB model (GB-Neck2) has better agreement to Poisson-Boltzmann (PB) in terms of reproducing solvation energies for a variety of systems ranging from peptides to proteins. Secondary structure preferences are also in much better agreement with those obtained from explicit solvent MD simulations. We also obtain near-quantitative reproduction of experimental structure and thermal stability profiles for several model peptides with varying secondary structure motifs. Extension to non-protein systems will be explored in the future.

## 2.1 Introduction

In order to accurately describe the properties of biomolecules in aqueous environment, solvent effects must be included in the Molecular Dynamics (MD) simulation. Solvation can be explicitly represented as atomistic solvent molecules or it can be implicitly represented by a model that calculates solvation effects using a continuum representation. Although implicit solvent is less realistic than explicit solvent model, it is still widely used[50] due to low computational cost, and many models directly provide solvation free energies as compared to the potential energies provided by explicit models. This has led to wide use in the drug discovery field of implicit solvent models in post-processing trajectories originally performed in explicit solvent.[51] In addition, the low viscosity in implicit solvent simulations can accelerate the rate of conformational sampling (such as protein folding) compared to explicit solvent.[1a]

Solvation free energy can be decomposed into two terms for the polar and nonpolar contributions. The present work focuses solely on the polar contribution. The nonpolar term is often approximated by the equation $\Delta G_{np}=\gamma A$ where $\gamma$ is the surface tension coefficient and A is the total solvent accessible area. The nonpolar term is frequently omitted in simulations due to the cost of calculating the surface area and its derivatives, and the fact that the magnitude of this term is typically much smaller than the polar contribution. Moreover, a simple solvent accessible surface area (SASA) based approximation that is commonly used to calculate the nonpolar term has several limitations.[9, 12-13, 52] Chen et al.[12] have shown that this nonpolar model tended to overestimate nonpolar interactions that shifted ensembles to non-native states. Despite these limitations, SASA-based approaches are widely used and available in the Amber program[22b], thus we evaluate the impact of their inclusion during simulations using our improved GB model.

Among all implicit solvent models, the Poisson- Boltzmann (PB) method[53] is considered the most accurate model for calculating polar solvation energy in MD. However, the computational cost of solving the PB equation and its derivatives, particularly on massively parallel computers, is high enough that it is not widely used in MD simulations.[14] Instead, most MD simulations use the GB equation (eq. 1), as was first introduced by Still et al.[15] and subsequently modified by other groups.

$$\Delta G_{GB} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}\right)\sum_{i,j}\frac{q_i q_j}{f_{ij}^{GB}(r_{ij})} \quad \text{(eq.1)}$$

Here $f_{ij}^{GB} = \sqrt{r_{ij}^2 + R_i R_j \exp(\frac{-r_{ij}^2}{4R_i R_j})}$; $q_i$, $q_j$ are the partial charges of atom $i$ and $j$; $r_{ij}$ is the distance between atom $i$ and $j$; $\varepsilon_{in}$ and $\varepsilon_{out}$ are interior and exterior dielectric constants respectively. $R_i$ and $R_j$ are the effective Born radii. It has been shown that accurate calculation of effective radii (or using 'perfect' radii calculated from PB method) is a key to close agreement between GB and PB solvation energies.[16] The effective radius is normally calculated by eq. 2

$R_i^{-1} = \rho_i^{-1} - I_i$ (eq. 2)

where $I_i$ is Coulomb integral derived from Coulomb Field Approximation (CFA)

$I_i = \frac{1}{4\pi}\int_{\Omega, r>\rho_i} \frac{1}{\mathrm{r}^4} d^3\mathrm{r}$ (eq. 3)

$\rho_i$ is the intrinsic radius of the atom $i$ and integral $I_i$ is calculated over the volume Ω outside atom $i$ but inside the molecule. $I_i$ can be calculated numerically[15] or analytically by using the pair-wise descreening approximation (PDA) method introduced by Hawkins et al. (the GB-HCT model).[21] Although GB-HCT is less computationally expensive than numerical methods,[14] it tends to underestimate the effective radii of buried atoms.[54] A modification based on GB-HCT was proposed by Onufriev et al.[23] (GB-OBC), in which effective radii for buried atoms are scaled up by an adjustable empirical parameter [$\alpha, \beta, \gamma$] set (eq. 4a).

$R_i^{-1} = \tilde{\rho}_i^{-1} - \rho^{-1}\tanh(\alpha\psi - \beta\psi^2 + \gamma\psi^3)$ (eq. 4a)

where $\tilde{\rho}_i = \rho_i - offset, \psi = \tilde{\rho}_i I$ (eq. 4b)

Importantly, these analytical models (GB-HCT and GB-OBC) use the van der Waals (VDW) surface to define the boundary between solvent and solute, instead of using more realistic but much more computationally demanding molecular surface (MS). Mongan et al.[24] introduced a "neck" correction to make the space defined by the VDW boundary closer to that defined by MS boundary, particularly at small interatomic distances where finite size explicit water is typically excluded (GB-Neck). (figure 2.0.1)

$I_{"MS"} = I_{vdw} + \int_{neck} \frac{1}{r^4} d^3 r$ (5)

where $I_{vdw}$ is the integral $I_i$ in eq. 3, using VDW volume for volume Ω . $I_{MS}$ is then applied as $I$ in eq. 4b.

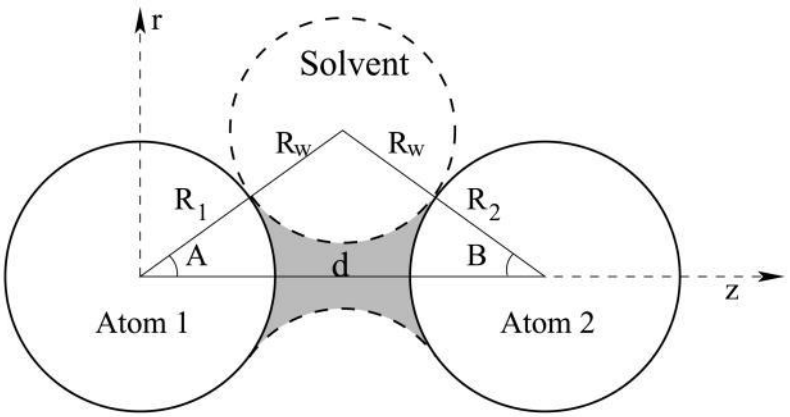


**Figure 2.0.1.** "Neck" correction (shaded) for GB-OBC model[23] with a simple two-atom system. GB-OBC only uses VDW volumes from atom 1 and 2 for effective radii calculation. The figure is reproduced from Mongan et al.[24]

All 3 of these PDA-based GB models have some advantages such as low computational cost,[14, 24] and in particular efficient parallel scaling compared to explicit solvent models.[4] These GB models have also been ported to GPU-based MD codes which accelerate MD up to 700 times faster than simulation on conventional CPUs.[3] The advantage of speed, however, comes with less accuracy in these GB models. GB-HCT and GB-OBC have apparent limitations such as high alpha helical content[31, 33-34, 49, 55] and overly strong ion interactions compared to TIP3P explicit solvent simulations.[30b, 34] Although GB-Neck introduced corrections to GB-OBC, this is not reflected in improved solvation energy accuracy.[24] Additionally, Dill et al.[49] and Roe et al.[31] have shown that GB-Neck tends to destabilize native peptide/protein structures, likely due to imbalance between intramolecular hydrogen bonds and interaction with implicit solvent.

Our goals for improving the GB model are to give more accurate solvation energy and effective radii calculation compared to PB method; to reduce secondary structure and salt bridge bias, and to better reproduce experimental structures and thermal stability for small proteins and peptides. We hypothesize that at least some of these weaknesses could be improved by more rigorous fitting of the many empirical parameters in these models. Since GB-Neck is more physically realistic than GB-HCT and GB-OBC, we decided to use it as the base model for our parameter refitting. The relatively poor performance of GB-HCT in many studies led us to omit it from the present comparisons.

In the original GB-Neck work,[24] 8 parameters were optimized by fitting GB solvation energies to PB solvation energies for a set of proteins and peptides. The GB-Neck parameters

include scaling factors $S_x$ (x=H, C, N, O) that were initially introduced in GB-HCT by Hawkins et al.[21] for analytically calculating the *I* integral in eq. 3, the [$\alpha, \beta, \gamma$] set used in eq. 4a that was initially introduced in GB-OBC by Onufriev et al,[23] and the neck scale factor $S_{neck}$ introduced by Mongan et al.[24] These describe properties related to gaps between atom pairs, and are thus likely dependent on size of the atoms involved. We therefore expanded the number of parameters from 8 to 18 (see method section) by making [$\alpha, \beta, \gamma$] atomic number dependent and making *offset* (eq. 4b) a free parameter as well. We recognized that the significant increase in the number of free parameters in the model necessitated use of much larger training and test sets than used previously, and thus much of the present work focuses on development of a large and broad data set for training and testing. Summary of parameter set difference between GB-Neck and new model is given in table 2.0.

**Table 2.0.** Parameters for the original GB-Neck and GB-Neck2 models.

| GB-Neck | GB-Neck2 |
|---|---|
| $S_H$ | $S_H$ |
| $S_C$ | $S_C$ |
| $S_N$ | $S_N$ |
| $S_O$ | $S_O$ |
| *offset* | *offset* |
| $S_{neck}$ | $S_{neck}$ |
| $\alpha$ | $\alpha_H$ |
| $\beta$ | $\beta_H$ |
| $\gamma$ | $\gamma_H$ |
| | $\alpha_C$ |
| | $\beta_C$ |
| | $\gamma_C$ |
| | $\alpha_N$ |
| | $\beta_N$ |
| | $\gamma_N$ |
| | $\alpha_O$ |
| | $\beta_O$ |
| | $\gamma_O$ |

In our training objective function, not only absolute solvation energy but also effective radii and relative solvation energy of peptide (or protein) conformations were included. PB solvation energies and 'perfect' radii of structures in the training set were used as benchmarks for fitting. The new GB-Neck parameter set (GB-Neck2) shows significant improvement in

accuracy of calculating solvation energy and effective radii compared to GB-OBC and original GB-Neck model for this training set. Importantly, the improvement is clearly transferable to test sets having thousands of structures for various proteins and peptides, including molecules not used in training.

The final goal of a GB model is to approximate the results (structure, stability, salt bridge profile etc.) obtained from more expensive explicit solvent simulations; thus we performed simulations of several peptides in GB-Neck2 as well as in explicit water to test if the improved agreement of GB-Neck2 to PB results (solvation energy and effective radii) led to improved agreement of structural ensembles compared to those obtained from explicit solvent simulation. Overall, the GB-Neck2 model does a much better job in reproducing ensemble data from explicit water (such as alpha-helical stability) as compared to GB-OBC and comparable to the original GB-Neck, with the exception of propensity to form ion pairs (salt bridges). We found that although salt bridges were specifically included in our training by fitting to PB solvation energies, they tended to remain too strong in GB-Neck2 when comparing to TIP3P simulation. A possible explanation is that PB also has too-strong ion interactions compared to TIP3P, perhaps arising from our use of the same set of intrinsic Born radii in our GB and PB calculations for consistency.[56] Salt bridge strength was thus adjusted in the same strategy as Geney et al.[30b] and Shang et at.[56] by empirically adjusting the Born radius of side-chain $H^{N+}$ of Arg to reproduce salt bridge PMF of TIP3P simulation. Unlike those earlier studies, we also adjusted the Born radius of side-chain Oε of Glu (and Oδ of Asp) to match PMF profiles of salt bridges and hydrogen bonds in TIP3P simulations. This radius modification was sufficient to reproduce the PMF of Lys salt bridge formation and we found no need to modify the Born radius of $H^{N+}$ in Lys.

We also tested the ability of GB-Neck2 in combination with widely used ff99SB force field[57] in reproducing experimental structure and thermal stability of different peptides from experiment by simulating a hairpin (HP5F)[58] system and a mini-protein with alpha, 3-10 and polyproline helices and a small hydrophobic core (trp-cage variant tc5b).[59] The effect of including a nonpolar solvation energy term in GB simulations was also tested. Although the agreement of melting temperature between simulation and experiment depends not only on the GB model but also on the protein force field, this testing is still valuable to confirm the robustness of the combination of specific GB model and force field. Dill et al.[49] evaluated various combinations of force fields and GB models for peptide and protein simulations and

found that GB-OBC[23] with the ff96 force field was the best combination for studying protein folding. However, this GB model and this force field both have well-known flaws, and thus it is likely that the combination benefits from significant fortuitous error cancelation. In the present case, we use the ff99SB protein force field (FF),[57] which has been shown by many studies to provide excellent results with explicit water.[60] We find that this single combined protein FF + solvent model is able to quantitatively reproduce the experimental thermal stability behavior of two tested peptide models with different secondary structures. Taken together, our results lead us to recommend this combination for simulations of peptides and proteins.

## 2.2 Materials and Methods

### 2.2.1 Training set for parameter fitting

We first designed test sets of between ~3,500 - 103,000 structures of each protein or peptide, and then took a subset of the structures for the training set. The subset was selected in a way that gives both training and test set similar absolute solvation energy root-mean-square-deviation (*abs_e*) between GB-OBC and PB solvation energies. For example, the Ala10 test set had 50000 structures with *abs_e* of 1.12 kcal/mol between GB-OBC and PB. The Ala10 training set had only 413 structures with *abs_e* of 1.14 kcal/mol. This reassures us that a small number of structures could represent a desired quality metric (*abs_e* in this case) of a larger number of structures. The assumption is tested by evaluating the model using the full test set, which was impractical during training due to the large number of parameter variations that were tested. An overall summary of the training sets and their contributions to the training objective function is given in Table 2.1. These sets are discussed in more detail below.

Roe et al.[31] used enhanced sampling Replica Exchange Molecular Dynamics (REMD)[2] simulations of Ala10 peptide to quantify the helical bias in GB-OBC and GB-HCT models. We used this system in our test and training sets; the training set Ala10_set_1 has 480 structures extracted from Thermodynamic Integration (TI) and REMD trajectories from Roe et al.[31] We first introduced 50 alpha and 50 hairpin structures from TI trajectories and then added 10 structures from each of the 20 most populated clusters sampled at 300K in 50 ns REMD using TIP3P[61] explicit water, as well as single representative structures for the next 180 clusters.

Okur et al.[34b] used the peptide sequence RAAE (Arg-Ala-Ala-Glu) to evaluate salt bridges in the GB-OBC model. Our RAAE set has 200 structures taken from the 300K trajectory

of TIP3P REMD simulation from Okur et al.[34b] We chose structures uniformly sampling the salt bridge distance (Cζ of Arg and Cδ of Glu) ranging from 3.6 Å to 14.5 Å with nearly equal interval of 0.05 Å.

We also added structures for two peptides having different secondary structures: we also added β-hairpin (trpzip2, PDB ID: 1LE1)[62] and α-helix (3Ai3)[63]; these also have more complicated side chains than Ala10 and RAAE. The trpzip2 set had 413 structures from Okur et al.[8] Those structure ensembles were chosen from cluster analysis of MD (or REMD) simulation trajectories, giving various types of backbone structures from helix, hairpin, PPII, and coil. The backbone RMSD for those structures to native trpzip2 is presented in Figure 2.S1. The 3Ai3 set had 200 structures of the peptide sequence Ac-YGG-$(KAAAA)_3$-K-$NH_2$, one of the helical peptides studied by simulation and NMR in Song et al.[63] We chose structures first by clustering the first 50ns of 300K trajectory data from GB-HCT REMD simulation of 3Ai3[63] and then picking 200 structures from the 20 most populated clusters (10 structures / cluster).

Because Ala10, RAAE, trpzip3, 3Ai3 were small peptides, we also added HP36 mini-protein[64] structures to train for structures having a hydrophobic core. These were extracted from the first 75 frames of 300K MD simulation in TIP3P from Wickstrom et al.[65] (the backbone RMSD to X-ray structure (PBD ID: 1YRF[66]) is given in Figure 2.S2).

Structure sets described above were used for training solvation free energy as compared to PB data. We also included two structure sets, Ala10_set_2 and HP1113, to train for effective Born radii. Ala10_set_2 has 200 Ala10 structures (50 structures for each alpha helix, hairpin, left handed helix ("left"), PPII) which were extracted from trajectories of TI calculations from Roe et al.[31] We added additional large protein structures to evaluate the effective radii underestimation of deeply buried atoms.[54] The HP1113 set has 6 large proteins having various secondary structure types, with PDB ID codes 1TSU,[67] 1BDD,[68] 1UBQ,[69] 1AEL,[70] 1FKG,[71] 3GB1[72] (details in Table 2.S0.1).

### 2.2.2 Test sets for evaluating the new model.

We designed two test set types. Test set type I (Ala10, trpzip2, 3Ai3, RAAE, HP36 sets) had proteins or peptides for which a smaller set of structures were included in the training stage. This tests the extension of the model to a broader set of structures (thousands rather than hundreds). Test set type II had proteins or peptides for which structures were <u>not</u> included in

solvation energy training (tc5b, DPDP, HIV1-PR, Lysozyme), testing the transferability to entirely different molecular systems. The large numbers of structures were chosen for testing to ensure local variation in structure as well as alternative folds. Summary for the test sets is given in Table 2.S0.2

**2.2.2.1 Test set type I.** Ala10 set has 50000 Ala10 structures taken from 50 ns of 300K trajectory of REMD simulation in TIP3P.[31] Trpzip2 [62] set has ~80000 structures having RMSD to native structure ranging from 0.2 to 7.6 Å (Figure 2.S3) which were taken from TIP3P and GB simulations.[8] The 3Ai3 set had 49000 structures taken from 49ns of 300K trajectory of GB-HCT REMD simulation.[63] The RAAE set had 50000 structures from 50ns of 300K trajectories of TIP3P REMD simulation of RAAE peptide.[34b] The HP36[64] set had 3500 structures extracted from the first 35ns (skipping every 10 frames) of TIP3P MD simulation at 300K from Wickstrom et al.[65] (Figure 2.S4).

**2.2.2.2 Test set type II.** Test set type II has 4 structure sets representing 4 different protein types, which are a helical mini-protein (trp-cage tc5b variant),[59] a small peptide having 3-stranded β-sheet (DPDP),[73] a larger, mainly helical protein (lysozyme)[74] and a larger protein having mainly β-sheet (HIV-1 protease).[67] The tc5b set had 103000 structures having backbone RMSD from 0.3 to 8.0 Å to native TC5b (Figure 2.S5), which were extracted from TIP3P and GB simulation ensembles.[30b] The DPDP set had 50000 structures from 150 ns GB-HCT REMD simulation trajectory at 300K.[75] HIV-1 PR had 1427 structures having closed, semi-open and wide open conformations of HIV-1 PR protein which were extracted from 600 ns trajectory at 300K of 1TSU [67] in TIP3P (Cα RMSD to closed X-ray structure is given in Figure 2.S6) [76]. The lysozyme set had 1000 structures taken from first 30 ns of 300K trajectory of TIP3P MD simulation (backbone RMSD to experimental native structure (PBD ID: 1IEE[74]) is given in Figure 2.S7).[57]

**2.2.3 PB calculations and intrinsic radii**. All PB calculations were performed using Delphi v2 and v4[77] with grid spacing of 0.25 Å and solvent probe of 1.4 Å (different Delphi versions were used due to the their availability in computer clusters). Interior and exterior dielectric constants of 1.0 and 78.5 respectively were used for solvation energy calculation. Exterior dielectric constant of 1000 was used for calculating 'perfect' radii, as suggested by Sigalov et al.[78]

Calculation of "perfect' radii was described by Onufriev et al.[16] The original GB-Neck suggested use of the bondi radii set,[79] however it was shown by Dill et al.[49] that this combination tended to destabilize protein native structure. Onufriev et al.[23] showed that GB-OBC worked quite well with mbondi2 radii, consistent with our previous observations with this combination.[6, 56] We reasoned that GB-Neck was an improvement of GB-OBC model and thus, mbondi2, instead of bondi, should be a good starting radii set. For consistency, the same radii were used in GB and PB. We therefore used mbondi2 intrinsic Born radii set[23] and charge set from ff99SB[57] in all PB and GB solvation energy calculations. Radii adjustment will be discussed below.

### 2.2.4 Fitting parameters and procedure

In the original GB-Neck model,[24] Mongan et al. fit only 8 parameters: $S_x$ (x=H, C, N, O), [$\alpha, \beta, \gamma$] and $S_{neck}$. We refit 18 parameters by allowing $S_x$, $\alpha, \beta, \gamma$ to vary for H, C, N, and O. We tried several parameter combinations for Sulfur (S) atom and found that S parameters have insignificant effect on correlation between GB and PB solvation energies due to a small number of S atoms in protein molecule. Thus, the S parameters were arbitrarily chosen to be the same as the ones for Oxygen (O). $S_x$ is a scaling parameter originally introduced by Hawkins et al. in the pairwise GB-HCT model[21] to avoid double counting of overlapping VDW volume. $S_x$ was conventionally considered to range from 0 to 1, but the search space was extended greater than 1 in the original GB-Neck paper.[24] In our optimization, we also extended the range of $S_x$ to [0.0, 2.0]. $[\alpha, \beta, \gamma]_x$ (x=H, C, N, O) are adjustable parameters used in eq. 4a. Onufriev et al.[23] and Mongan et al.[24] used one set of [$\alpha, \beta, \gamma$] for all atoms, but we allowed different elements to adopt their own parameter to allow for atomic-size dependence of the interstitial gaps for which these parameters empirically correct. We chose [0.0-10.0] as potential range for these parameters. We also attempted fitting with one parameter set for all elements like Onufriev et al.[23] or Mongan et al.[24] did, but found no significant improvement in solvation energy calculation compared to GB-OBC and GB-Neck (data not shown). The *offset* parameter (eq. 4b) was originally used by Still et al.[15] to decrease atomic radii to maximize the agreement between GB and experimental solvation energies for a set of small molecules, and it has been used in several GB models such as GB-HCT,[21] GB-OBC[23] and GB-Neck[24] as a conventional constant (*offset=0.09*). In our study, we treated *offset* as an adjustable parameter with possible range of [-0.2, 0.2]. $S_{neck}$ is the scaling factor introduced by Mongan et al.[24] to avoid the overlap of neck regions in nearby pairs, and

thus reducing the over calculating of neck integral (eq. 5). We kept the original range of [0.0, 1.0] for $S_{neck}$. In summary, the search range of each parameter set is $S_x \in [0.0, 2.0]$, $[\alpha_x, \beta_x, \gamma_x] \in [0.0, 10.0]$ ($x$= *H, C, N, O*), $offset \in [-0.2, 0.2]$ and $S_{neck} \in [0.0, 1.0]$.

GB MD simulation biases have been shown to correlate with differences between GB and PB solvation energies.[31] In this work, we define "absolute solvation energy root-mean-square-deviation" (*abs_e*) as RMSD of solvation energies for a set of conformations, where the error is the difference between GB and PB energies. "Relative solvation energy root-mean-square-deviation" (*rel_e*) was calculated as the RMSD for GB and PB energy differences for all pairs of structures. "Effective radii root-mean-square-deviation" (*eff_rad_rmsd*) is defined as RMSD of GB effective Born radii from those calculated using PB ('perfect' radii).[16] Typically, only *abs_e* is considered when optimizing GB models.[23-24] However, we consider *rel_e* and *eff_rad_rmsd* as important additional targets. The *rel_e* is included since it is the relative energy of alternate conformations that determines thermodynamic populations, such as those sampled in MD simulations. The *eff_rad_rmsd* is included since Onufriev et al.[16] showed that the best agreement between GB and PB solvation energy is obtained when using 'perfect' radii from PB calculation in GB, later confirmed by Honig et al.[55] We set our objective function for training as the sum of weighted *abs_e*, *rel_e* and *eff_rad_rmsd*. The objective function is shown in eq. 6, where $w_i$ is weighting factor and $x_i$ is contribution of each component $i$ (each set in Table 2.1).

$$obj\_funct = \sum_i w_i x_i \,, \quad (6)$$

We weighted *abs_e*, *rel_e*, and *eff_rad_rmsd* so that they contributed roughly equally to the objective function. We first minimized the objective function with $w_i = 1$ for all contributions and calculated how far each contribution value could be decreased from those calculated by GB-OBC model. The *abs_e* values for trpzip2, 3Ai3, HP36 decreased a few kcal/mol but the *abs_e* of Ala10_set_1 and RAAE set decreased only ~0.5 kcal/mol. Thus, we used $w_i = 1$ for *abs_e* of trpzip2, 3Ai3 and HP36 sets while used $w_i = 10$ for *abs_e* of Ala10_set_1 and RAAE set. We set other weighting factors in a similar way. Weighting factors for different contributions are summarized in Table 2.1. Our choice of weighting factors is not unique and thus others could choose different $w_i$.

**Table 2.1**. *abs_e* (kcal/mol), *rel_e* (kcal/mol) and *eff_rad_rmsd* (Å) to PB results for each training set after optimization, compared to GB-OBC and GB-Neck models. *w* is the weighting

factor for each component. The objective function for training is the sum of the weighted contributions from each column.

| | Ala10_set_1 *abs_e* *w=10* | Ala10_set_1 *rel_e* *w=10* | trpzip2 *abs_e* *w=1* | trpzip2 *rel_e* *w=10* | 3Ai3 *abs_e* *w=1* | 3Ai3 *rel_e* *w=10* | RAAE *abs_e* *w=10* | RAAE *rel_e* *w=10* |
|---|---|---|---|---|---|---|---|---|
| GB-OBC | 1.1 | 1.6 | 9.5 | 4.8 | 7.1 | 4.5 | 1.5 | 1.3 |
| GB-Neck | 2.7 | 2.3 | 8.3 | 7.0 | 10.6 | 4.1 | 1.1 | 1.1 |
| GB-Neck2 | 0.8 | 1.0 | 2.8 | 3.9 | 3.7 | 3.7 | 0.9 | 1.2 |

| | HP36 *abs_e* *w=1* | HP36 *rel_e* *w=10* | HP1113 *eff_rad_rmsd* *w=50* | Ala10_set_2 *eff_rad_rmsd* *w=250* | *obj_funct* |
|---|---|---|---|---|---|
| GB-OBC | 21.6 | 6.5 | 1.8 | 0.16 | 381.2 |
| GB-Neck | 28.3 | 5.1 | 2.3 | 0.19 | 444.7 |
| GB-Neck2 | 4.3 | 4.8 | 1.5 | 0.10 | 273.8 |

The search space for fitting 18 parameters is vast, thus we did not expect to locate the global minimum for our objective function. Our goal instead is to have a parameter set showing significant improvement in solvation energy and effective radii calculation relative to PB when comparing to GB-OBC and original GB-Neck models, and one that simultaneously accounts for many aspects of the training data. In the beginning of this project, we used the local search method UOBYQA[80]. UOBYQA is an unconstrained minimization method that does not require objective function derivatives, and allows optimization with large number of variables. It took about 1 minute for the objective function calculation and we spent about 20 days (~30000 function evaluations) for each optimization. Because of the computational expense, we performed only 5 optimization runs, each run starting with an initial random guess. We then performed further extensive search of combinations of various parameter subsets by using a parallel Genetic Algorithm (GA)[81] (code implemented by Metcalfe et al.).[82] The GA is one of the most popular global search methods[83] and particularly well suited for this task because it is likely

that some of the parameters are weakly coupled, and thus mating of genes with independent improvements located in different parameters could be productive. GA options such as mutation and crossover rate were set to default values [82]. Each optimization run had population size of 120 and the objective function was allowed to be evaluated up to 2500 generations. We performed 31 runs in total. Initial populations of most runs were randomly created. Parameters of GB-OBC, GB-Neck and previous UOBYQA results were also included in some runs as initial guesses.

### 2.2.5 Structures used for testing the new parameters in MD simulations

All of the fitting described above was performed relative to PB solvation energy calculations. This is consistent since both lack description of the hydrophobic and van der Waals components of the aqueous solvation, and thus by design our fitting did not modify the electrostatic component of GB to empirically correct for these missing terms as it would if we fit directly to reproduce data from explicit solvent simulations, which would likely lead to reduced transferability. However, we hypothesized that improved agreement with PB would also result in improved agreement with explicit water.[31] To test this hypothesis, we compared simulations generated with the re-optimized GB-Neck parameter set (named GB-Neck2) with those from TIP3P explicit water for several systems.

**2.2.5.1 Ser-Ala-Ala-Glu Model Peptide (SAAE).** SAAE was used to compare hydrogen bond (Hγ of Ser and Oε of Glu) PMFs between GB and TIP3P models. The potential of mean force for hydrogen bond formation was a useful independent measure of model quality, but was also performed because one of our intermediate models showed much too high propensity of forming such interactions as compared to MD in explicit water. Since the solvation energy profile matched that in PB using the same intrinsic Born radii (data not shown), we built on our previous work[30b, 56] that showed adjusting intrinsic radii could improve fit to explicit solvent data. Additionally, the strong salt bridge interaction between side chains of Asp (or Glu) and Arg could be adjusted by modifying the radius of either the $H^{N+}$(Arg) or the carboxyl oxygen[34b]. Thus, the SAAE model was built to compare the H-bond PMFs of GB models to TIP3P model and allow adjustment the radii of carboxyl oxygen atoms independent from subsequent adjustment of $H^{N+}$ to refine salt bridge strength.

We performed 3 REMD simulations using TIP3P, GB-OBC, GB-Neck and 2 REMD simulations for GB-Neck2 with original and modified carboxyl oxygen intrinsic radius (1.5 and 1.4 Å respectively). GB-OBC and GB-Neck runs were used as controls. SAAE was solvated in a truncated octahedron box with 8 Å buffer by using 459 TIP3P water molecules. TIP3P REMD simulation was performed for 60 ns while GB REMD simulations were extended to 50 ns. Because Glu had two symmetric Oε in the side chain, we used PMFs of distance between Hγ (Ser) and Cδ (Glu) to define PMFs of H-bond instead of distance between Hγ (Ser) and Oε (1 and 2) of Glu. The choice of Cδ (Glu) for PMF calculating was consistent with previous reports.[30b, 34b, 56]

**2.2.5.2 Arg-Ala-Ala-Glu Model Peptide (RAAE).** RAAE was used as test system due to its small size and the availability of TIP3P data from Okur et al.[34b] Okur and coworkers[34b] demonstrated that Arg salt bridge strength in RAAE was 2.5-3 kcal/mol stronger in GB-OBC than in TIP3P. Shang et al.[56] later corrected this GB-OBC overestimation by reducing radii of $H^{N+}$( Arg) from 1.3 Å to 1.1 Å. We had initially hypothesized that simply including RAAE structures in training would help reduce salt bridge strength, and performed GB-Neck2 simulations using original mbondi2 radii. However, the stability was still significantly overestimated compared to TIP3P, and we thus performed several simulations with various combinations of modified radii to identify a value that better reproduced the PMF in TIP3P. Particularly, we performed REMD simulations of GB-Neck2 by using unmodified mbondi2 H radii and also using 4 modified radii for $H^{N+}$: using a radius of 1.4 Å Oε (Glu) with 1.3 Å, 1.2 Å, 1.17 Å or 1.1 Å for $H^{N+}$ (Arg). 300 K trajectories from TIP3P and GB-OBC mbondi2 REMD simulations from Okur et al.,[34b] GB-OBC 1.1 $H^{N+}$( Arg) REMD simulation from Shang et al.[56] and GB-Neck mbondi2 REMD simulation were used for comparison with GB-Neck2 modified mbondi2 simulations. RAAE protocols and initial structures were taken from Okur et al.[34b] We used one 40ns run for GB-Neck and GB-Neck2 simulations.

**2.2.5.3 Lys-Ala-Ala-Glu Model Peptide (KAAE).** In addition to Arg, Lys can also participate in salt bridge interactions. We compared the KAAE salt bridge PMF of GB-Neck2 simulation to that from TIP3P simulation. As with RAAE, it was built in a helical backbone conformation to allow favorable salt bridge orientation.[34b] We performed REMD simulations for TIP3P, GB-

OBC, GB-Neck as controls and two simulations for GB-Neck2 (with mbondi2 and with mbondi3 (Table 2.S1)) to see which simulation of GB-Neck2 could best reproduce TIP3P salt bridge PMF.

The RAAE protocol was adopted for KAAE. All simulations of GB models were run up to 50 ns while TIP3P simulation was run for 30ns. The distance between Nζ of Lys and Cδ of Glu was defined as salt bridge distance. A large solvation buffer length was defined to minimize periodicity artifacts in the PMF[34b], using a truncated octahedron box with 16 Å buffer and 2433 TIP3P water molecules.

The SAAE, RAAE and KAAE peptides were built using the tleap program in Amber10 with acetylated and amidated N- and C-termini. The radii obtained from these optimizations is denoted mbondi3 (Table 2.S1).

**2.2.5.4 Ala10 Model Peptide.** Alanine decapeptide (Ace-$Ala_{10}$-$NH_2$) was used to compare secondary structure content (DSSP)[84] and local structural propensities between GB and TIP3P simulations following Roe et al.[31] To test if the improvement observed in our training would translate to better secondary structure balance in MD, we repeated Roe's protocol with GB-Neck2. DSSP and local structure propensities from GB-Neck simulation were compared with the ones from TIP3P, GB-OBC and GB-Neck models. Eight replicas were used for REMD simulations, starting from extended conformations. One 50ns REMD run was performed for GB-Neck2 and compared to GB-OBC, GB-Neck and TIP3P data from Roe et al.[31]

**2.2.5.5 HP-1 Model Peptide.** Because Ala10 structures were used in fitting GB-Neck2 parameters, we desired an independent test of the change in helical propensity. HP-1 is good candidate since it is nearly the same size as Ala10 and showed moderate Because Ala10 structures were used in f[85]. HP-1 (MLSDEDFKAVFGM) is adopted from the N-terminal helix of HP36, a 36-residue helical subdomain of the villin headpiece. As with Ala10, we compared DSSP and local conformational propensities between GB simulations (GB-OBC, GB-Neck and GB-Neck2) and TIP3P. 300K trajectories from TIP3P and GB-OBC REMD simulations were taken from Wickstrom et al.[85] We performed 2 REMD simulations up to 50ns for GB-Neck and GB-Neck2. Nonhelical structures extracted from TIP3P REMD simulation were used as initial

structures for REMD. Because HP-1 peptide has Lys salt bridge potential, we the optimized mbondi3 radii set (Table 2.S1) for GB-Neck2 simulation.

We further tested the robustness of GB-Neck2 by evaluating the ability to reproduce the experimental thermal stability for 2 small peptides for which experiments indicate different secondary structure motifs different from the unstructured Ala10 and helical HP-1: a hairpin structure (HP5F),[58] and the trp-cage tc5b mini-protein[59] that has alpha and 3-10 helix as well as a PPII strand and a small hydrophobic core. The short length and the availability of experimental melting temperature (and melting curve for tc5b) made these two sequences ideal for testing. For each protein, we performed 2 REMD simulations starting from folded and linear structures. Simulated melting curves for those proteins were generated by calculating fraction folded (the fraction of the number of frames having native structure over the total number of frames from simulation) versus temperature. When comparing melting temperature between simulation and experiment, it is important to note that the agreement depends on not only on the solvation model but also the protein force field. As discussed above, we employed the widely used ff99SB force field, but disagreement with experiment can arise from many sources outside GB model accuracy so one must use caution when interpreting the results.

**2.2.5.6 HP5F model peptide.** HP5F[58] is a short peptide with sequence KKYTWNPATGKFTVQE.[58] We first simulated an extended structure in GB-Neck2 using REMD, and then extracted the representative structure for most populated cluster at 300K. This representative "folded" structure was used to initiate a second independent REMD run. REMD simulations were run to 150 ns, 75 ns and 90 ns for GB-OBC, GB-Neck and GB-Neck2 respectively. We also performed an additional 70 ns run for GB-Neck2 with a SASA (solvent accessible surface area) based nonpolar solvation term (*gbsa = 1* in Amber) to test the effect. An experimental atomic structure of HP5F has not been reported, but as it is expected to adopt the same fold as the GB1p peptide,[58, 86] we used the GB1p backbone for calculating RMSD during HP5F trajectories. The GB1p structure was derived from the C-terminal hairpin of protein G (PDB ID: 3GB1,[72] residue 41-56). Residues 2 to 15 were chosen for RMSD to avoid the flexible termini. Structures having backbone RMSD smaller than 2.0 Å were defined as folded. We chose this cutoff based on the position of the minimum separating folded and unfolded regions in the simulated RMSD histogram (Figure 2.S8). A full experimental melting curve for HP5F was not

available, thus we compared our results to the experimental melting temperature and folded population at 298K.[58]

**2.2.5.7 Trp-Cage tc5b model peptide.** The tc5b[59] variant of trp-cage is a 20 residue peptide having sequence of **NLYIQWLKDGGPSSGRPPPS**.[59] The first model from the NMR structure ensemble, and a linear structure built by tleap, were used as starting structures for 2 REMD simulations for each GB model (340 ns, 240 ns, 160 ns and 72 ns for GB-OBC, GB-Neck, GB-Neck2 and GB-Neck2 with nonpolar solvation term, respectively). Different GB models have different simulation lengths since they have different time scale for convergence (having small error bars from two runs). As with HP5F, we defined folded structures by using a backbone RMSD cutoff of 2.0 Å based on the RMSD histogram (Figure 2.S9). RMSD to native tc5b was calculated for backbone atoms from residue 3 to 18 to avoid flexible termini.[30b] We compared melting curves from GB-OBC, GB-Neck, GB-Neck2 and GB-Neck2 SASA to the ones from NMR and CD experiments.[59]

### 2.2.6 Protocols for simulations and data analysis

**2.2.6.1 REMD Simulation Protocols.** All simulations used to compare GB simulations to TIP3P simulations and experiments presented in Results were carried out with AMBER 10[22b] and the ff99SB force field.[57] The AMBER 10 code was modified to support GB-Neck2; it is now available in AMBER version 11 or later by specifying *igb = 8*. All simulations used REMD[2] for enhancing sampling. The time step was 2 fs for all REMD simulations. SHAKE[87] was used for constraining all bonds to hydrogen. For small protein/peptide simulations, we did not employ a surface-area based nonpolar solvation term. Temperature was controlled by using Berendsen thermostat[88] in TIP3P[61] simulations with a time constant of 1.0 $ps^{-1}$, or by using a Langevin thermostat in GB simulations with a collision frequency of 1.0 $ps^{-1}$. Unless noted, GB-Neck simulation used mbondi2 radii,[23] GB-OBC simulations used mbondi2 with 1.1 $H^{N+}$ (Arg)[56] while GB-Neck2 simulation used mbondi3 (Table 2.S1). Further details of each simulation are given in the context. In explicit solvent simulations, peptide models were solvated with TIP3P[61] water in a truncated octahedron box. PME[89] was used for treating long range electrostatic interactions and nonbonded interaction cutoff was 8 Å. No cutoff was used in GB simulations.

Exchanges in REMD simulations were attempted every 1 ps. 32 replicas were used for TIP3P REMD while only 8 replicas were needed for Ala10, HP-1, tc5b, HP5F GB REMD and 6 replicas were used for GB REMD of RAAE, SAAE, KAAE. Temperature distributions were chosen to give 15-25% exchange success, with actual temperatures reported in Table 2.S2. The TIP3P REMD simulation protocol was adopted from Okur et al.[34b] For REMD simulations of RAAE, SAAE, KAAE model peptides, backbone atoms were restrained with weak positional restraints (1.0 kcal/mol*Å) to the starting helical conformation, as discussed in Okur et al.[34b] There were two runs for tc5b and HP5F REMD simulations, starting from extended and folded conformations. We discarded first 25 ns of tc5b trajectories and 40 ns of HP5F trajectories to avoid initial structure bias. The error bars in these 2 cases were calculated from two runs. For other REMD simulations (Ala10, HP-1, RAAE) only 1 run was performed since the convergence time under these conditions had already been reported.[31, 34b, 85] In the case of SAAE and KAAE, we assumed that converged simulation time for side chain sampling should be comparable to that reported for RAAE[34b]. For Ala10, HP-1, SAAE, RAAE and KAAE REMD simulations, the first 10 ns of each run was discarded and error bars were estimated from first and second half of data.

**2.2.6.2 Data analysis.** PMFs were calculated based on the assumption of Boltzmann-weighted populations. Data were extracted from histograms of RMSD or distance, using $\Delta G = -RT\ ln(N_i/N_0)$ where $N_0$ was the population of the most populated bin and $N_i$ was the population of $i^{th}$ bin. Calculation of RMSD, DSSP[84] and φ/ψ values were done using the ptraj program in Amber10. For proteins taken from the Protein Data Bank, all ligands, water molecules and ions were removed and missing hydrogen atoms were added by tleap program in Amber10. Local secondary structure assignment for Ala10 was previously defined by Roe et al.[31] based on φ/ψ angle values (alpha (-70°/-25°), left (50°/30°), PP2 (-70°/150°), or extended (-150°/155°)). We retained this definition for HP-1 to be consistent with Ala10.

**2.2.6.3 Cluster analysis.** Cluster analysis was done by the Moil-View program[90] following the protocol described by Okur et al.[34b] We used a similarity cutoff of 2.5 Å for all backbone atoms of Ala10, trpzip2, and 3Ai3 and HP5F trajectories.

## 2.3 Results and Discussion

**2.3.1 Parameter fitting.** The 18 parameters were refit to minimize the objective function. As stated in Methods, we performed 5 runs for UOBYQA in which each run started from initial random parameters and each UOBYQA run converged at different local minima. We then performed 31 runs for parallel GA in which most of runs started with random parameters while some runs started by including in initial population GB-OBC parameters, GB-Neck parameters or parameter sets from UOBYQA runs. Although convergence among all GA runs would be ideal, this was not achieved in ~2500 generations. Each GA run had its own final objective function and optimized parameters. Lack of convergence could be due to the choice of GA parameters (population size, mutation rate…), or the attempt to explore too large a parameter space. Attempting to vary GA parameters such as population size and mutation rate were not successful in producing better objective function values. However, our aim was not to get the global minimum of the objective function, but to find a parameter set showing significant improvement compared to GB-OBC and GB-Neck models. Figure 2.S10 shows the objective function evolution during optimization. The best objective function was achieved through UOBYQA optimized parameters; these are denoted hereafter as GB-Neck2. The optimized values of the 18 refit parameters are given in Table 2.2. Objective function values for GB-OBC, GB-Neck and GB-Neck2 are provided in Table 2.1. The objective function of GB-Neck2 is 274, which is much smaller than 381 and 445 for GB-OBC and GB-Neck models, respectively.

It should be noted that although the scaling parameters $S_X$ were initially introduced to correct for overlap of van der Waals spheres and so might be expected to remain less than or equal to 1.0, there is no formal reason that they cannot be greater than 1.0, as pointed out by Hawkins et al.[21] Indeed, it can be argued that since the purpose of the majority of the parameters introduced into the GB formalism is to allow a better fit to higher levels of theory, the overall agreement of the model is more important than assigning a physical meaning to the parameters. When the $S_X$ values are considered free parameters it allows them to correct for other errors in the model, such as those introduced by the CFA.

**Table 2.2.** Optimized parameters for GB-Neck2 model.

| Parameter | Value | Parameter | Value | Parameter | Value |
|---|---|---|---|---|---|
| $S_H$ | 1.426 | $\alpha_H$ | 0.788 | $\alpha_N$ | 0.503 |
| $S_C$ | 1.059 | $\beta_H$ | 0.799 | $\beta_N$ | 0.317 |
| $S_N$ | 0.734 | $\gamma_H$ | 0.437 | $\gamma_N$ | 0.193 |
| $S_O$ | 1.061 | $\alpha_C$ | 0.734 | $\alpha_O$ | 0.868 |
| *offset* | 0.195 | $\beta_C$ | 0.506 | $\beta_O$ | 0.877 |
| $S_{neck}$ | 0.827 | $\gamma_C$ | 0.206 | $\gamma_O$ | 0.388 |

### 2.3.2 Comparison with PB solvation energies and effective radii

#### 2.3.2.1 Results on training sets

Table 2.1 shows the contributions to the objective function of the absolute solvation energies, relative solvation energies, and effective Born radii. For small systems like Ala10 or RAAE where most atoms are solvent-exposed, these pairwise GB models perform reasonably well,[31] and only modest improvement in these metrics is obtained with refitting. Particularly, *abs_e* and *rel_e* to PB calculation of GB-Neck2 for Ala10_set1 are 0.8 and 1.0 kcal/mol, compared to 1.1 and 1.6 kcal/mol of GB-OBC or 2.7 and 2.3 kcal/mol of GB-Neck, respectively. In larger systems like trpzip2, 3Ai3 or HP36, more substantial improvement was seen in GB-Neck2. For example, *abs_e* for 3Ai3 was reduced from 10.1 (GB-Neck) or 7.6 (GB-OBC) to 3.7 kcal/mol (GB-Neck2). For a given molecule, obtaining more accurate GB absolute solvation energies was easier than relative solvation energy. HP36, for instance, has *abs_e* of 4.3 kcal/mol (GB-Neck2), which is 24.1 kcal/mol lower than *abs_e* of GB-Neck (85.2 % reduction in error), while *rel_e* of GB-Neck2 for this training set is only improved by 0.3 kcal/mol (5.9 % reduction). This result suggests that refitting leads to improvements in systematic error of GB-Neck across all conformations (see Figure 2.1). GB-Neck2 also has better agreement with PB in calculating effective Born radii. The GB-Neck2 *eff_rad_rmsd* to 'perfect' radii of Ala10 (0.10 Å) was smaller than GB-OBC and GB-Neck (0.16 Å and 0.19 Å respectively). Once again, the improvement is more significant for larger systems. The *eff_rad_rmsd* for the protein HP1113 set is 1.47 Å for GB-Neck2 as compared to 1.82 Å for GB-OBC or 2.27 Å for GB-Neck.

Overall, there is improvement of absolute solvation energy, relative solvation energy and effective radii calculation for GB-Neck2 for all training sets as compared to GB-OBC and original GB-Neck model.

**2.3.2.2 Results on test sets.** In this section, we employ larger test sets to gauge the transferability of the new parameters. As stated in Methods, we designed two test categories: type I and II. Type I had a peptide / protein system that was used in training, but with many more conformations, while test set type II had entirely different molecules than those in the training sets.

**Comparison with PB solvation energy.** The *abs_e* and *rel_e* to PB data are presented in Table 2.3. The trend observed for the type I test sets is consistent with results for the training data, indicating that the structure variation was sufficient in the training data to permit application to more structure variety while retaining the improvement compared to the older GB models. For instance, *abs_e* for Ala10 test set from GB-OBC, GB-Neck and GB-Neck2 are 1.1, 2.2 and 1.0 kcal/mol respectively while *rel_e* are 0.7, 0.7 and 0.5 kcal/mol for GB-OBC, GB-Neck and GB-Neck2. For more complex molecules, the test set results closely match those from the training set: the *abs_e*, for example, of trpzip2 test set from GB-OBC, GB-Neck and GB-Neck2 are 9.2, 8.4 and 3.2 kcal/mol which are close to 9.5, 8.3, 2.8 kcal/mol for trpzip2 training set, respectively.

Results for type II test sets (Table 2.3) indicate that the improvements are transferable to independent systems, with lower abe_e and *rel_e* for GB-Neck2 as compared to GB-OBC and GB-Neck. There is little improvement for very small proteins like tc5b and DPDP. However, larger proteins show quite dramatic improvement. For example, *abs_e* of GB-Neck2 for the AIDS drug target HIV1-PR was 17.2 kcal/mol, eliminating 85% - 87% of the error as compared to GB-OBC (115.0 kcal/mol ) or GB-Neck (133.1 kcal/mol). Additionally, *rel_e* of GB-Neck2 for HIV1-PR was 16.8 kcal/mol, significantly improved as compared to 20.1 kcal/mol error with GB-OBC and GB-Neck. Since relative energies control the equilibrium populations, this improvement would be expected to have a significant impact on the ensemble sampled in MD simulations.

**Table 2.3.** *abs_e* and *rel_e* (kcal/mol) between each GB and PB calculation for type I and II test sets, shown for multiple GB models. Type II test sets are indicated in bold.

| (A) *abs_e* (kcal/mol) | GB-OBC | GB-Neck | GB-Neck2 | (B) *rel_e* (kcal/mol) | GB-OBC | GB-Neck | GB-Neck2 |
|---|---|---|---|---|---|---|---|
| Ala10 | 1.1 | 2.2 | 1.0 | Ala10 | 0.7 | 0.7 | 0.5 |
| Trpzip2 | 9.2 | 8.4 | 3.2 | Trpzip2 | 1.6 | 1.9 | 1.2 |
| 3Ai3 | 7.2 | 10.6 | 4.0 | 3Ai3 | 2.1 | 2.0 | 1.9 |
| RAAE | 1.3 | 1.6 | 1.4 | RAAE | 0.6 | 0.7 | 0.5 |
| HP36 | 21.3 | 29.7 | 6.6 | HP36 | 6.0 | 6.0 | 5.4 |
| **tc5b** | 7.4 | 13.4 | 5.3 | **tc5b** | 1.8 | 2.6 | 1.8 |
| **DPDP** | 3.4 | 12.7 | 3.6 | **DPDP** | 2.0 | 2.2 | 1.9 |
| **HIV1-PR** | 115.0 | 133.1 | 17.2 | **HIV1-PR** | 20.1 | 20.1 | 16.8 |
| **Lysozyme** | 72.2 | 88.4 | 13.1 | **Lysozyme** | 13.4 | 13.5 | 11.9 |

**Comparison with PB 'perfect' radii.** In order to test the transferability in improvement of effective Born radii from training to testing stage, we randomly extracted 100 tc5b structures having backbone RMSD to native structure smaller than 2.5 Å to compare effective radii from GB to PB. Since calculating PB 'perfect' radii for large proteins is computationally expensive,[16] we chose tc5b as a system large enough to have buried atoms and small enough to be computationally tractable. In addition, native-like structures were chosen to have a wide range of effective radii from atoms in molecule's surface to deeply buried atoms. The inverse of effective radii is used in calculating forces, thus it makes sense to compare inverse effective radii.[24] The RMSD between GB and PB inverse effective radii for GB-OBC, GB-Neck and GB-Neck2 are 0.068, 0.052 and 0.054 respectively. GB-Neck2 and GB-Neck have nearly the same RMSD and the performance of these models is somewhat better than GB-OBC.

Figure 2.1 shows 2D histograms for the TC5b set of inverse effective radii from GB models compared to inverse of 'perfect' radii derived from PB. The effective radii of buried atoms (lower left region) were still underestimated in GB-OBC while GB-Neck and GB-Neck2 had less degree of underestimation. However all three models seemed to overestimate effective radii of atoms near surface of the molecules (upper right region). GB-Neck2 is somewhat improved for effective radii calculation of atoms in the middle region of the plot in which the most populated bins lie close to the diagonal. For atoms in this region, GB-OBC and GB-Neck

tend to overestimate effective radii, perhaps leading to the dramatic improvement in systematic error with GB-Neck2 seen in Tables 2.1 and 2.3.

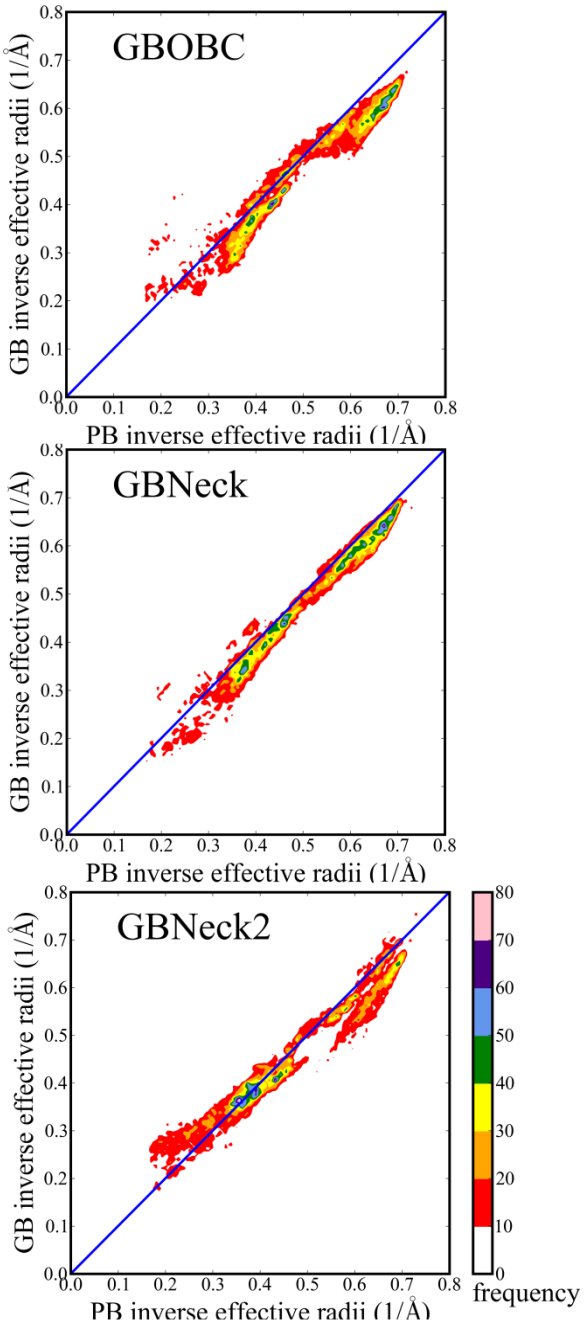


**Figure 2.1**. 2D histograms of inverse effective Born radii of each GB model versus PB 'perfect' radii for tc5b. Perfect agreement is shown by the diagonal line. The color indicates the frequency (number of atoms) in each bin.

In summary, the results from the test sets confirm the improvement from the new parameters is transferable from a set of structures used for training to different set of structures not used in training as well as to entirely different proteins.

### 2.3.3 Comparison with explicit water MD: hydrogen bonds, salt bridges and secondary structure

A particular goal of this work was to reduce the errors in secondary structure bias and salt bridge strength previously reported for GB models compared to results from explicit water models such as TIP3P. The results presented in previous sections showed significant improvement of GB-Neck2 model in calculating solvation energy and effective Born radii when using PB as a benchmark. It is of interest to determine if better match to PB also results in improved correspondence of GB with reference simulations in explicit water, which tend to be much more computationally demanding. We hypothesized that fitting to the more accurate PB continuum water model could help improve agreement between GB and explicit water.

We next tested GB-Neck2 to see if this improvement could translate to improvement in balancing secondary structure populations and improving salt bridge strength. The comparison between GB and TIP3P simulations, however, does not solely depend on the GB model. Firstly, performance of a GB model is heavily affected by the intrinsic Born radius set that defines the boundary between solute and solvent. Secondly, GB only calculates the polar part of solvation free energy and simulation results also depend on the accuracy of nonpolar solvation contributions, such as cavity formation and van der Waals interactions with solvent. Since the simple SASA-based nonpolar solvation approach currently available in Amber has known limitations,[9, 12-13] the main focus of this work is improved agreement in polar solvation free energy. However, in order to roughly estimate the affect that including this term has on results, simulations were performed with and without the commonly used surface area based nonpolar term for the HP5F and tc5b systems.

#### 2.3.3.1 Strength of hydrogen bonds and salt bridges

The salt bridge is formed by oppositely charged side chains of Arg (or Lys) and Glu (or Asp). Conventionally, the ion pair interaction could be adjusted by changing the radii of $H^{N+}$ of Arg. For example, Geney et al.[30b] and Shang et al.[56] empirically decreased the radius of $H^{N+}$ of Arg from 1.3 Å to 1.1 Å to match the salt bridge PMF from GB to that from TIP3P simulation.

However, we recognize that the radii (and hence desolvation penalty) of the carboxyl oxygen atoms can also be modified to change the balance of desolvation and Coulombic contributions. One approach to determine which group to adjust is to examine carboxyl interactions in the absence of a positively charged partner. We first chose a simple peptide system, SAAE, to investigate H-bond strength between the ionized side chain of Glu with the side chain of Ser. Original mbondi2 radii[23] were used for all GB models. Figure 2.2 shows the distance PMFs of Hγ (SER) and Cδ (Glu) for TIP3P, GB-OBC, GB-Neck, and GB-Neck2 models. The H-bond is thermodynamically unstable in all cases, meaning that the H-bonded distance is a local and not the global free energy minimum. All of these GB models fail to reproduce the solvent-separated local minimum near 5 Å; such behavior is expected for continuum models. The H-bond in GB-Neck2 is 0.7 kcal/mol stronger than in TIP3P, while GB-Neck and GB-OBC H-bond strength are comparable to TIP3P. We empirically decreased the carboxyl oxygen radii from 1.5 Å to 1.4 Å to reproduce the profile obtained in TIP3P. The modified carboxyl oxygen radii should be applied to charged carboxyl groups in Asp and Glu sidechains as well as C-terminal residues.

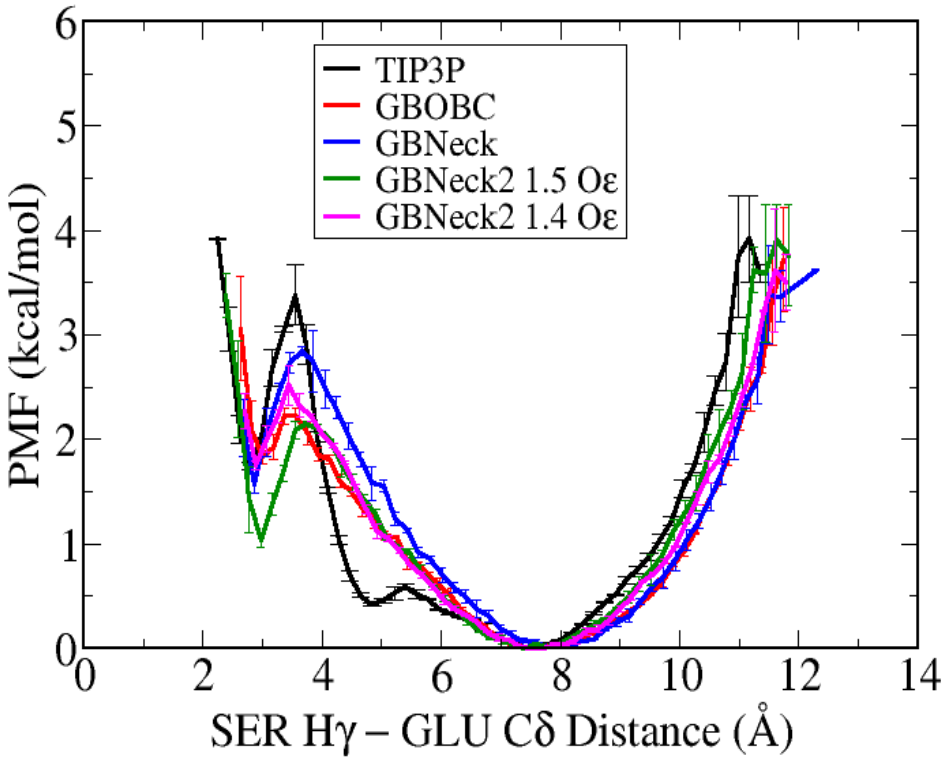


**Figure 2.2.** PMFs for side chain H-bond formation in the SAAE model peptide for various solvent models. The 2 GB-Neck2 curves used different Born radii for the Glu side chain carboxyl oxygen atoms, indicated in Å in the legend.

Having carboxyl oxygen radii of Glu sidechain adjusted, we next investigated the salt bridge formed by Arg and Glu. We originally included a set of RAAE structures in our training set (Table 2.1) for explicitly training the salt bridge. The fitting resulted in modest improvement in solvation energy calculation for these ion pairs as compared to GB-OBC and GB-Neck. We therefore expected improved agreement between salt bridge PMFs from GB-Neck2 and TIP3P REMD simulations for this RAAE system. Salt bridge PMFs from REMD runs for different GB models and TIP3P are shown in Figure 2.3A, with variation of the Arg $H^{N+}$ Born radii in the mbondi2 set; PMFs for GB-OBC, GB-Neck, GB-OBC 1.1 Å $H^{N+}$, GB-Neck2 with 1.4 Å Oε and GB-Neck2 with 1.4 Å Oε + 1.17 Å $H^{N+}$ are shown. All GB profiles have a global minimum slightly shifted from the one in TIP3P due to the difference in salt bridge geometry between GB and TIP3P, discussed in more detail by Okur et al.[34b] With standard mbondi2 radii (1.5 Å Oε of Glu and 1.3 Å $H^{N+}$ of Arg) the PMF indicates salt bridges from GB-Neck2 simulation are ~3.5 kcal/mol stronger than in TIP3P, significantly worse than the 1.0 kcal/mol and 2.0 kcal/mol stronger with GB-OBC and GB-Neck, respectively. This implies that fitting to PB solvation energies did not help improve salt bridge profile (RMSD between GB and PB absolute energies for RAAE test set (Table 2.3) is 1.4 kcal/mol). We thus hypothesize that PB with mbondi2 radii may also have too strong salt bridge compared to TIP3P, as indicated by Shang et al.[56] With new carboxyl oxygen radii (1.4 Å) fit to SAAE PMFs and standard $H^{N+}$ radii (1.3 Å), the salt bridge with GB-Neck2 is still ~2.0 kcal/mol stronger than in TIP3P (Figure 2.S11). Thus, radii of $H^{N+}$ of Arg were empirically reduced from 1.3 Å to 1.17 Å to match the TIP3P PMF curve (Figure 2.3). The PMF from GB-Neck2 with 1.17 $H^{N+}$ (Arg) also matches well to that from GB-OBC with modified $H^{N+}$ radii as reported in Shang et al.[56], suggesting that modification of this radius is a general way to improve salt bridges in GB models. The physical justification for adjusting these radii is discussed in detail by Geney et al.[30b]

We next addressed whether primary amines (Lys and N-term) needed comparable corrections to Arg. Figure 2.3B shows the PMFs for KAAE from GB-OBC, GB-Neck and GB-Neck2 (all with mbondi2) and GB-Neck2 with mbondi3 radii. As discussed above, none of the GB models reproduce the solvent-separated minimum seen with explicit water. In GB-Neck2 with mbondi2 radii, the salt bridge was ~1.0 kcal/mol stronger than TIP3P while the GB-OBC salt bridge was ~0.5 kcal/mol stronger. In contrast, the salt bridge with GB-Neck mbondi2 was ~0.5 kcal/mol weaker than in TIP3P. GB-Neck2 with mbondi2 and modified carboxyl oxygen

showed near-quantitative match to TIP3P PMF, suggesting that our caboxyl changes were sufficient and no adjustment of radii is needed for $H^{N+}$ of Lys side chain or N-terminal amines.

The new radii set with modified carboxyl oxygen and Arg $H^{N+}$ is denoted mbondi3 (Table 2.S1). Overall, mbondi3 appears to be the best radii set for use with GB-Neck2 in reproducing TIP3P PMFs of salt bridge interactions. In the remainder of this work manuscript, all simulations of GB-Neck2 used mbondi3 intrinsic Born radii unless noted otherwise.

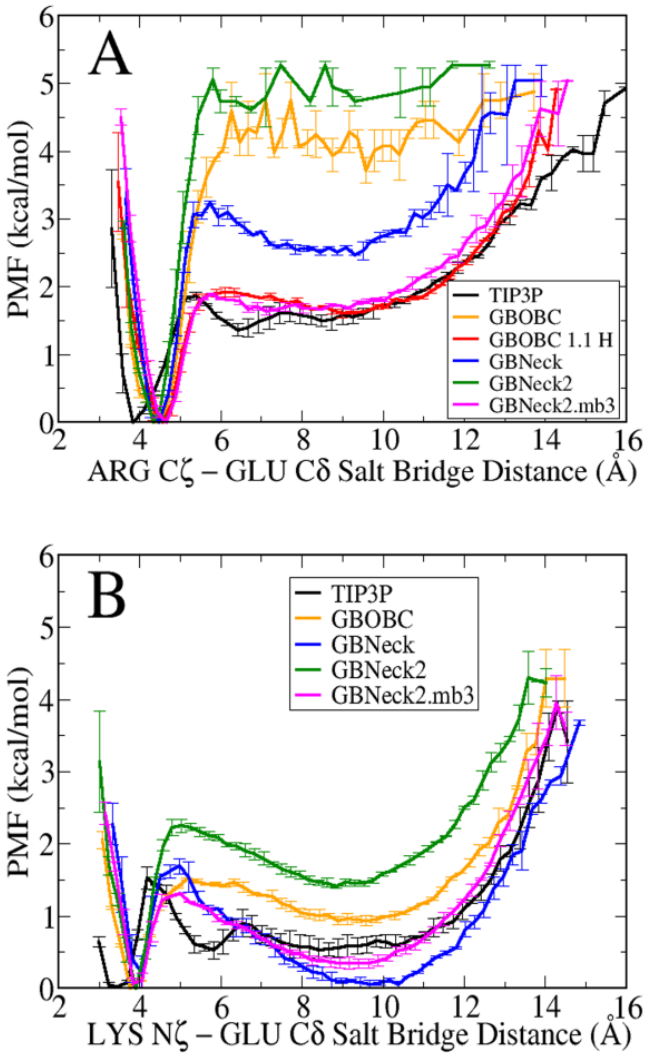


**Figure 2.3.** Salt bridge PMFs for various solvent models. Panel A shows the PMF profiles for RAAE (Arg salt bridge) while panel B shows PMFs for KAAE (Lys salt bridge). GB-OBC, GB-

Neck and GB-Neck2 used original mbondi2 radii set while GB-OBC 1.1 $H^{N+}$ used mbondi2 with modified $H^{N+}$(Arg). GB-Neck2.mb3 used the optimized radii set denoted mbondi3 (Table 2.S1).

### 2.3.3.2 Evaluating α-helical bias

**Ala10 Model Peptide.** Roe et al.[31] showed that the ability of a GB model to reproduce PB solvation energies for Ala10 was well correlated with the extent of helical bias obtained in simulations compared to TIP3P simulation. We therefore hypothesize that our new GB model, with better agreement to PB, should also better reproduce secondary structure preferences as compared to TIP3P. Roe et al.[31] quantified the accuracy by comparing DSSP and local conformational propensity between GB and TIP3P simulations. We repeated these analyses for our GB-Neck2 model, using GB-OBC and GB-Neck results as controls (Figure 2.4, with numerical data provided in Table 2.S3).

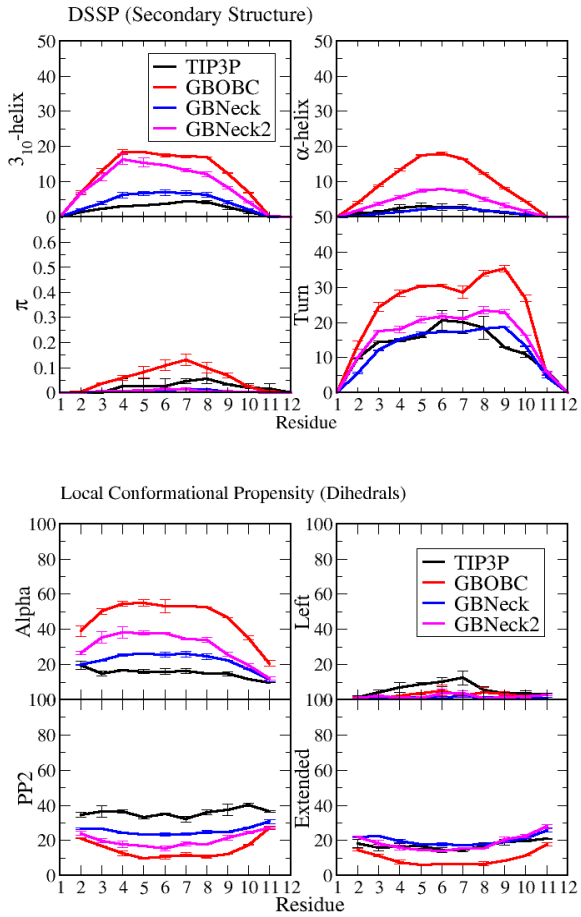


**Figure 2.4**. Secondary structure (upper) and local conformational propensities (lower) for each residue of Ala10 at 300K from REMD simulations using different solvent models.

GB-Neck2 has reduced alpha and turn content as compared to GB-OBC (4.4 % vs. 10.1 % in OBC for alpha content; 16.2 % vs. 25.5 % in OBC for turn content). However, the original GB-Neck still has somewhat better agreement to TIP3P data (1.4 % vs. 2.5 % in TIP3P for alpha content; 4.6 % vs. 2.9 % in TIP3P for turn content). Although 3-10 helix content was reduced for GB-Neck2, the population is still somewhat too large compared (9.3 % in GB-Neck2 and 12.7 % in OBC vs. 2.9 % in TIP3P). GB-Neck2 also has higher preference for residue to sample the helical region of the Ramachandran map (30.3 % in GB-Neck2 and 22.6 % in GB-Neck vs. 6.2 % in TIP3P). Although GB-Neck2 better reproduced absolute and relative solvation energies for Ala10 training set and Ala10 test set than GB-Neck, this improvement seems not to transfer to better agreement with TIP3P simulation. There might be several reasons for this. First, the mbondi2 radii set (mbondi3 is the same as mbondi2 for systems that do not have Arg, Glu, Asp or charged C-termini) was not specifically optimized for use with GB-Neck, and the improved agreement to TIP3P for this combination may be fortuitous cancellation of error. This same cancellation of error may make the performance of GB-Neck better than PB in this particular case; however it is difficult to get converged REMD data when using PB solvation, and such calculations are out of the scope of the present work. In addition, the small improvement in energy compared to PB may not be enough to improve structure results compared to TIP3P simulation for this system. This seems reasonable since we have seen significant improvement for larger systems like HP5F or tc5b, which will be shown below.

**HP-1 Model Peptide.** Because Ala10 structures were used in training GB-Neck2, we repeated the same analyses as we did for Ala10 but for a different peptide system (HP-1) to confirm the results in balancing secondary structure (**Figure 2.5**). Furthermore, unlike Ala10, HP-1 is known to adopt modest helical content in solution.[85] Similar trends to Ala10 were observed in DSSP data and local alpha content (Table 2.S4**).** Particularly, the alpha content from GB-Neck was slightly smaller than TIP3P content while GB-Neck2 alpha content was somewhat larger (23.8 % in GB-Neck2 vs. 18.9 % in GB-Neck vs. 21.6 % in TIP3P). GB-OBC had too much alpha content (43.9 %). Although all GB models had close average turn content compared to TIP3P, GB-Neck and GB-Neck2 had better agreement as indicated by DSSP. This shows that performance of GB-Neck and GB-Neck2 on alpha content is somewhat system dependent, likely

due to the role of side chain interactions in helix formation of HP-1.[85] However, the trend remains that GB-Neck tends to destabilize alpha conformations, as demonstrated above and as previously reported by Dill et al.[49] and Roe et al.[31] Overall, the good performance of GB-Neck2 in balancing secondary structure can be transferred from training system (Ala10) to testing system (HP-1).

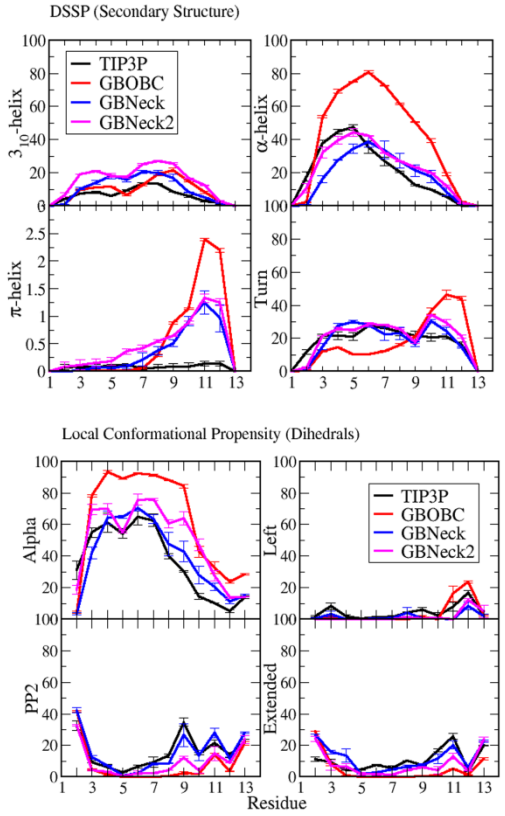


**Figure 2.5**. Secondary structure (upper) and local conformational propensities (lower) at 300K for each residue of HP-1, obtained from REMD simulations using different solvent models.

**2.3.4 Folding of HP5F and tc5b: Comparison with experimental melting temperature.** The above GB simulation results were compared to TIP3P simulations using the same protein force field. However, one of the main purposes of improving a GB model is to get closer agreement between computational and experimental data, particularly for simulations that are currently difficult or intractable in explicit water. However, such comparisons are more complex than the comparison between GB and TIP3P simulation because they also depend on the protein force field used. Deviations from experiment may not be a result of weakness in the GB part of the model, and accurate reproduction of experimental data could arise from fortuitous cancellation of

error and may not provide proof of an accurate solvent model. Nonetheless, the comparison to experiment provides a useful measure of the quality one might expect from this particular combination. For the purpose of this testing, we used the combination of the GB-Neck2, mbondi3 radii and the ff99SB force field.[57] This widely adopted force field was used since it has been shown to well balance secondary structure.[57, 60c, 91]

We compared equilibrium thermal stability between different GB models and experiment (NMR or CD) for HP5F[58] and tc5b,[59] which adopt different structure motifs (hairpin and helix-turn-PPII).[5a, 30b, 92] Simulations with GB-Neck2 were also repeated including a SASA-based nonpolar solvation term in order to ascertain its impact on results.

Figure 2.6A shows the simulated melting curves for HP5F for GB-OBC, GB-Neck, GB-Neck2 and GB-Neck2 SASA models compared to experiment data. The melting temperature and fold population of GB-Neck2 at 298 K (317K and 74%, respectively) are in excellent agreement with experimentally determined values (326 K and 82%).[58] For the tc5b mini-protein, the melting curves for GB-Neck2 and NMR and CD experiments[59] are shown in **Figure 2.6B**. GB-Neck2 predicts a melting temperature of 302 K, which is again close to the experimental value of 315 K[59] and to the reported value of 321 K from TIP3P REMD simulation[93] with ff99SB force field. The excellent agreement between GB-Neck2 simulation and experiment is promising since several groups reported significantly elevated simulated melting temperatures for tc5b.[94] Pitera et al.[94a] reported a melting temperature of ~400K from REMD simulation of GB-HCT model + ff94 force field. Zhou et al.[94b] also obtained a melting temperature above 400K when using TIP3P model + OPLS-AA force field. Compared to GB-Neck2 simulations, GB-OBC and GB-Neck significantly underestimate melting temperature for both testing systems (GB-OBC: ~307 K and ~264 K; GB-Neck: <275K and ~290K for HF5F and tc5b respectively). GB-Neck especially destabilizes the native hairpin even at very low temperature.

GB-Neck2 runs with and without the nonpolar term both produce reasonable estimations of melting points for HP5F and tc5b (317 K and 335 K for HP5F; 302 K and 324 K for tc5b for simulations with and without nonpolar term respectively). Inclusion of the SASA-based nonpolar term provides small increases in stability but does not dramatically impact the results for these systems. It is likely that use of a better nonpolar model (such as that in AGBNP2[28]) could improve results even further, however that is beyond the scope of the current work, which focuses on the polar component of solvation.

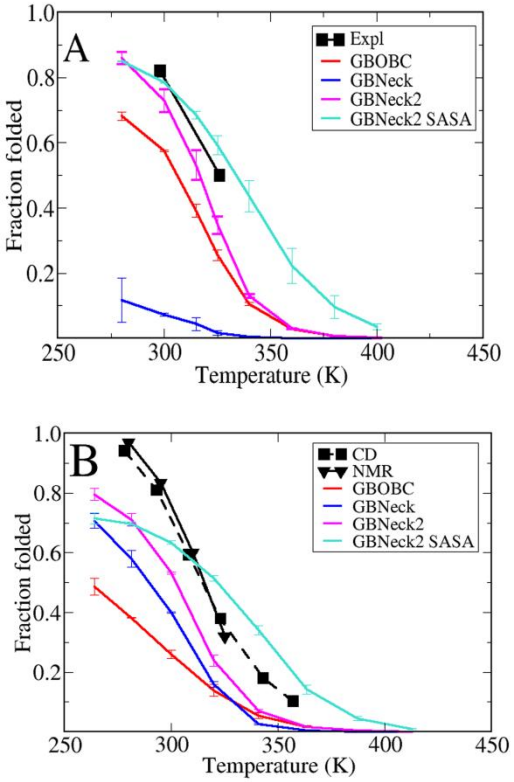


**Figure 2.6**. Panel A and B show the thermal stability profiles for the HP5F and tc5b respectively in GB-OBC, GB-Neck and GB-Neck2 (with and without SASA) REMD simulations, compared to experimental data.[58-59]

## 2.4 Conclusion

Pairwise GB solvation models remain desirable due to their high computational efficiency, but many weaknesses have been reported. We propose a new parameter set for the GB-Neck model, obtained by making several key parameters that relate to interstitial cavities dependent on chemical element. Adding more parameters called for use of a much larger training set than employed in the past, therefore we developed conformation libraries containing thousands of structures for peptide and protein sequences of various lengths and structure propensities. Our objective function for training included absolute and relative solvation free energies compared to PB, as well as accuracy of effective Born radii of the atoms. Final empirical adjustments were made to some of the intrinsic radii to improve agreement with

explicit solvent simulations. These modifications help GB-Neck reproduce the H-bond and salt bridge PMFs of TIP3P simulations. The new GB-Neck2 model not only shows better results for the training systems, but for a variety of tests systems that measure solvation free energy, secondary structure propensity and even thermal stability profiles compared to experimental data. Thus the combination of GB-Neck2 model, radii set, force field used here is recommended for future study of peptide or protein simulations.

Our GB-Neck2 model shows significant improvement in solvation energy and effective radii calculation as compared to GB-OBC and GB-Neck. This model, however, is still based on the CFA integral calculation which has been shown to overestimate effective radii,[17] compared to much slower numerical models such as GBMV[95] or GB-R6[17] using non-CFA integrals. Through parameter fitting, our approach thus has attempted to empirically compensate for the CFA as much as possible. Onufriev et al.[27] recently developed an analytical form of GB-R6 (named AR6) but the resulting accuracy was substantially decreased from the numerical form (NR6), and performed worse than GB-Neck2 on our training and test sets.[96] We believe that our strategy in fitting parameters, as well as use of the training and test sets we have developed, could help to improve the performance of AR6 and future solvation models.

Our results also show that despite not including a nonpolar term, GBNeck2 is still able to improve agreement to TIP3P as well as experiment, and it is likely that further improvement will be seen with the addition of a more accurate term for nonpolar solvation free energy.

Future work will include optimization of additional parameters for nucleic acid simulations.[97]

**Appendix 2. Supporting document**

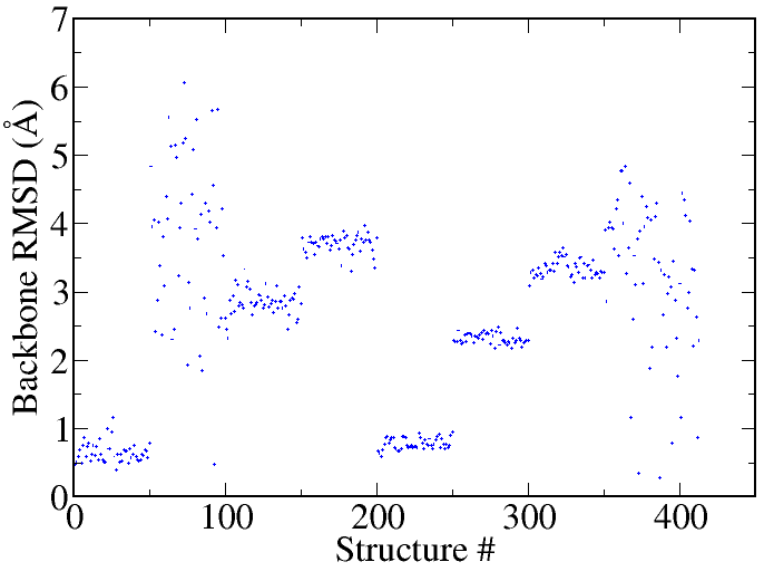


**Figure 2.S1**. Backbone RMSD (Å) to native trpzip2 for structures in trpzip2 set used for training GBNeck2 parameters. Residues 3-12 were used to calculate RMSD to avoid the flexibility of termini.

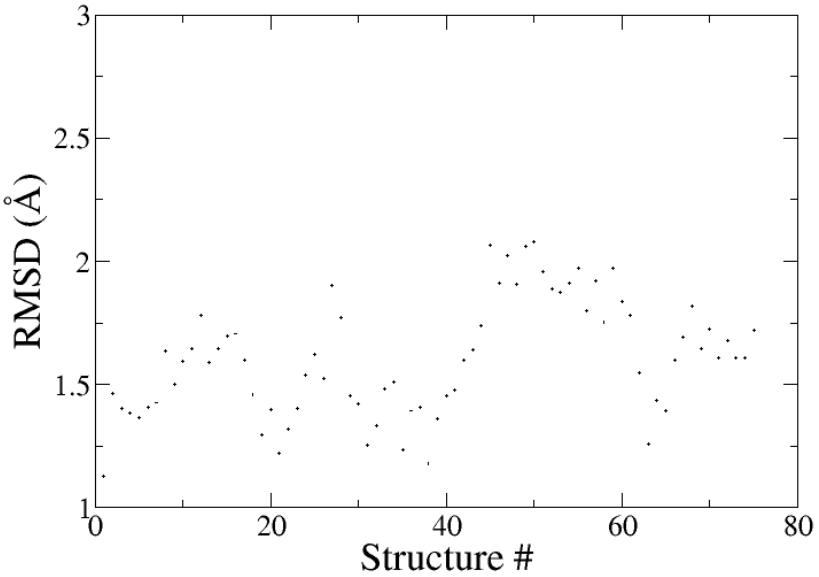

**Figure 2.S2**. Backbone RMSD (Å) to X-ray structure of 75 HP36 structures used in training GBNeck2 parameters. Residues 3 to 34 were used to calculate RMSD to avoid the flexibility of termini.

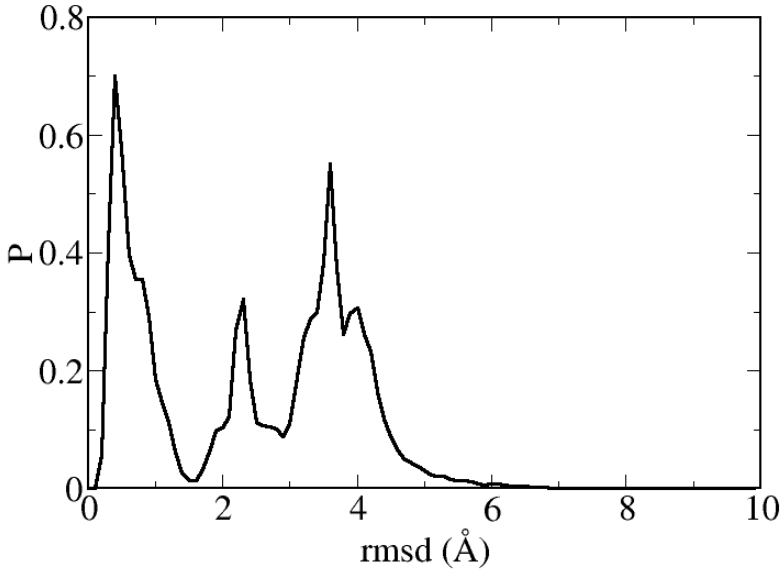


**Figure 2.S3**. Backbone RMSD histogram of trpzip2 structures used in testing GB and PB solvation energies. RMSD histogram is shown instead of RMSD plot for trpzip2 due to the large number of structures.

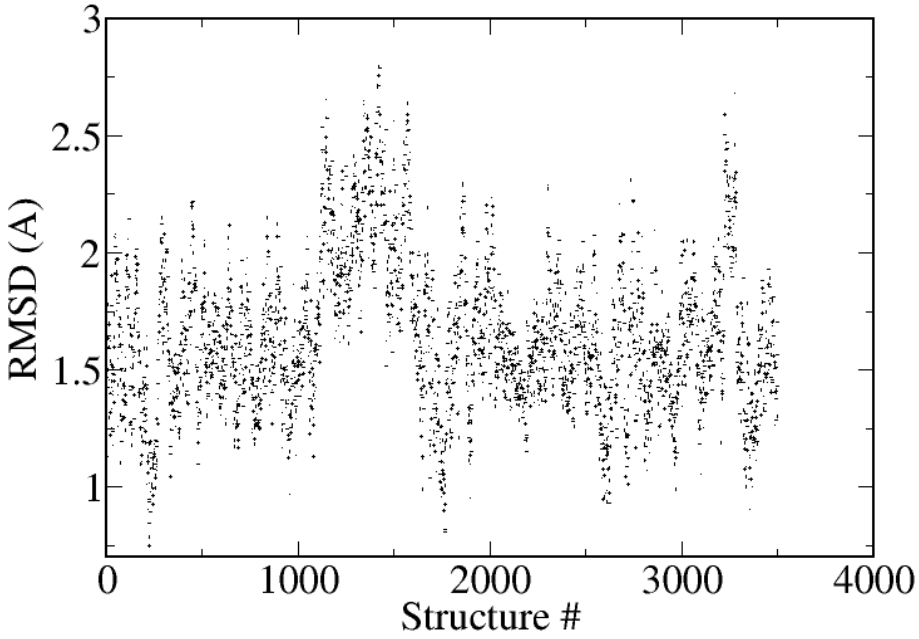

**Figure 2.S4**. Backbone RMSD (Å) to X-ray structure of 3500 HP36 structures used in testing GBNeck2 parameters. Residues 3 to 34 were used to calculate RMSD to avoid the flexibility of termini.

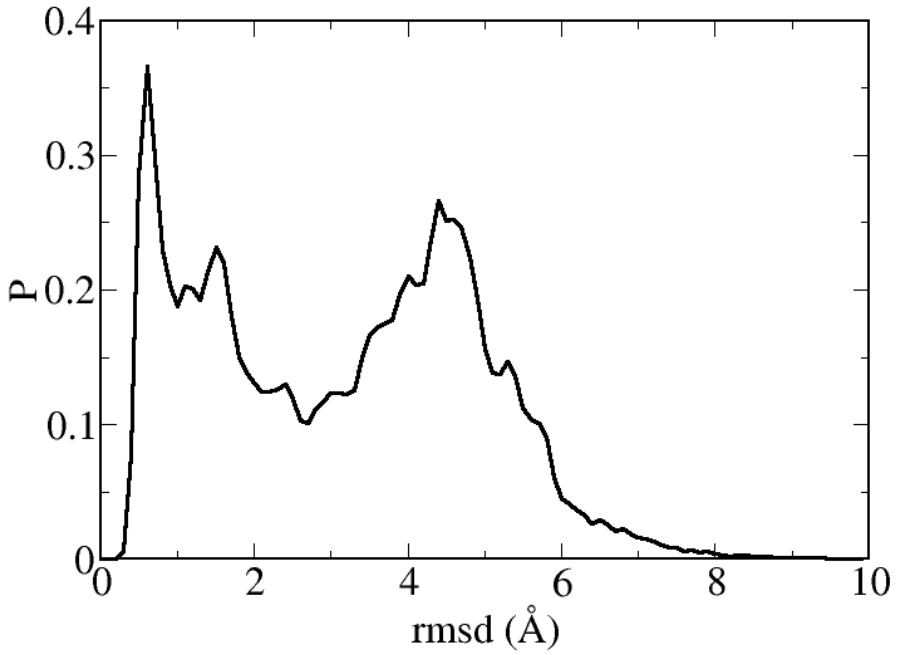


**Figure 2.S5**. Backbone RMSD (Å) histogram of Tc5b structures used in testing GB and PB solvation energies. Residues 3 to 18 were used to calculate RMSD to avoid the flexibility of termini.

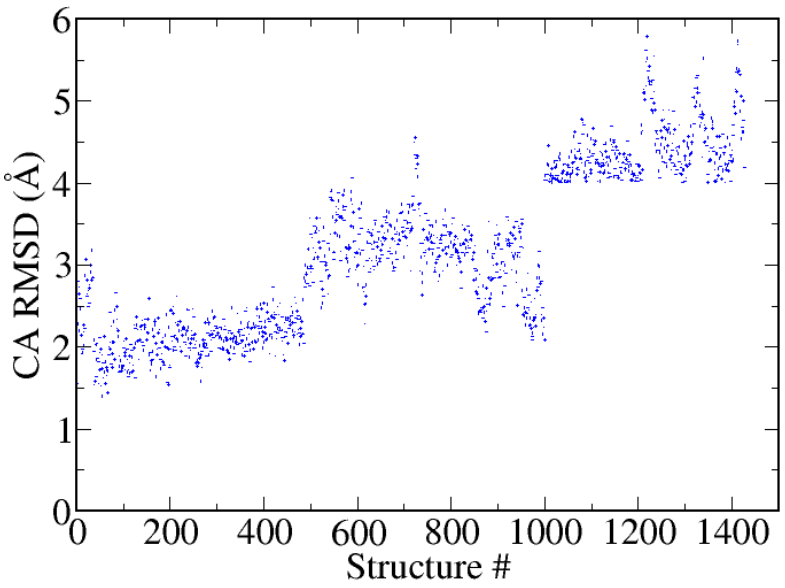


**Figure 2.S6**. Cα RMSD (Å) to closed X-ray structure for HIV-PR test set.

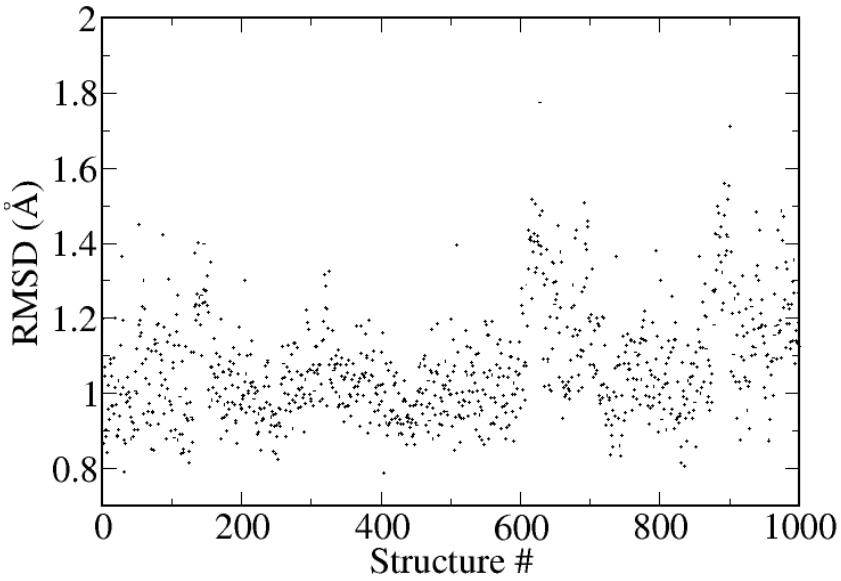


**Figure 2.S7**. Backbone RMSD (Å) to native lysozyme structure (PBD ID: 1IEE) of 1000 lysozyme structures used for testing GBNeck2 parameters. Residues 4-125 were used to calculate RMSD to avoid the flexibility of termini.

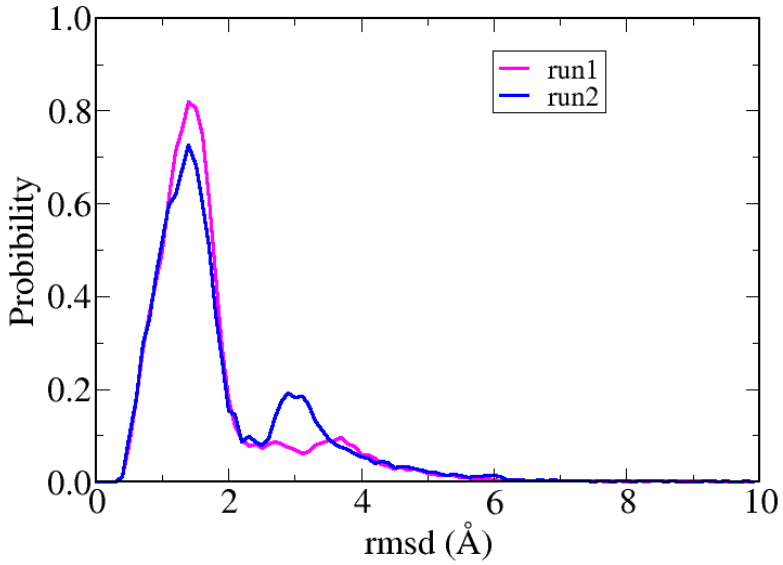


**Figure 2.S8**. Histogram of backbone RMSD (Å) of HP5F from 2 GBNeck2 simulation runs. 300K trajectories were used for calculating RMSD. 2 Å minimum separated folded and unfolded region.

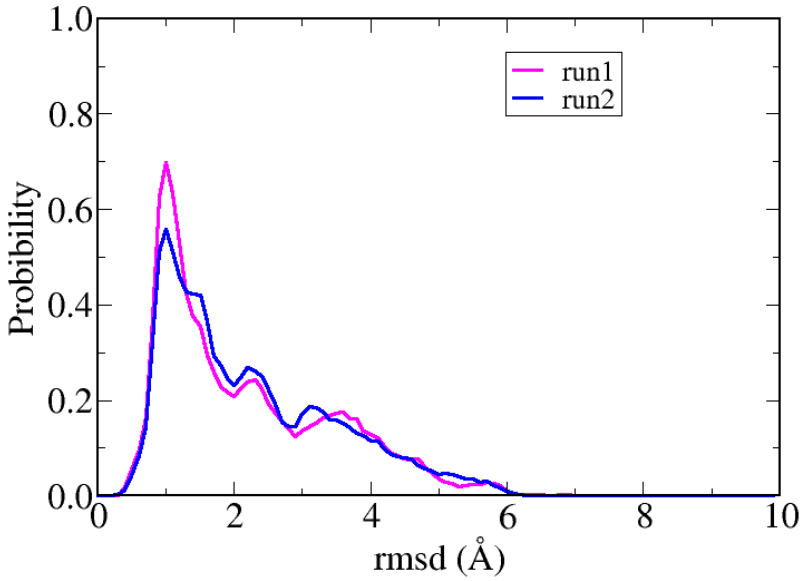


**Figure 2.S9**. Histogram of RMSD (Å) of TC5b from 2 GBNeck2 simulation runs. 300K trajectories were used for calculating RMSD. 2 Å minimum separated folded and unfolded region.

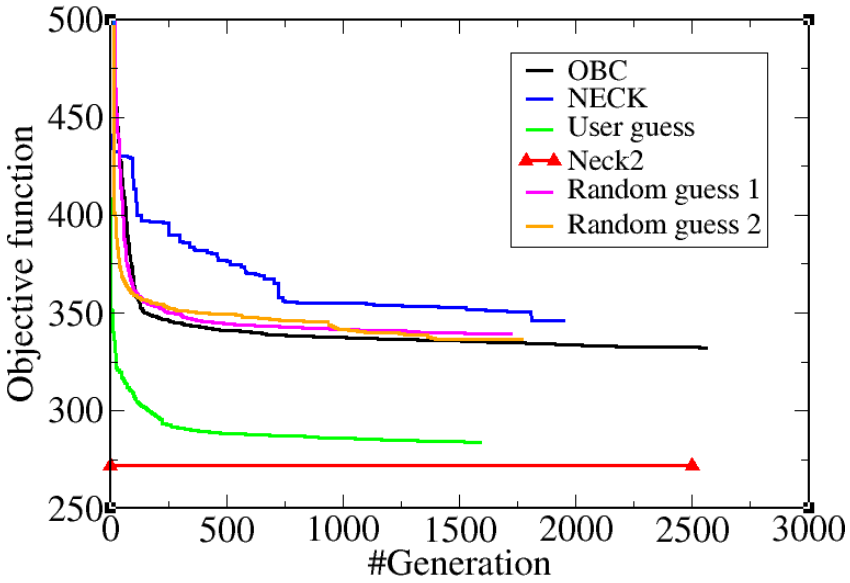


**Figure 2.S10.** Objective function value versus the number of generations in for several Genetic Algorithm (GA) runs. Because there is large number of runs, we are showing only 6 runs in Figure 2.S10 in which 2 runs included GBNeck and GBOBC parameters (named as "Neck" and "OBC" runs), 2 runs had random initial populations (names as "Random guess 1" and "Random guess 2" runs) and 2 runs included parameters from UOBYQA runs (named as "User guess" and "Neck2" runs). For "Random guess 1", "Random guess 2" and "OBC" runs, objective functions were reduced dramatically after ~200 generation while for "Neck" run; objective function was reduced significantly after ~750 generations. All of objective functions above were then reduced insignificantly during generation from 750 to 2500. This means that those GA runs stuck in the local minima. Per "Neck2" run, objective function was not reduced during optimization; this means that GA run did not help objective function escape from local minima found by UOBYQA. This could be due to that our choice of GA parameter was not really good or due to GA method itself. Parameters from "Neck2" run, however, were still chosen for our final working parameter because they had lowest objective function among all GA runs. The result indicates that UOBYQA might be a good optimization program for this kind of parameter fitting.

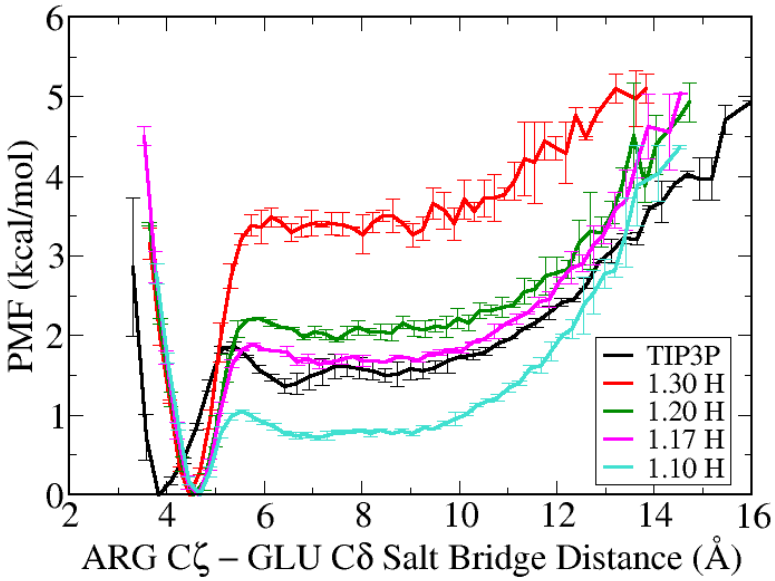


**Figure 2.S11.** Salt bridge PMFs for various GBNeck2 simulations compared to TIP3P data. We used new Oε (Glu) radius (1.4 Å) with different $H^{N+}$(Arg) radii (1.3, 1.2, 1.7 and 1.1 Å) . The PMF from GBNeck2 run using $H^{N+}$ radius of 1.17 Å shows the best match to TIP3P PMF.

**Table 2.S0.1** HP1113 set, having 6 large proteins for training GBneck2 effective radii

| **Family** | **PDB ID** | **Number of residues** |
|---|---|---|
| HIV-1 protease | 1TSU | 198 |
| Protein A | 1BDD | 60 |
| Ubiquitin | 1UBQ | 76 |
| Fatty acid-binding protein | 1AEL | 131 |
| FK506 binding protein | 1FKG | 107 |
| B1 domain of protein G | 3GB1 | 56 |

**Table 2.S0.2** Summary for test sets used for comparing GB and PB solvation energies. Type II test sets are indicated in bold.

| Test set | Number of structures | Number of residues | Net charge (C) |
|---|---|---|---|
| Ala10 | 50000 | 10 | 0 |
| Trpzip2 | 80000 | 12 | +2 |
| 3Ai3 | 49000 | 19 | +4 |
| RAAE | 50000 | 4 | 0 |
| HP36 | 3500 | 36 | +3 |
| tc5b | 103000 | 20 | +1 |
| DPDP | 50000 | 20 | +2 |
| HIV1-PR | 1427 | 198 | +8 |
| Lysozyme | 1000 | 129 | +8 |

**Table 2.S1**. New radii set mbondi3 for GB-Neck2 compared to mbondi2 radii set. H(X) means H bound to X atom. H($N^+$, Y) means $H^{N+}$ of amino acid Y (Y=Arg, Lys). O($COO^-$; Glu, Asp or C-terminal) is Oxygen of charged carboxyl group of Glu, Asp or C-terminal. The differences between mbondi3 and mbondi2 are bold.

| Atom | mbondi2 | mbondi3 |
|---|---|---|
| H(C) | 1.2 | 1.2 |
| H(N) | 1.3 | 1.3 |
| H($N^+$, Arg) | 1.3 | 1.17 |
| H($N^+$, Lys or N-terminal) | 1.3 | 1.3 |
| H(O) | 1.2 | 1.2 |
| H(S) | 1.2 | 1.2 |
| C | 1.7 | 1.7 |
| N | 1.55 | 1.55 |
| O | 1.5 | 1.5 |
| O($COO^-$; Glu, Asp or C-terminal) | 1.5 | 1.4 |
| S | 1.8 | 1.8 |

**Table 2.S2**. Temperatures (K) for each peptide system from GB and TIP3P REMD simulations

| SAAE GB | SAAE TIP3P | RAAE GB | KAAE GB | KAAE TIP3P | Ala10 GB | HP-1 GB | HP5F GB | TC5b GB | |
|---|---|---|---|---|---|---|---|---|---|
| 256.5 | 291.4 | 300.0 | 261.2 | 296.3 | 246.2 | 277.4 | 280.0 | 264.0 | |
| 300.0 | 300.0 | 348.7 | 300.0 | 300.0 | 271.8 | 300.0 | 300.0 | 281.4 | |
| 350.9 | 308.8 | 405.3 | 344.6 | 303.8 | 300.0 | 324.4 | 315.0 | 300.0 | |
| 410.5 | 318.0 | 471.0 | 395.9 | 307.6 | 331.2 | 350.8 | 325.0 | 319.8 | |
| 480.2 | 327.3 | 547.5 | 454.7 | 311.5 | 365.5 | 379.4 | 340.0 | 340.9 | |
| 561.7 | 337.0 | 636.3 | 522.3 | 315.5 | 403.5 | 410.3 | 360.0 | 363.3 | |
| | 346.9 | | | 319.5 | 445.4 | 443.7 | 380.0 | 387.3 | |
| | 357.2 | | | 323.5 | 491.6 | 479.8 | 400.0 | 412.9 | |
| | 367.7 | | | 327.6 | | | | | |
| | 378.5 | | | 331.7 | | | | | |
| | 389.7 | | | 335.9 | | | | | |
| | 401.2 | | | 340.2 | | | | | |
| | 413.0 | | | 344.5 | | | | | |
| | 425.2 | | | 348.9 | | | | | |
| | 437.8 | | | 353.3 | | | | | |
| | 450.7 | | | 357.7 | | | | | |
| | 464.0 | | | 362.3 | | | | | |
| | 477.6 | | | 366.8 | | | | | |
| | 491.7 | | | 371.5 | | | | | |
| | 506.2 | | | 376.2 | | | | | |
| | 521.2 | | | 380.9 | | | | | |
| | 536.5 | | | 385.8 | | | | | |
| | 552.4 | | | 390.6 | | | | | |
| | 568.6 | | | 395.6 | | | | | |
| | 585.4 | | | 400.6 | | | | | |
| | 602.7 | | | 405.7 | | | | | |
| | 620.5 | | | 410.8 | | | | | |
| | 638.8 | | | 416.0 | | | | | |
| | 657.6 | | | 421.3 | | | | | |
| | 677.0 | | | 426.6 | | | | | |
| | 697.0 | | | 432.0 | | | | | |
| | 717.5 | | | 437.4 | | | | | |

**Table 2.S3.** Average Percent Secondary Structures and Local Conformational Propensities from Ala10 REMD Simulations

| | TIP3P | GB-OBC | GB-Neck | GB-Neck2 |
|---|---|---|---|---|
| (A) DSSP (Secondary Structure) | | | | |
| 3-10 helix | 2.89 ± 0.06 | 12.66 ± 0.07 | 4.64 ± 0.09 | 9.33 ± 0.23 |
| alpha-helix | 2.45 ± 0.63 | 10.06 ± 0.08 | 1.37 ± 0.01 | 4.38 ± 0.07 |
| pi-helix | 0.01 ± 0.01 | 0.09 ± 0.02 | 0.01 ± 0.01 | 0.01 ± 0.00 |
| turn | 14.26 ± 0.18 | 25.54 ± 0.09 | 14.21 ± 0.30 | 16.20 ± 0.44 |
| (B) Local Conformational Propensity (Backbone Dihedrals) | | | | |
| alpha | 16.20 ± 0.33 | 45.85 ± 0.20 | 22.63 ± 0.15 | 30.26 ± 0.02 |
| left | 6.00 ± 0.28 | 2.58 ± 0.03 | 1.29 ± 0.04 | 1.99 ± 0.23 |
| PP2 | 34.65 ± 0.29 | 15.14 ± 0.09 | 25.45 ± 0.04 | 20.35 ± 0.30 |
| extended | 17.61 ± 0.38 | 9.87 ± 0.10 | 19.83 ± 0.15 | 18.79 ± 0.05 |

**Table 2.S4.** Average Percent Secondary Structures and Local Conformational Propensities from HP-1 REMD Simulations

| | TIP3P | GB-OBC | GB-Neck | GB-Neck2 |
|---|---|---|---|---|
| (A) DSSP (Secondary Structure) | | | | |
| 3-10 helix | 6.76 ± 0.40 | 9.93 ± 0.05 | 11.04 ± 0.21 | 16.05 ± 0.14 |
| alpha-helix | 21.59 ± 0.41 | 43.90 ± 0.05 | 18.91 ± 1.99 | 23.75 ± 0.88 |
| pi-helix | 0.09 ± 0.01 | 0.59 ± 0.25 | 0.37 ± 0.03 | 0.50 ± 0.01 |
| turn | 19.42 ± 0.50 | 18.77 ± 0.26 | 19.50 ± 1.00 | 21.58 ± 0.14 |
| (B) Local Conformational Propensity (Backbone Dihedrals) | | | | |
| alpha | 36.90 ± 0.79 | 62.43 ± 0.25 | 39.49 ± 1.88 | 49.31 ± 1.70 |
| left | 4.61 ± 0.19 | 3.72 ± 0.37 | 1.66 ± 0.33 | 1.72 ± 0.29 |
| PP2 | 15.57 ± 0.23 | 7.73 ± 0.17 | 15.89 ± 1.32 | 9.56 ± 0.48 |
| extended | 10.55 ± 0.11 | 4.84 ± 0.02 | 11.74 ± 0.75 | 8.27 ± 0.55 |

## Chapter 3. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent

**Acknowledgments.** This chapter is direct excerpt with minor change from "Nguyen, H.; Maier, J.; Huang, H; Simmerling, C., Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent." Nguyen and Maier are co-first authors.

**Abstract.** The millisecond timescale needed for molecular dynamics simulations to approach quantitative study of protein folding is not yet routine. One approach to extend the simulation time scale is to perform long simulations on specialized and expensive supercomputers such as Anton. Ideally, however, folding simulations would be more economical while retaining reasonable accuracy, and provide feedback on structure, stability and function rapidly enough to partner directly with experimental protocols. Various approaches to this problem typically involve different compromises between accuracy, precision and cost; the goal of this work is to address whether simple implicit solvent models have become sufficiently accurate for their weaknesses to be offset by their ability to rapidly provide much more precise conformational ensembles as compared to explicit solvent. Here, we demonstrate that our recently developed physics-based model performs well on this challenge, enabling accurate all-atom simulated folding of proteins with a variety of sizes, secondary structure and topologies. The simulations were carried out using the Amber software on inexpensive GPUs, providing ~1μsec/day per GPU and >2.5 milliseconds overall. We also show that native conformations are preferred over misfolded structures for most of the proteins. For 3 of the 17 proteins tested, however, folding is successful but misfolded structures are thermodynamically preferred, suggesting opportunities for further improvement.

## 3.1 Introduction

Proteins typically function properly only after folding into a specific three-dimensional structure. Experimental techniques can very accurately determine folded structures, as evidenced by >90,000 structures available in the protein data bank[98]. However, this remains a small subset of the number of known sequences[99]. Moreover, protein folding is a dynamic process, involving nanosecond to millisecond timescale transitions among many unfolded states.[3, 37a] Insight to the factors controlling the folding landscape is crucial to designing proteins with new or enhanced functionality, determining the structures of proteins not yet characterized experimentally, or understanding detrimental effects of protein misfolding and aggregation.

Atomistic simulation models could potentially elucidate folding with spatial, temporal and energetic resolution, but millisecond-scale simulation is far from routine.[100] One way around the timescale problem is approaches like Rosetta, where a combination of empirical and physical rules aids in the prediction of the final native coordinates[101]. Limited experimental data can also be used to focus the search and eliminate inconsistent structures[102]. The structural optimization approach, however, does not provide physical details about the folding process itself, and may be less useful for disordered or dynamic proteins where physics-based approaches may be more successful. Folding@Home[44] makes use of otherwise idle computer time to harvest numerous but relatively short simulations, which can then be assembled into models describing folding.[103] Recently, Shaw and colleagues used the specialized Anton supercomputer[104] to directly fold 12 proteins. [46] This brute-force calculation spanning ~8 ms remains state of the art.

Is there a way to simulate protein folding dynamics in atomic resolution using inexpensive computer hardware that would make these protocols more widely accessible? Implicit solvent models can dramatically accelerate folding due to lower viscosity that facilitates chain diffusion[1a]. Pairwise variants of the generalized Born (GB) model[21] perform particularly well on inexpensive GPUs[105], leveraging a vast consumer video game market to make folding simulations more widely accessible. However, many fast GB models are inaccurate[31], often with incorrect secondary structures preferences and salt bridge strength, thus succeeding in anecdotal cases but lacking broad transferability. GBMV2[20] is arguably the most accurate GB model, but at

a cost of reduced performance.[14] The best performing combination of implicit solvent and protein force field can result from fortuitous cancellation of error in models that have significant but compensating weaknesses[49].

Recently, we reported development of a new fast pairwise GB model that was trained to reproduce more accurate Poisson-Boltzmann solvation across a broad range of peptide and protein systems[106]. Here, we combine it with our widely used ff99SB protein force field[107], along with our recently updated protein side chain parameters. [108] The solvent and protein energetics were trained for independent accuracy, in an attempt to avoid cancellation of error and improve transferability. In this report, we demonstrate that this new physics-based combined model is an attractive tradeoff, enabling accurate folding for all but 1 of a set of 17 proteins ranging from 10 to 92 amino acids. We address two key issues in detail: the sampling problem (whether simulations can fold to the correct structure) and the accuracy problem (whether the preferred structure in the simulated ensemble is native-like).

### 3.2 Methods

#### 3.2.1 Peptides and proteins studied

Seventeen systems were simulated (Table 3.S1), including 12 studied by Shaw and colleagues[46]: CLN025, Trp-cage, BBA, villin HP36, WW domain GTT, $NTL9_{39}$, BBL, protein B, homeodomain 2P6J, the NuG2 variant of protein G, α3D, and λ-repressor. We added a second WW domain (Fip35),[109] and several larger systems: $NTL9_{52}$, cold shock protein A (CspA), hypothetical protein 1WHZ, and Top7. Unless otherwise noted, RMSD values are for Cα atoms regions well defined in structures based on experiments.

#### 3.2.2 Simulation details

All MD simulations were carried out using the GPU implementation[105] of the pmemd program in AMBER14[22a] with the combination of GB-Neck2,[106] mbondi3 intrinsic radii,[106] and ff14SBonlysc, which includes ff99SB[107] with new side chain dihedral parameters from ff14SB[108]. We did not use the backbone dihedral modifications from ff14SB, since they are empirical adjustments aimed at improving agreement between experiment and simulations in explicit water. The protocol delivered 0.6 to 1.4 μs/day (Table 3.S2).

There are many potential limitations of simple implicit solvent models, such as lack of structured water and ions. In addition, nonpolar solvation contributions were not included in this work, as methods for their accurate treatment are less well developed, and their treatment *via* surface area (as done in Amber) is overly simplistic, significantly slows the calculations, and has been reported to bias nonpolar interactions.[12] Although the hydrophobic effect plays a major role in protein folding, we note that neglecting nonpolar solvation also omits the attractive dispersion interaction with solvent, partially compensating for the hydrophobic effect[110]. On the whole, it seems reasonable to test our model without the nonpolar term, since we showed previously that simulations without the nonpolar term performed well on smaller peptide systems.[30] In the section for each system below, we provide figures showing the SASA as a function of RMSD for each system, which provides a qualitative indication of the potential impact of including the SASA term in the simulations.

Initial structures were built using the LEaP module of AmberTools[111] then minimized and equilibrated in three 250 ps stages: heating from 100 K to the production temperature with heavy atom positional restraints of 10 kcal $mol^{-1}$ $Å^{-2}$, reducing force constant from 10.0 to 1.0 and then to 0.1 kcal $mol^{-1}$ $Å^{-2}$. A time step of 4 fs was used with hydrogen mass repartitioning.[112] Bonds involving hydrogen were constrained by the SHAKE algorithm[87] with a tolerance of 0.00001. Temperature was controlled with a Langevin thermostat with collision frequency $\gamma = 1.0$ $ps^{-1}$. We used 300 K except as follows. We initially used 300 K for Fip35; as the native structure was stable 3.for 10 μs, however, the temperature was raised to 325 K to aid folding. We used the same temperature for GTT, which is a variant of Fip35. HP36 and BBL unfolded within tens of ns at 300 K, so these systems were simulated at 290 K. Our subsequent use of REMD avoids the need for selecting a single optimal folding temperature.

### 3.2.3 ff14SBonlysc

Our ff14SBonlysc force field, freely available as part of AmberTools 14 from the Amber web site at ambermd.org, used the backbone dihedral corrections of ff99SB [113] with updated dihedral side chain corrections fit to MP2[114] /6-31+G**[115] //HF/6-31G*[115] ab initio side chain energy surfaces of dipeptides at α (-60°, -45°) and β (-135°, 135°) backbone conformations; all other parameters were from ff94[116]. To limit variability to predominantly side chain motions and to limit backbone-side chain hydrogen bonding that may be incorrectly modeled by fixed charges

in vacuo, all backbone dihedrals were restrained. Valine was fit to 10° χ1 scans. Aspartate (ionic and neutral), asparagine, cysteine, isoleucine, leucine, serine, threonine, phenylalanine, tyrosine, tryptophan, and histidine (δ-, ε-, and doubly-protonated) were fit to 20° χ1 and χ2 two-dimensional scans. Glutamate (ionic and neutral), glutamine, and methionine were fit to randomly distributed conformations extracted from high temperature simulations. Quantum calculations of the one- and two-dimensional scans employed GAMESS (US) (1 MAY 2012 (R1))[117] whereas quantum calculations of the structures from high temperature simulation employed Gaussian 98[118]. Molecular mechanics calculations were performed using Amber 11 and 12[22c, 119]. Fitting was performed by a genetic algorithm[120] using GAlib[121], with parameters restrained to phase shifts of 0 or π to permit simulation of different enantiomers. A complete description of the parameter development will be published elsewhere.

### 3.2.4 Clustering

Means algorithm was used with distance defined by Cα RMSD to generate 50 clusters using default settings in ptraj[122]. The clustering for REMD was performed for the lowest temperature trajectory. Snapshots were used from 5 ns intervals, but this interval was adjusted to ensure between 4000 and 7000 frames.

### 3.2.5 Nonpolar Solvation Analysis

Structures were extracted every 1 ns from extended and native MD simulations. The combined set was postprocessed in SANDER to calculate the cavity contribution to nonpolar solvation (gbsa=2 in Amber), which is proportional to the solvent-accessible surface area determined by recursively optimizing spheres around each atom starting from icosahedra[123]. We then generated population histograms of surface area contribution versus RMSD to the native structure with grid spacing of 0.5 kcal $mol^{-1}$ in nonpolar solvation energy and 0.5 Å in RMSD.

### 3.2.6 Protein folding events (Fip35)

Folded and unfolded conformational cutoffs were assigned by visual inspection of two-dimensional RMSD population histograms (RMSD values for hairpin 1 and for hairpin 2). The Fip35 folded cutoffs were 2.7 and 1.2 Å RMSD for hairpins 1 and 2, respectively. The Fip35

unfolded cutoffs were 5.0 and 4.5 Å RMSD for hairpins 1 and 2, respectively. These numbers were empirically selected to reflect visual boundaries in population around the two states. Whenever a structure went above the two unfolded cutoffs (both hairpins unfolded), it was considered unfolded. Whenever a structure went below the two folded cutoffs, it was considered folded. The total simulation period between an unfolded conformation and a folded conformation was considered a folding path. Each path was plotted in two-dimensional RMSD, with lines colored by time through red, yellow, green, cyan, and blue. The sequence of folding was determined by manually evaluating which RMSD dropped first—visually, whether the folded state was reached from the top (metric on x-axis folded first), the side (metric on y-axis folded first), or diagonally from a conformation where neither was pre-formed.

### 3.2.7 Order parameter calculations

Lipari-Szabo NH librational order parameters $S^2$ [124] were calculated using the cpptraj[122] implementation of iRED [125], which does not require separation of internal and external motions, over 8 ns windows for lysozyme, as done elsewhere[126], and 5 ns for cold shock protein, consistent with its tumbling time[127], in each trajectory. Uncertainties were determined by standard errors in the average $S^2$ for each trajectory. Simulations for order parameter calculations were performed with a 1 fs timestep. GB-Neck2 simulations used Langevin dynamics with a constant of 91 $ps^{-1}$ to mimic water viscosity[128]. TIP3P simulations did not use barostat or thermostat following equilibration. Lysozyme was extended to 96 ns simulation time, and cold shock protein to 60 ns, yielding 12 windows per simulation.

TIP3P[129] simulations used the particle-mesh Ewald approximation[130] with a direct non-bonded cutoff of 8 Å. Equilibration proceeded by minimization of the experimental structure with 100 kcal/mol/$Å^2$ restraints on protein heavy atoms, followed by 100 ps of restrained heating at constant volume from 100 K to 300 K using the weak-coupling (Berendsen) thermostat[131]. Following 100 ps at 300 K and constant volume, the pressure was equilibrated to 1 bar with isotropic position scaling, for 100 and 250 ps with time constants of 100 fs and then 500 fs and restraints of 100 kcal/mol/$Å^2$ and then 10 kcal/mol/$Å^2$. Then the N, Cα, and C were restrained during minimization, followed by three 100 ps simulations with temperature and pressure time constants of 500 fs, reducing restraints from 10 kcal/mol/$Å^2$ to 1 kcal/mol/$Å^2$, and then 0.1 kcal/mol/$Å^2$. Finally, the volume of the unrestrained system was equilibrated with time

constants of 1 ps with a 2 fs time step, removing center-of-mass translation every ps, for 1 ns.

NVE production simulations used a direct sum tolerance of $10^{-6}$ and SHAKE[132] applied to bonds to hydrogen with a tolerance of $10^{-6}$ Å.

### 3.2.8 General REMD setup

Temperature ranges for REMD were chosen for an acceptance ratio of ~0.25. The number of replicas ranged from 6 to 24, depending on system size (Table 3.S2). Exchanges were attempted every 1 ps. Snapshots were saved every 100 ps.

**Extended REMD (extREMD)**

Extended REMD refers to REMD simulations initiated from fully extended initial structures (built by LEaP in Amber) for all replicas. We performed 17 REMD runs for 17 systems starting from extended conformations Temperatures are indicated in Table 3.S3. We performed additional REMD calculations for hypothetical protein 1WHZ using final snapshots from extended MD run (named extMDREMD).

**Seeded REMD**

The goal of running seeded REMD was to indicate which structure (folded or unfolded) is preferred at low temperatures, under conditions in which all of the structures of interest are present in the replica set at the same time. Even though all clusters may have been sampled in the REMD run, they may not have been sampled at the same time, and thus the temperature of the replicas (or the population of the clusters at various temperatures) does not indicate relative favorability. Although the seeding procedure does provide REMD with the opportunity to rank the structures, the resulting "melting" behavior is artificial, since it depends on the numbers of structures of each type used in seeding. We performed seeded REMD for NuG2 variant, CspA, Lambda Repressor, 1WHZ, and Top7. Temperatures are indicated in Table 3.S4.

We performed two seeded REMD simulations with NuG2 variant. In the first, we continued the extREMD calculation, but adding 2 native structures (from an MD run of the crystal structure) at 2 new temperatures in the middle of the previous temperature ladder: 309.0 K and 334.0 K. In the second seeded REMD, we alternated the most populated cluster from

extREMD (11.4 Å RMSD) and a native-like structure (1.0 Å RMSD) through twelve temperatures beginning at 250.0 K.

- For CspA, we alternated misfolded (10.0 Å), near-native (4.7 Å), and native-like (1.2 Å) cluster structures through twelve temperatures beginning at 250.0 K.
- For λ-repressor, we alternated misfolded (12.0 Å), the lowest RMSD from extREMD, and native structures through twelve temperatures beginning at 250.0 K.
- For hypothetical protein 1WHZ, we alternated 2 unfolded (10.6 Å, 10.0 Å), 2 partly folded (3.1 Å, 4.2 Å), and 1 native-like (1.5 Å) replicas through twenty temperatures beginning at 242.0 K.
- For Top7, we alternated partly folded (2.7 Å), unfolded (11.2 Å), and native-like (1.5 Å) replicas through eighteen temperatures beginning with 240.0 K.

Seeding REMD was run for ~40ns for all cases. This was determined to be adequate to sort the replicas such that all replicas starting from the same structure were grouped in a continuous temperature range, as compared to the alternation that was used at the start. The simulation length was also short enough that they generally did not sample large structure changes, since this is not the goal in these calculations.

### 3.2.9 RMSD calculation

RMSD calculations and cluster analysis were performed with ptraj[122] in AmberTools[111]. RMSD calculations excluded flexible termini or other regions, such as loops, that were not well defined in the crystal structure or family of NMR structures (as described below and tabulated in table 3.S3). The reference structure was the experimentally derived structure or, where none was available, the structure of a homologue as described.

**CLN025**

We simulated full-length CLN025. All the Cα atoms in the x-ray structure[133] were used to calculate RMSD.

**Trp-cage**

We simulated full-length Trp-cage tc5b. We calculated RMSD against the first model of the NMR ensemble[59], excluding residues 1 to 2 and 19 to 20 as flexible termini.

**BBA**

We simulated full-length BBA. We calculated RMSD against the first model of the NMR ensemble[134], excluding residues 1 to 3 and 27 to 28 as flexible termini.

**Pin1 WW domain mutants: Fip35 and GTT**

2F21 is a fast-folding Pin1 WW domain mutant[109]. Fip35 is a faster folding (13 μs) mutant based on residues 6–38 of 2F21[135]. GTT is an even faster folding (4 μs) mutant based on Fip35 plus the two prior residues in 2F21[136]. We simulated full-length Fip35 and GTT. We calculated RMSDs against 2F21, using residues 10 to 32—the first residue of the first β-strand to the final residue in the last β-strand.

**HP36**

We simulated the thermostable C-terminal fragment (residues 41 to 76) of the chicken villin headpiece (HP36). Shaw and colleagues[46] used the HP35 variant of villin peptide with norleucine double mutant[64] to accelerate folding. We chose the HP36 variant[64], which includes only standard amino acids. We calculated RMSDs against an averaged NMR structure (PDB ID: 1VII)[64], using residues 43 to 72 regions to exclude flexible termini of the NMR ensemble of a G34L mutant[137].

**NTL9 (39 AA)**

NTL9 (39) is an N-terminal truncation (residues 1 to 39) of N-terminal domain of 50S ribosomal protein L9 (NTL9). We simulated the K12M mutant. We calculated RMSDs against residues 1 to 39 of the crystal structure of the full-length K12M sequence (PDB ID: 2HBA)[138].

**BBL**

We simulated the H142W mutant of BBL, residues 124 to 170—the residues in a solution structure ensemble[139] (PDB ID: 2WXC). We calculated RMSDs against the experimental structure skipping the flexible N-terminal residues and the flexible loop from residue 152 to 158. Our mask thus included residues 133–151 and 159–170.

**Protein B**

We simulated a K5I/K39V double mutant of truncated Protein B (residues 7–53) of the NMR structure (PDB ID: 1PRB), as done previously[7]. We calculated RMSDs against the NMR structure using residues 8–50, including the start of the first helix to the end of the last helix.

**Homeodomain**

We simulated a computationally re-designed variant of Drosophila Melanogaster Engrailed homeodomain[140]. We calculated RMSDs against the first model of the NMR ensemble (PDB ID: 2P6J)[140], using residues 5 to 48 to exclude flexible termini.

**NTL9 (52 AA)**

In addition to the 39 residue NTL9 described above, we also simulated the full length N-terminal domain of the 50S ribosomal protein L9 (NTL9) K12M[138], denoted as NTL9 (52 AA). We calculated RMSDs against the first monomer in the crystal structure (PDB ID: 2HBA).

**NuG2 variant**

We simulated residues 6 to 61 of a N37A/A46D/D47A mutant of NuG2 (PDB ID: 1MI0[141]), as done previously[136]. We calculated RMSDs against the crystal structure of unmodified NuG2, including all simulated residues from 6 to 61.

**CspA**

We simulated the major cold shock protein of *Escherichia coli*, CspA, excluding the first residue missing from the x-ray structure (PDB ID: 1MJC[142]). We calculated RMSD against all structured regions in the experimental structure (residues 4–14, 16–23, 29–36, 48–56, and 62–70), as the loops are flexible in an NMR ensemble of the same sequence (PDB ID: 3MEF[127]).

**Hyp protein 1WHZ**

We simulated full-length hypothetical protein from *Thermus thermophilus* HB8. We calculated RMSDs against all residues in the crystal structure (PDB ID: 1WHZ[143]).

**α3D**

We simulated full-length α3D, a de novo designed three-helix bundle[144]. We calculated RMSDs against all residues in the solution structure (PDB ID: 2A3D[144]).

**λ-repressor**

We simulated truncated (residues 6–85) monomeric D14A/Y22W/Q33Y/G46A/G48A mutant of λ-repressor studied previously [136]. We calculated RMSD against the unmutated homologue (PDB ID: 1LMB[145]), however. In the x-ray structure, dimeric λ-repressor binds DNA, along with multivalent ions. We calculated RMSDs against all simulated residues (6–85) in the first protein chain in this complex.

**Top7**

We simulated residues 3 to 94 of Top7, which was computationally designed with a novel fold[146]. We calculated RMSDs against the x-ray structure (PDB ID: 1QYS[146]), using residues 3 to 94.

### 3.3 Results

Our goal in this study is to investigate feasibility of simulating all-atom folding for a variety of proteins with a single force field and solvent model combination, using inexpensive, widely available computer hardware and software. System sizes range from short peptides to proteins of nearly 100 amino acids, with topologies including all α-helix, all β-sheet, and combinations. Experimental folding times vary from microseconds to seconds (Table 3.S2). Overall, the set comprises a challenging benchmark for any folding study. We first performed a baseline study on native state dynamics of 2 proteins, CspA and lysozyme, and compared backbone order parameters from the resulting simulations to those obtained from experiment as well as from simulation with explicit water (Figure 3.S1). We obtained excellent quantitative agreement (0.05 and 0.02 RMSD to experimental and TIP3P $S^2$ for CspA, and 0.02 RMSD to both experimental and TIP3P $S^2$ for lysozyme), suggesting that more challenging tests were warranted.

We separate our analysis of protein folding below along two general goals. First, we address sampling: in spite of the limitations of the implicit solvent model, can standard MD simulations properly fold to the correct experimental structure when starting from a fully extended conformation? Second, we address accuracy: is the experimental structure also the most favorable in our model? The latter goal is significantly more challenging; the physics must be accurate enough to reproduce the correct global free energy minimum for a variety of topologies and secondary structure combinations, and the populations of the minima must be well converged in order to make precise predictions. For several of the larger systems studied here, convergence was not readily achieved in standard MD, and thus we used replica exchange (REMD[2]). The ability to use REMD further supports our premise that disadvantages of implicit solvent are in some cases offset by significant advantages, since REMD on proteins in explicit solvent is largely intractable due to computational cost[147].

### 3.3.1. Can simulations fold to native conformations?

Simulations starting in extended conformations were able to locate structures in excellent agreement with experiment for 16 of the 17 systems (Table 3.S2). All of the proteins smaller than 50 amino acids fold well in standard MD on this timescale, reaching Cα RMSD values below 2 Å, except BBL which reaches 3.2Å (all-time series data are in Supporting Information). This includes systems with β-sheet (the hairpin CLN025 and the 3-stranded sheets Fip35 and GTT), α-helix (Trp-Cage, HP36 and protein B) and mixed α/β (BBA and the 39 amino acid version of NTL9). In REMD, these systems all fold to <2.1 Å Cα RMSD, with minimum RMSD values often below 1Å. While it is beyond the scope of this work to fully analyze side chain packing accuracy, the heavy atom RMSD of the Fip35 conformation with the lowest Cα RMSD is 1.8 Å, while that of NTL9 (39 AA) is 1.4 Å, suggesting that highly accurate folding is achievable with our protocol.

The larger proteins (50-92 amino acids) tend to become kinetically trapped in standard MD on the microsecond timescale, with only homeodomain (1.9 Å), α3D (2.5 Å) and λ-repressor (4.4 Å) finding native-like conformations. The enhanced sampling provided by REMD provides notable benefit, with 16 of the 17 proteins now folding to structures with RMSD values under 3 Å (Figure 3.1). Only the NuG2 variant was still unable to sample the correct conformation; the minimum RMSD value is 4.8 Å, with the first hairpin and helix correctly formed, but the second hairpin not yet formed. In contrast, NuG2 simulations initiated from the experimental structure underwent significant unfolding to ~10 Å RMSD, followed by refolding to an accurate native state (< 1.0 Å RMSD, Figure 3.S40).

One advantage of direct simulation of folding is that it is possible to analyze folding pathway(s). Direct comparison of kinetic data to experiments is precluded by our use of low viscosity to enhance sampling. Instead, we consider the relative flux through various folding pathways, presenting a single example since a comprehensive analysis is beyond the scope of the present manuscript. We analyzed which of the 2 hairpins in Fip35 WW folded first in the12 independent folding events that we observed in the MD run from extended structure (Figure 3.S16), obtaining a 4:1 ratio favoring initial folding of hairpin 1. This ratio is in excellent agreement with the 4:1 ratio reported for explicit solvent simulations of the same system[45a].

### 3.3.2. Do the simulations show the correct structure preferences?

Next, we address the more challenging issue of accuracy, and whether our model could predict a qualitatively reasonable structure if it were not already known, by comparing the experimental structure to the most populated cluster from simulation. For 10 of the 17 systems, multiple (> 3) folding and unfolding events were observed in the standard MD runs; however, many remained poorly converged even on the μs timescale, particularly for the longer proteins. We therefore use the REMD ensembles to obtain qualitative estimates of the preferred conformations for each protein.

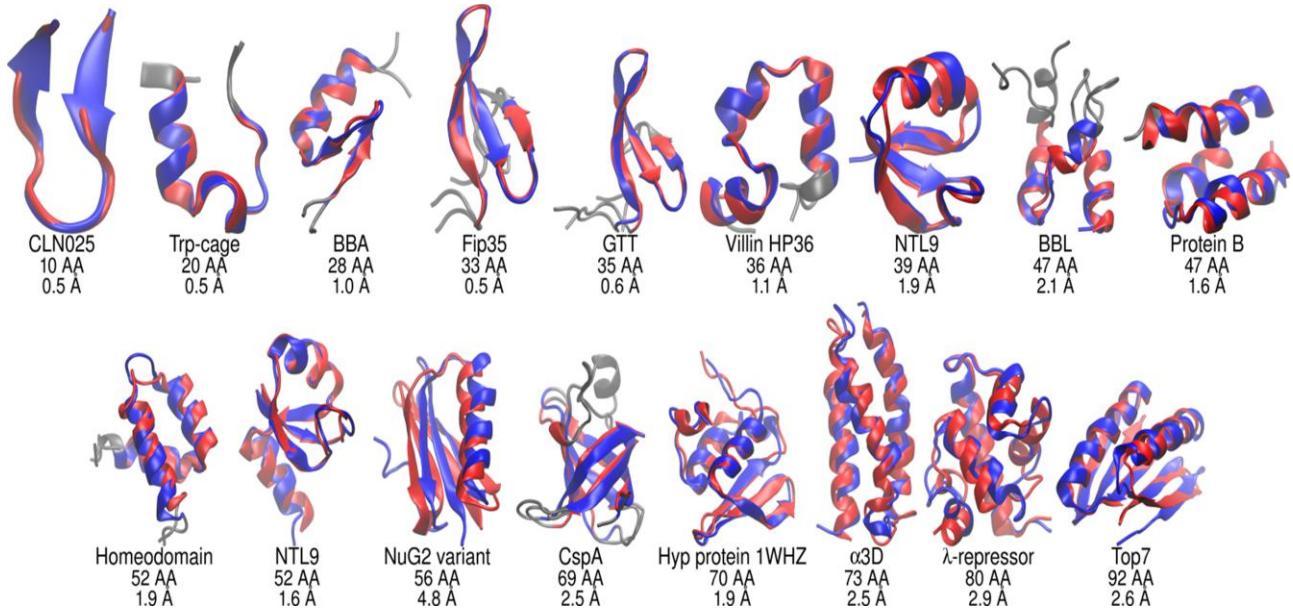


**Figure 3.1.** Comparison of structures based (red) on experiment and (blue) lowest RMSD sampled in simulations started from extended conformations. Gray regions were excluded from RMSD calculations. Under each structure is the protein name, chain length and Cα RMSD value.

The cluster with the largest population was in good agreement with conformations based on experiment for roughly half (8 of 17) of the proteins studied (RMSD values are provided in Table 3.S2, with structures shown in Figure 3.S2). Once again, performance tended to be better for proteins under 50 amino acids, with **CLN025**, **Trp-cage**, **Fip35**, **GTT** and **HP36** all

preferring the correct structure, with representative structure RMSD values of 0.6-3.0 Å. For **protein B**, the representative structure has an RMSD value of 4.2 Å: properly folded but with a slight rotation of the middle helix relative to the core. In the case of **BBA**, the native zinc finger fold is present in the ensemble, but with lower population than the preferred alternate structure with RMSD of 4.6 Å, in which the hairpin and helix are both still present, but with somewhat longer hairpin and shorter helix. Although the 39 and 52 amino acid long variants of **NTL9** both fold properly in the simulations, the large RMSDs for the most populated cluster in both systems (6.1 and 6.0 Å, respectively) reflect higher population of an otherwise properly folded structure with an alternate conformation of the loop connecting β-strands 1 and 2. Neglecting this loop, the RMSD of the largest clusters become a more reasonable 4.1 and 4.2 Å, respectively. The only protein under 50 amino acids that prefers an incorrect fold is **BBL**, which locates the correct fold from extended structures, but favors a conformation with 8.3Å RMSD in which the ends of the short N- and C-terminal and intervening helices become disordered. However, the 2$^{nd}$ and 3$^{rd}$ most populated clusters have more reasonable RMSD values of 4.3 and 4.8Å. Lindorff-Larsen et al.[46] estimated a very low melting temperature in BBL simulations (270±10 K), suggesting that BBL challenges not only our model, but also MD with explicit water.

For the other 7 proteins larger than 50 amino acids, only **homeodomain** and ***α*3D** have most populated clusters (23% and 33%, respectively) that are close to those observed experimentally (3.2 and 4.0 Å, respectively). The second most populated cluster of homeodomain (8%) is even closer to experiment (2.3 Å). For both systems, the differences are predominantly in the surface loop regions, and the RMSDs for the 3 helices are 2.5 Å for homeodomain and 2.1 Å for α3D.

As discussed above, the **NuG2** variant was the only system that never sampled the native conformation, thus the cluster populations cannot report on whether the correct structure would be preferred if folding had occurred. To explore this further, we carried out an additional ~40ns "seeded" REMD simulation continuing from the end of the previous one, but adding 2 equilibrated native structures at 2 new temperatures in the middle of the previous temperature ladder (see Supporting Information for complete details, including descriptions of misfolded structures, seeded REMD setup and results). Our expectation was that the REMD exchanges would perform sorting, placing the more favorable structures at the lower temperatures. The simulations showed a strong preference for the native fold over the other structures, moving both

low RMSD structures to low temperatures (Figure 3.S43). We then competed 6 native and 6 misfolded structures from the initial REMD run. The native structures were again strongly preferred at low temperature (Figure 3.S44), suggesting that our model correctly identifies the NuG2 native fold, and misfolding represents a sampling failure.

The other four systems for which the largest cluster in REMD was non-native (RMSDs of 10–12 Å) were **CspA**, **1WHZ, λ-repressor** and **Top7**. In each case, examination of RMSD history for each of the replicas in REMD showed that only a few replicas properly folded, and likewise, only a few misfolded. The data suggest that even though the structures are reproducibly sampled, REMD remains unreliable for distinguishing the relative stability of these alternate conformations. We again turned to a seeded REMD approach for gaining additional insight into the conformational preferences of our model. In each case, native-like structures were alternated in the temperature ladder with representative structures from misfolded clusters with large populations (**Figures** S48, S53, S60, S68). The results suggest that, among the 4 proteins with unconverged ensembles, our model can accurately identify the native conformation for **CspA** and **Top7**. For **CspA**, only 2 replicas misfolded in REMD, and 2 others located a near-native fold, suggesting poor population convergence even after ~30 μseconds of REMD, which is perhaps not surprising given the experimental folding rate of ~5 milliseconds[148]. REMD seeded with native, near-native and misfolded structures showed a strong preference for the native structure at the lowest temperatures. **Top7** showed similar behavior, with the highest population misfolded structure only being sampled by 1 REMD replica. Seeded REMD combining the misfolded and correctly folded structures showed a strong preference for the correct fold, moving all misfolded structures to higher temperatures. Interestingly, two of the **Top7** replicas that were initially misfolded underwent spontaneous refolding to the correct structure during this run. The results provides additional evidence that that our model prefers the native fold and that the variety of kinetic traps that the **Top7** simulations encountered was a result of the non-cooperative, seconds-timescale folding experimentally observed for this system[149].

In contrast to the other systems, the seeded REMD results suggest that the model fails to accurately recognize the native conformation of **λ-repressor** and **1WHZ**, preferring misfolded over native structures at low temperature. **λ-repressor** shows transient folding to the native structure in REMD, but prefers a misfolded structure with the 5 α-helices largely present, but packed against the first helix in a clockwise fashion, rather than counterclockwise as seen in the

native fold. **1WHZ** also folds to the correct structure with a 3-stranded β-sheet and 3 helices, but the preferred structure replaces the first β-strand with a helix and the last two helices with two β-strands. Otherwise, the RMSDs of the first helix and N-terminus (residues 1 to 18) and the second and third β-strands (residues 28 to 44) are both 1.8 Å.

To summarize the analysis of our second goal (conformational preferences), all 11 proteins smaller than 55 amino acids were reasonably converged and all except BBL preferred the native fold, with some differences in loop regions. For the 6 larger proteins, only α3D appears well converged in the REMD runs, with the others all sampling multiple clusters and having populations that indicated the model favors non-native folds. We used a seeded REMD approach to evaluate the relative populations of native *vs.* non-native folds, and found that NuG2, CspA and Top7 prefer native conformations, while the model prefers misfolded structures for λ-repressor and 1WHZ. Overall, the data suggest correct preference for the native fold in 14 of the 17 proteins that we studied (Figure 3.S2).

### 3.4 Conclusions

We presented *ab initio* folding for a set of 17 proteins, ranging from 10 to 92 amino acids, with different topologies and secondary structure content. We used an efficient implicit solvent model[106] combined with an accurate protein force field, using the Amber software running on GPUs. This largely solves the sampling aspect of folding proteins of this size; we demonstrated that folding to the correct structure is achievable for all but 1 of the systems that we studied, within run times of several days to 3 weeks. For the larger proteins where convergence was challenging, we used REMD to evaluate the extent to which our model could correctly predict preference of native over misfolded structures; such analysis remains highly challenging in explicit water. Despite being able to fold to correct structures, some of the systems showed stronger preference for alternate, non-native structures, ranging from misfolded loops to incorrect topologies. Future detailed analysis of possible trends in misfolding, quality of side chain packing, and overall protein stability, along with application to a larger range of systems, could provide crucial insight into the limitations in accuracy of our models, and possible routes for further improvement.

**Appendix 3. Supporting document**

**NH order parameters**

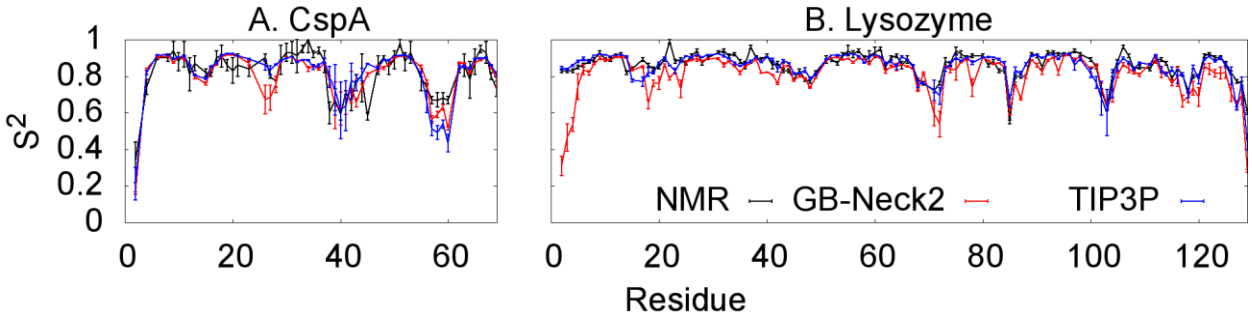


**Figure 3.S1.** Order parameters measuring the NH librational motions of (A) CspA and (B) lysozyme according to NMR[127, 150] (black), GB-Neck2 (red), and TIP3P (blue). All simulation data used force field 14SBonlysc with order parameters backcalculated by iRED. Error bars reflect the standard deviation of the averages from windows in the simulation.

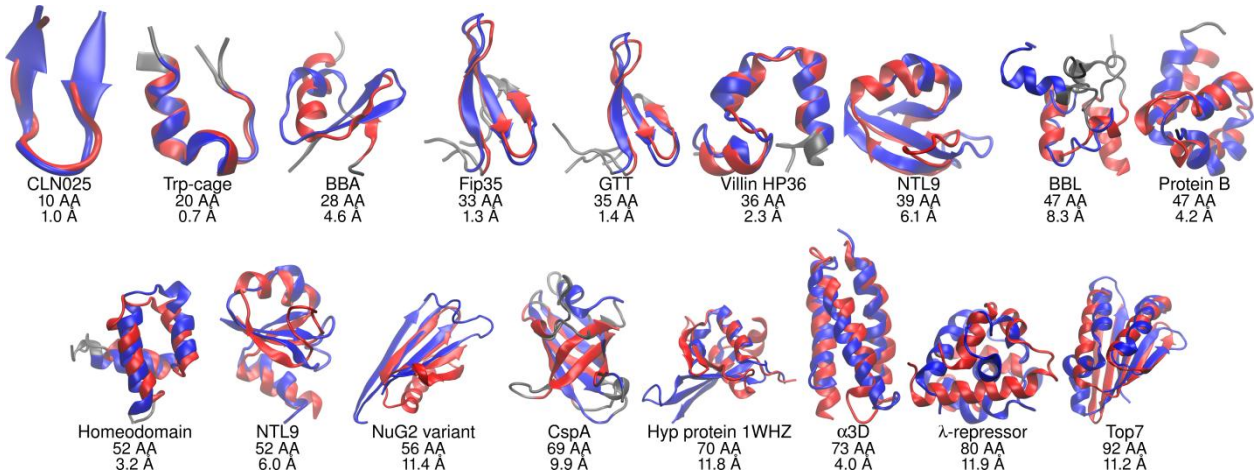


**Figure 3.S2**. The most populated cluster of each protein starting from extended REMD simulations, in blue, aligned to the experimental structure, in red. In the cases of BBA, BBL, NuG2 variant, CspA, Hyp protein 1WHZ, and Top7, the alignment above reflects the parts of the structure best reproduced by the simulations, rather than the alignment yielding the lowest RMSD. Below each rendering is the system name, the number of amino acids (AA), and the RMSD between the two structures (neglecting the flexible gray regions, as in Figure 3.1).

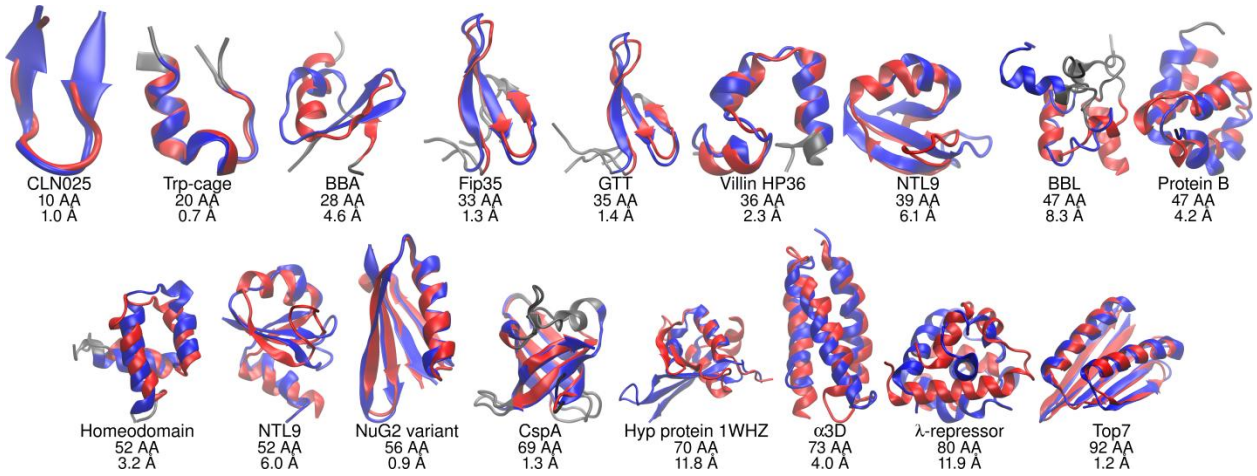


**Figure 3.S3**. The structure of each protein preferred by the force field, either: the centroid of the most populated cluster from extended REMD; or, as in NuG2 variant, CspA, and Top7, the preferred cluster in seeded REMD (see main text for details). The color code follows Figure 3.S2.

**Table 3.S1.** Sequence of peptides and proteins simulated in this work. $H^{\delta}$, $H^{\varepsilon}$, and $H^{\delta\varepsilon}$ stand for Histidine that is protonated at $N^{\delta}$, $N^{\varepsilon}$ or both $N^{\delta}$ and $N^{\varepsilon}$, respectively. All His protonation states were used as indicated in the experimental studies.

| System name | Sequence |
|---|---|
| CLN025 | YYDPETGTWY |
| Trp-cage | NLYIQWLKDGGPSSGRPPPS |
| BBA | EQYTAKYKGRTFRNEKELRDFIEKFKGR |
| WW domain Fip35 | KLPPGWEKRMSRDGRVYYFNH$^{\delta}$ITNASQFERPSG |
| WW domain GTT | GSKLPPGWEKRMSRDGRVYYFNH$^{\varepsilon}$ITGTTQFERPSG |
| Villin HP36 | MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF |
| NTL9 (39 AA) | MKVIFLKDVKGMGKKGEIKNVADGYANNFLFKQGLAIEA |
| BBL | GSQNNDALSPAIRRLLAEWNLDASAIKGTGVGGRLTREDVEKH$^{\delta\varepsilon}$LAKA |
| Protein B | LKNAIEDAIAELKKAGITSDFYFNAINKAKTVEEVNALVNEILKAH$^{\varepsilon}$A |

| | |
|---|---|
| Homeodomain | MKQWSENVEEKLKEFVKRH$^{\delta}$QRITQEELH$^{\delta}$QYAQRLGLNEEAIRQFFEEFEQRK |
| NTL9 (52 AA) | MKVIFLKDVKGMGKKGEIKNVADGYANNFLFKQGLAIEATPANLKALEAQKQ |
| NuG2 variant of Protein G | DTYKLVIVLNGTTFTYTTEAVDAATAEKVFKQYANDAGVDGEWTYDAATKTFTVTE |
| CspA (1MJC) | SGKMTGIVKWFNADKGFGFITPDDGSKDVFVHFSAIQNDGYKSLDEGQKVSFTIESGAK GPAAGNVTSL |
| Hyp protein 1WHZ | MWMPPRPEEVARKLRRLGFVERMAKGGHRLYTHPDGRIVVVPFHSGELPKGTFKRILR DAGLTEEEFHNL |
| α3D | MGSWAEFKQRLAAIKTRLQALGGSEAELAAFEKEIAAFESELQAYKGKGNPEVEALRK EAAAIRDELQAYRH$^{\delta}$N |
| λ-repressor | PLTQEQLEAARRLKAIWEKKKNELGLSYESVADKMGMGQS AVAALFNGINALNAYNAALLAKILKVSVEEFSPSIAREIY |
| Top7 | DIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELMDYIKKQGAKRVRISITARTKKEAE KFAAILIKVFAELGYNDINVTFDGDTVTVEGQL |

**Table 3.S2.** System details. The protein name, PDB ID, number of amino acids of simulated system, overall topology, residues in RMSD mask, MD speed (μs/day), MD temperature, native MD length (μs), extended MD length (μs), lowest RMSD in extended MD (Å), RMSD of extended MD largest cluster centroid (Å), extended REMD simulation time (μs), lowest RMSD in extended REMD (Å), RMSD of extended REMD largest cluster centroid (Å), experimental folding time. Asterisks following PDB IDs indicate differences between the system in the crystal and in the simulation. RMSD region is listed as amino acid residue IDs in PDB from RCSB[151].

| Protein | PDB ID | # AA | Secondary Structure type | RMSD region (amino acid numbers) | μs/day | MD T, K | MD-native length, μs | MD-extended length, μs | Lowest RMSD, Å | Largest cluster RMSD, Å | REMD-extended length, μs | Lowest RMSD, Å | Largest cluster RMSD, Å | Experimental folding time (μs) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLN025 | Honda et al.[133] | 10 | beta | 1-10 | 1.4 | 300 K | 1.2 | 2.4 | 0.5 | 1.0 | 0.8 | 0.3 | 1.0 | ~0.1(300 K)[152] |
| Trp-cage | 1L2Y[59] | 20 | alpha | 3-18 | 1.3 | 300 K | 1.7 | 1.0 | 0.5 | 0.6 | 0.4 | 0.3 | 0.7 | ~ 4 (298 K)[153] |
| BBA | 1FME[134] | 28 | mixed | 4-26 | 1.4 | 300 K | 5.2 | 7.8 | 1.0 | 1.9 | 9.1 | 0.9 | 4.6 | N/A (low stability)[134] |
| Fip35 | Freddolino et al.[109] | 33 | beta | 5-27 | 1.4 | 325 K | 29.0 | 25.6 | 0.5 | 1.6 | 3.0 | 0.4 | 1.3 | 13 (337 K)[135] |
| GTT | 2F21*[154] | 35 | beta | 10-32 | 1.4 | 325 K | 12.4 | 21.6 | 0.6 | 1.5 | 3.3 | 0.5 | 1.4 | ~ 4 (353 K)[136] |
| Villin HP36 | 1VII[64] | 36 | alpha | 43-72 | 1.4 | 290 K | 22.1 | 26.7 | 1.1 | 2.4 | 4.2 | 1.1 | 2.3 | ~ 10 (330–350 K)[155] |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NTL9 (39 AA) | 2HBA*[156] | 39 | mixed | 1-39 | 1.4 | 300 K | 47.6 | 65.8 | 1.9 | 6.4 | 30.1 | 0.4 | 6.1 | ~ 700 (298 K)[138] |
| BBL | 2WXC[157] | 47 | mixed | 133-151,159-170 | 1.2 | 290 K | 14.1 | 17.1 | 3.2 | 8.5 | 2.2 | 2.1 | 8.3 | ~ 14 (283 K)[139] |
| Protein B | 1PRB*[158] | 47 | alpha | 8-50 | 1.0 | 300 K | 4.6 | 10.3 | 1.6 | 4.2 | 1.9 | 1.6 | 3.3 (4.2) | ~ 1 (298 K)[159] |
| Homeodomain | 2P6J[140] | 52 | alpha | 5-48 | 1.0 | 300 K | 7.2 | 17.3 | 1.9 | 3.0 | 3.5 | 1.6 | 3.2 | ~ 13 (308 K)[160] |
| NTL9 (52 AA) | 2HBA[156] | 52 | mixed | 1-52 | 1.0 | 300 K | 11.8 | 10.2 | 6.1 | 11.4 | 21.2 | 1.6 | 6.0 | ~1400 (298 K)[138] |
| NuG2 variant | 1MI0*[141] | 56 | mixed | 6-61 | 1.0 | 300 K | 51.3 | 54.7 | 7.5 | 9.6 | 28.8 | 4.8 | 11.4 | ~ 60 (298 K)[161] |
| CspA | 1MJC[142] | 69 | beta | 4-14,16-23,29-36,48-56,62-70 | 0.8 | 300 K | 2.7 | 6.9 | 8.7 | 10.1 | 29.4 | 2.5 | 9.9 | ~5000 (298 K)[148] |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hypothetical protein from *Thermus thermophilus* 1WHZ | 1WHZ[143] | 70 | mixed | 6-70 | 0.8 | 300 K | 14.3 | 22.5 | 5.9 | 9.7 | 9.0 | 1.9 | 11.8 | not available |
| α3D | 2A3D[144] | 73 | alpha | 1-73 | 0.8 | 300 K | 6.6 | 20.5 | 2.5 | 3.7 | 1.2 | 2.9 | 4.0 | > 3.2 (323 K)[162] |
| λ-repressor | 1LMB*[145] | 80 | alpha | Chain3 6-85 | 0.7 | 300 K | 26.6 | 39.3 | 4.4 | 10.5 | 24.0 | 2.9 | 11.9 | ~ 10 (350 K)[163] |
| Top7 | 1QYS[146] | 92 | mixed | 3-94 | 0.6 | 300 K | 8.0 | 5.4 | 12.0 | 14.7 | 18.2 | 2.6 | 11.2 | > $10^5$ (295 K)[149] |

**Table 3.S3.** Temperatures used for extended REMD simulations

| System | Extended REMD temperatures (K) |
|---|---|
| CLN025 | 275.1 · 300.0 · 327.2 · 356.8 · 389.1 · 424.3 |
| Trp-cage | 264.0 · 281.4 · 300.0 · 319.8 · 340.9 · 363.3 · 387.3 · 412.9 |
| BBA | 243.8 · 256.8 · 270.4 · 284.8 · 300.0 · 316.0 |
| Fip35 | 285.4 · 300.0 · 315.4 · 331.5 · 348.5 · 366.3 · 385.0 · 404.7 · 425.5 · 447.2 · 470.1 · 494.2 · 519.5 · 546.0 · 574.0 · 603.4 |
| GTT | 285.4 · 300.0 · 315.4 · 331.5 · 348.5 · 366.3 · 385.0 · 404.7 · 425.5 · 447.2 · 470.1 · 494.2 · 519.5 · 546.0 · 574.0 · 603.4 |
| Villin HP36 | 250.0 · 262.2 · 275.0 · 288.4 · 300.0 · 317.3 · 332.8 · 349.0 |
| NTL9 (39) | 273.3 · 286.3 · 300.0 · 314.3 · 329.3 · 345.1 · 361.6 · 378.8 |
| BBL | 274.9 · 287.2 · 300.0 · 313.4 · 327.4 · 342.0 · 357.2 · 373.2 · 389.8 · 407.2 · 425.3 · 444.3 · 464.1 · 484.8 · 506.5 · 529.1 |
| Protein B | 290.0 · 300.0 · 316.0 · 329.8 · 344.2 · 359.3 · 375.0 · 391.5 · 408.6 · 426.5 · 445.2 · 464.7 · 485.0 · 506.3 |
| Homeodomain | 288.7 · 300.0 · 311.7 · 323.9 · 336.6 · 349.8 · 363.5 · 377.7 · 392.4 · 407.8 · 423.8 · 440.3 · 457.6 · 475.5 · 494.1 · 513.4 |
| NTL9 (52) | 280.0 · 291.6 · 300.0 · 316.2 · 329.2 · 342.9 · 357.0 · 371.8 · 387.2 · 403.2 |
| NuG2 variant | 280.0 · 291.4 · 303.3 · 315.7 · 328.6 · 342.0 · 355.9 · 370.5 · 385.6 · 401.3 |
| CspA | 290.0 · 300.0 · 311.8 · 323.3 · 335.3 · 347.7 · 360.5 · 373.8 · 387.6 · 401.9 |
| Hypothetical protein 1WHZ | 280.0 · 289.6 · 300.0 · 309.9 · 320.5 · 331.6 · 343.0 · 354.7 · 366.9 · 379.6 · 392.6 · 406.1 |
| α3D | 289.9 · 300.0 · 310.4 · 321.2 · 332.3 · 343.9 · 355.8 · 368.1 · 380.9 · 394.1 · 407.8 · 422.0 · 436.6 · 451.7 · 467.4 · 483.6 · 500.4 · 517.8 · 535.8 · 554.3 · 573.6 · 593.5 · 614.1 · 635.4 |
| λ-repressor | 290.4 · 300.0 · 309.9 · 320.1 · 330.7 · 341.6 · 352.9 · 364.5 · 376.6 · 389.0 |

| | |
|---|---|
| Top7 | 280.0 · 288.5 · 300.0 · 306.3 · 315.7 · 325.3 · 335.1 · 345.3 · 355.8 · 366.7 · 377.8 · 389.3 |

**Table 3.S4.** Temperatures used for seeded REMD simulations

| System | Seeded REMD temperatures (K) |
|---|---|
| NuG2 variant (1) | 250.0 · 260.2 · 270.8 · 281.9 · 293.4 · 305.3 · 317.8 · 330.8 · 344.3 · 358.3 · 372.9 · 388.2 |
| NuG2 variant (2) | 280.0 · 291.4 · 303.3 · 315.7 · 328.6 · 342.0 · 355.9 · 370.5 · 385.6 · 401.3 · 309.0 · 334.0 |
| CspA | 250.0 · 259.2 · 268.8 · 278.7 · 289.0 · 299.7 · 310.8 · 322.2 · 334.1 · 346.5 · 359.3 · 372.6 |
| Hypothetical protein 1WHZ | 242.0 · 250.0 · 258.6 · 267.5 · 271.0 · 276.7 · 286.2 · 296.0 · 306.2 · 311.0 · 316.7 · 327.6 · 338.9 · 350.5 · 356.0 · 362.6 · 375.1 · 388.0 · 401.3 · 410.0 |
| λ-repressor (1) | 250.0 · 260.0 · 270.0 · 280.0 · 290.4 · 300.0 · 309.9 · 320.1 · 330.7 · 341.6 · 352.9 · 364.5 · 376.6 · 389.0 |
| λ-repressor (2) | 258.3 · 266.8 · 275.6 · 284.7 · 294.1 · 303.8 · 313.8 · 324.2 · 334.9 · 345.9 · 357.3 · 369.1 |
| Top7 | 240.0 · 247.3 · 254.8 · 262.6 · 270.6 · 278.8 · 287.3 · 296.0 · 305.0 · 314.3 · 323.8 · 333.7 · 343.8 · 354.3 · 365.1 · 376.2 · 387.6 · 399.4 |

**System data**

**CLN025**

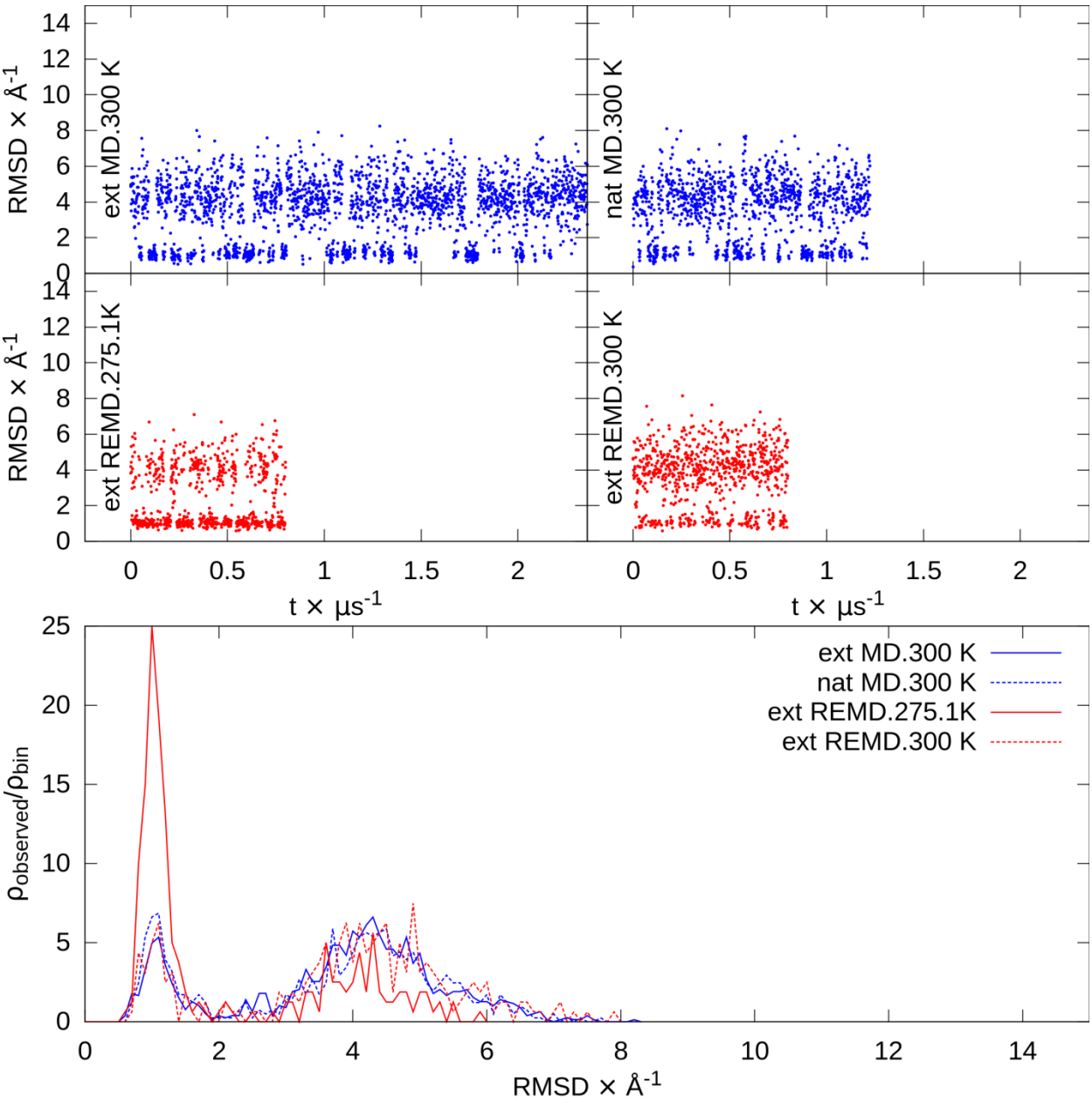


**Figure 3.S4.** CLN025 RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

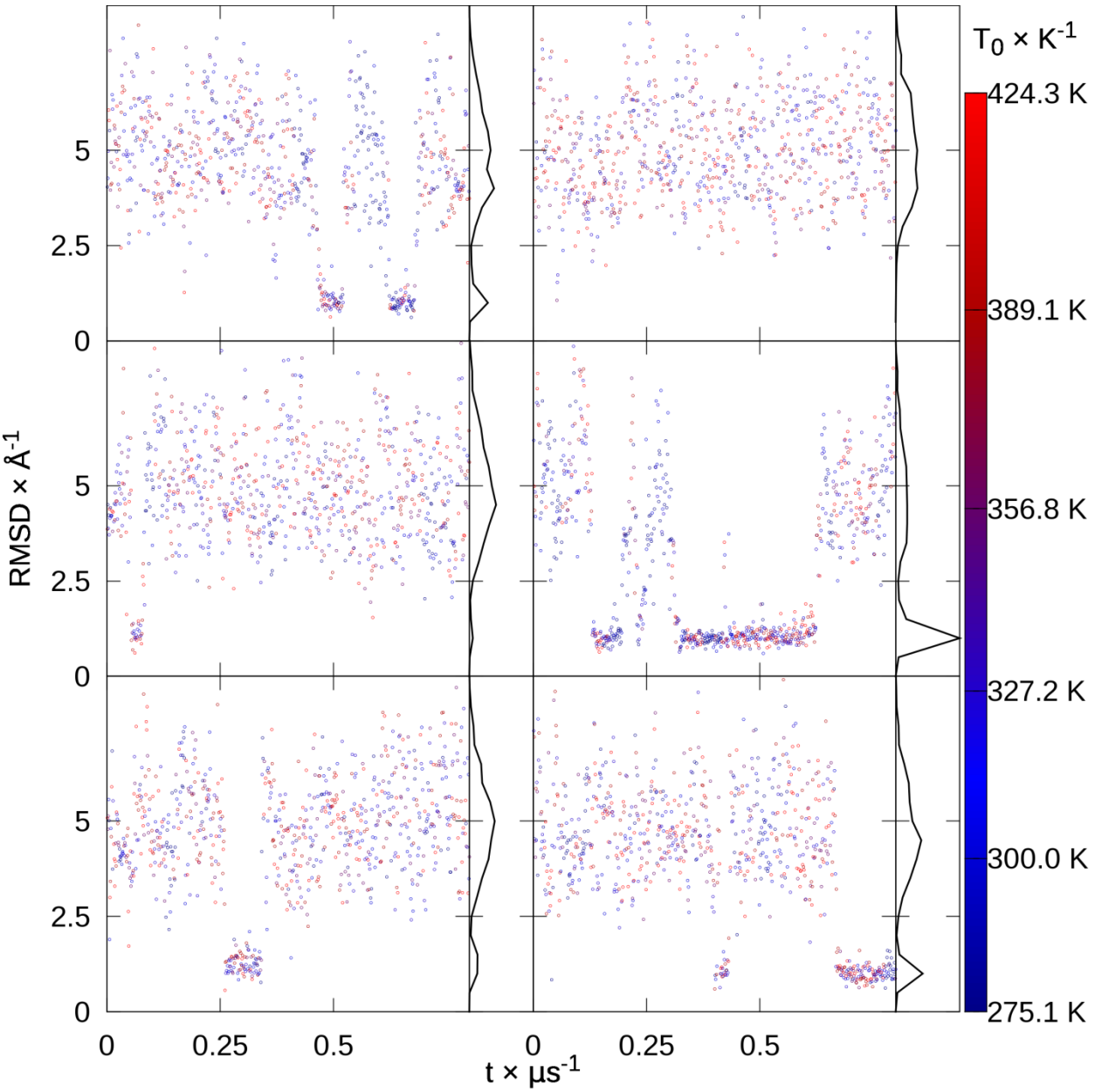


**Figure 3.S5**. CLN025 replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms shown on the right.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population | 57.6 | 7.8 | 2.9 | 2.0 | 1.9 |
| Centroid Cα RMSD (Å) | 1.0 | 4.2 | 3.7 | 5.1 | 3.4 |

**Table 3.S5.** CLN025 top 5 extended REMD cluster populations and centroid Cα RMSDs.

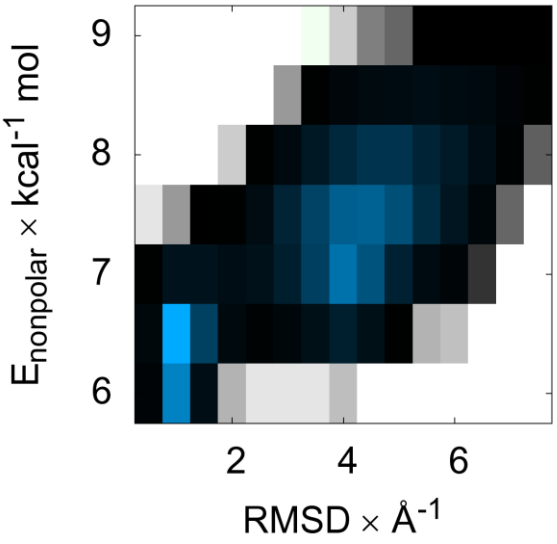


**Figure 3.S6.** CLN025 surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal $mol^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is modestly more favorable at low (around 1 Å) than medium (around 4 Å) RMSDs.

**Trp-cage**

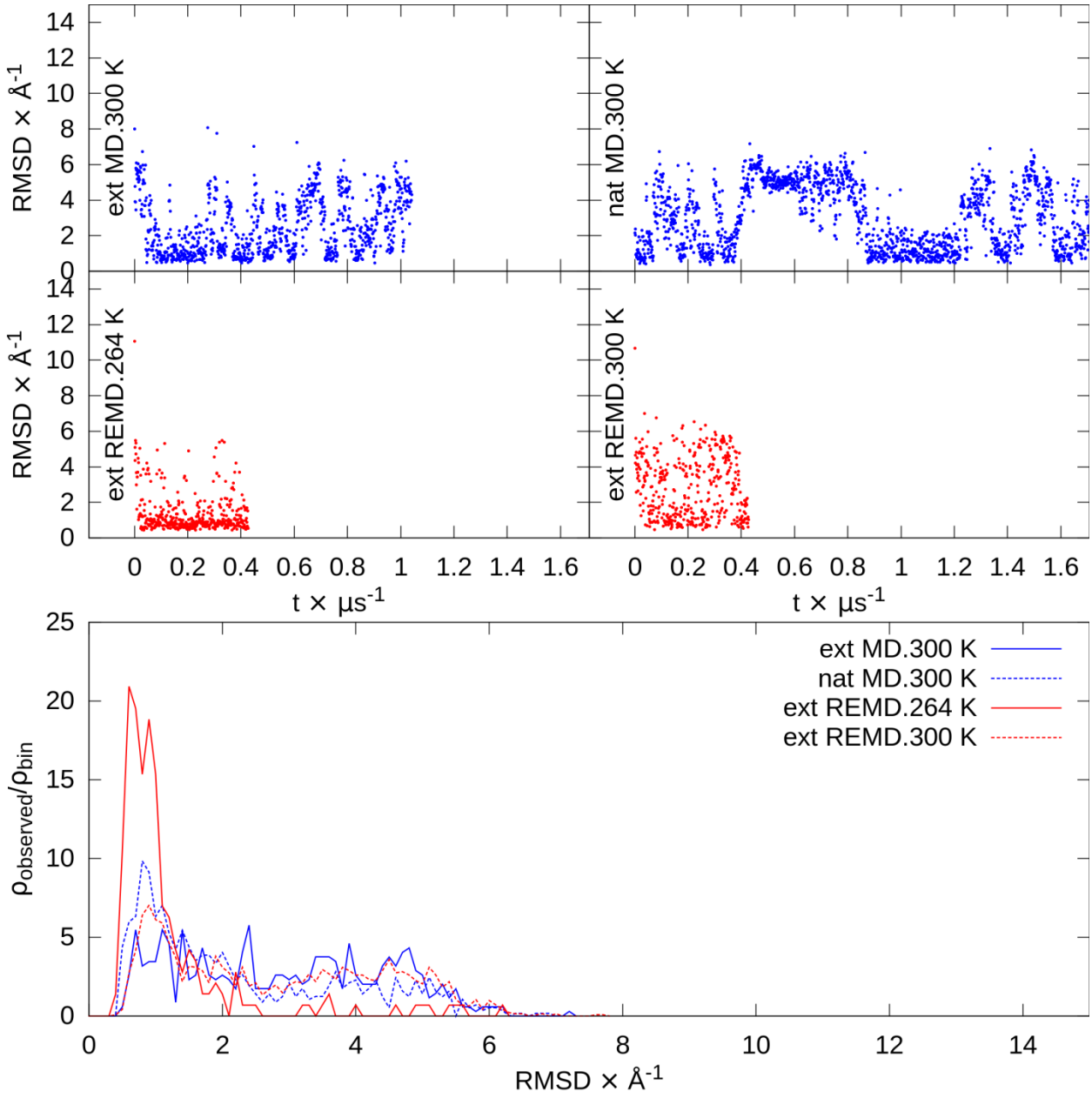


**Figure 3.S7.** Trp-cage RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

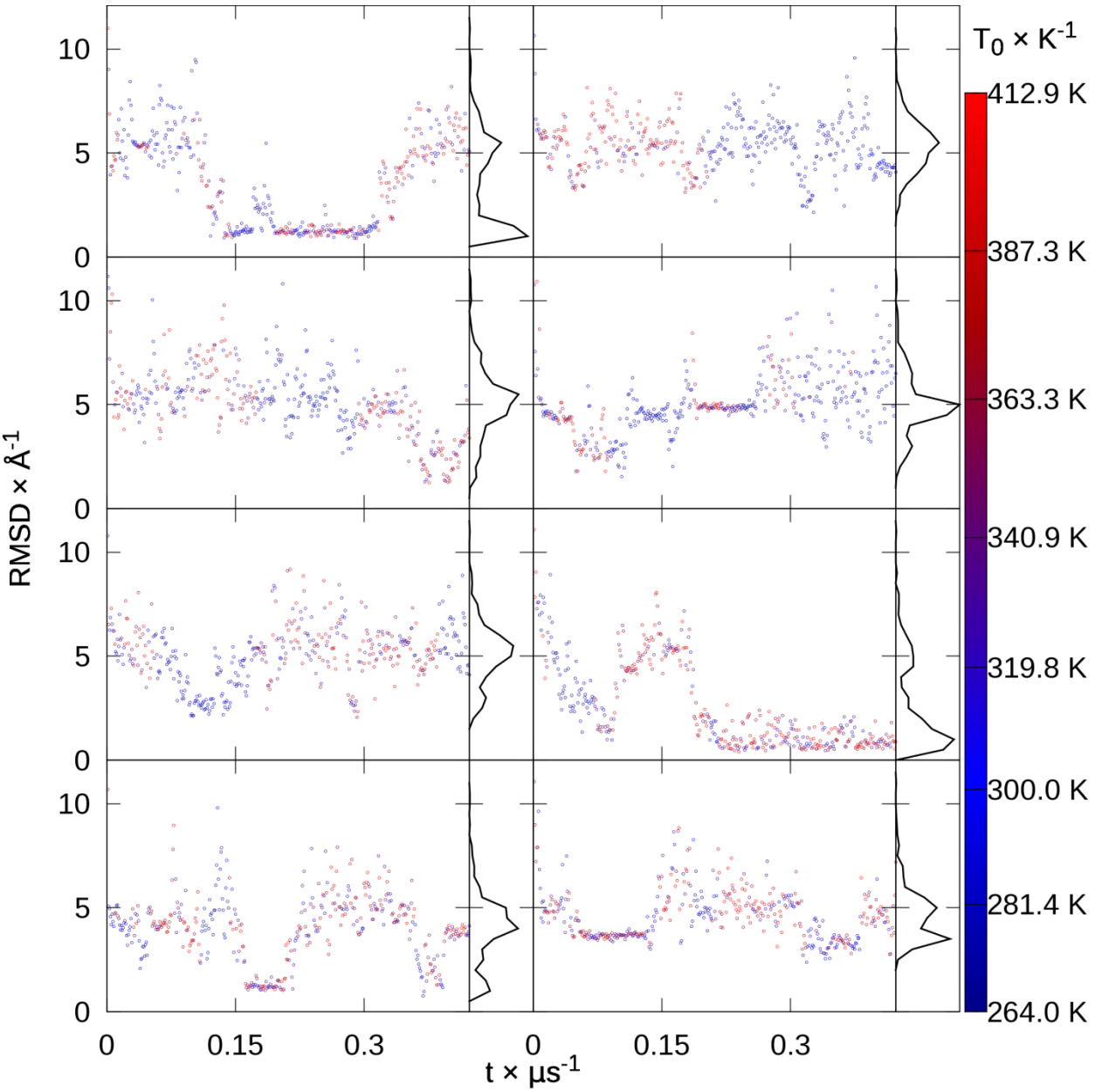


**Figure 3.S8.** Trp-cage replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 28.7 | 25.0 | 14.9 | 13.8 | 8.1 |
| Centroid Cα | 0.7 | 0.5 | 0.7 | 1.6 | 0.9 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S6.** Trp-cage top 5 extended REMD cluster populations and centroid Cα RMSDs.

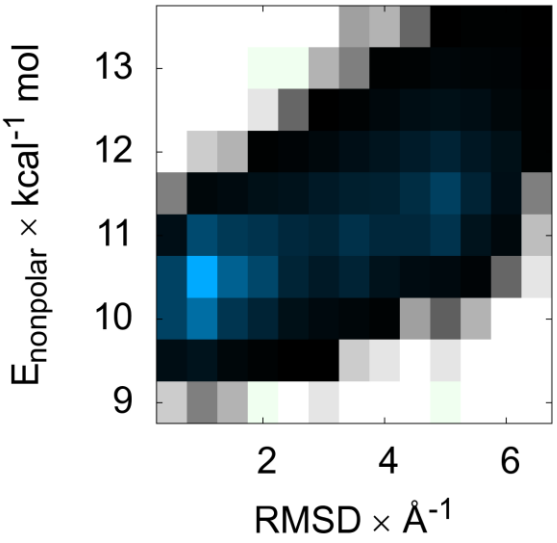


**Figure 3.S9.** Trp-cage surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal $mol^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is similarly to slightly more favorable at low (around 1 Å) than medium (3 -5 Å) RMSDs.

**BBA**

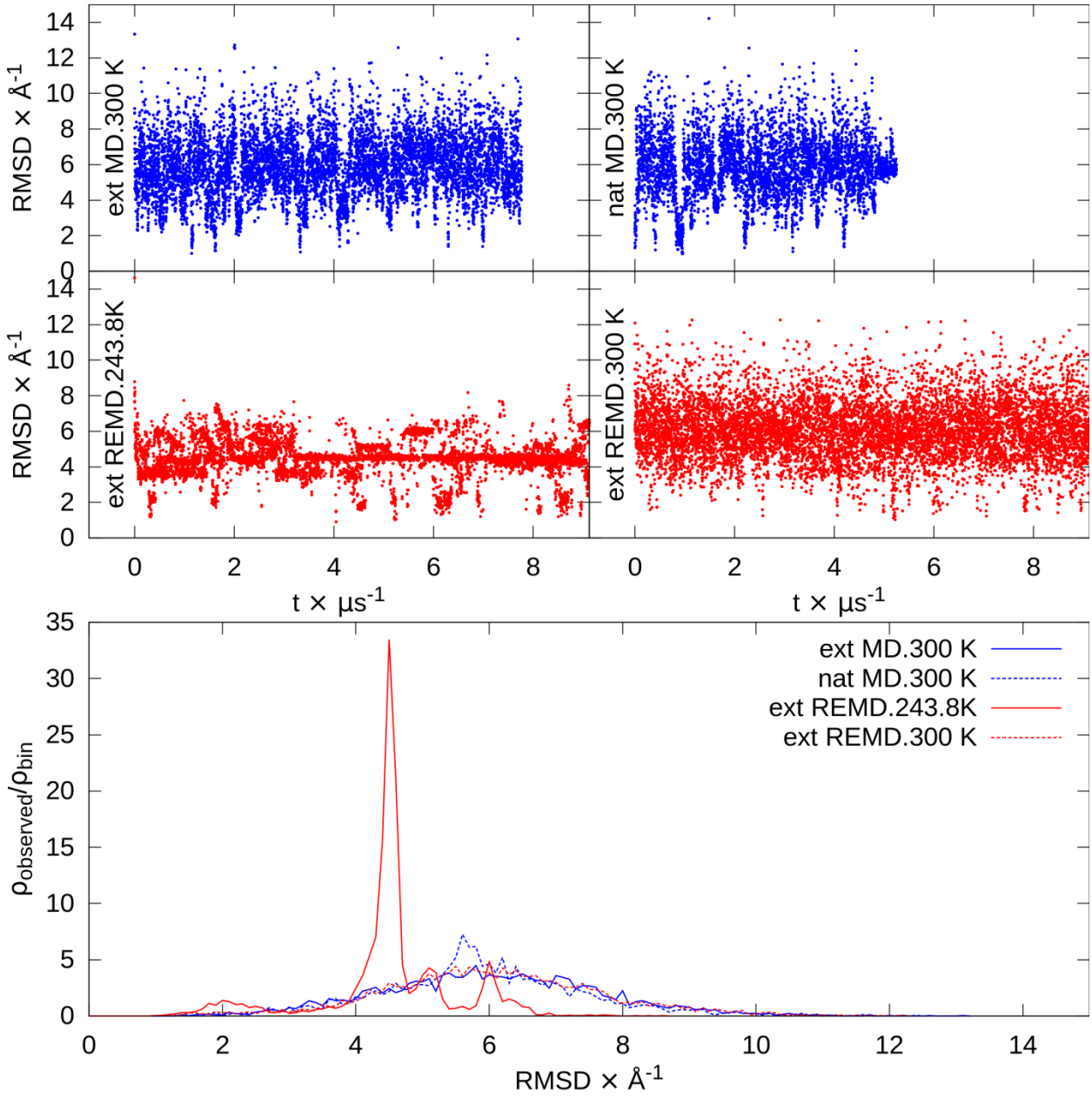


**Figure 3.S10.** BBA RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

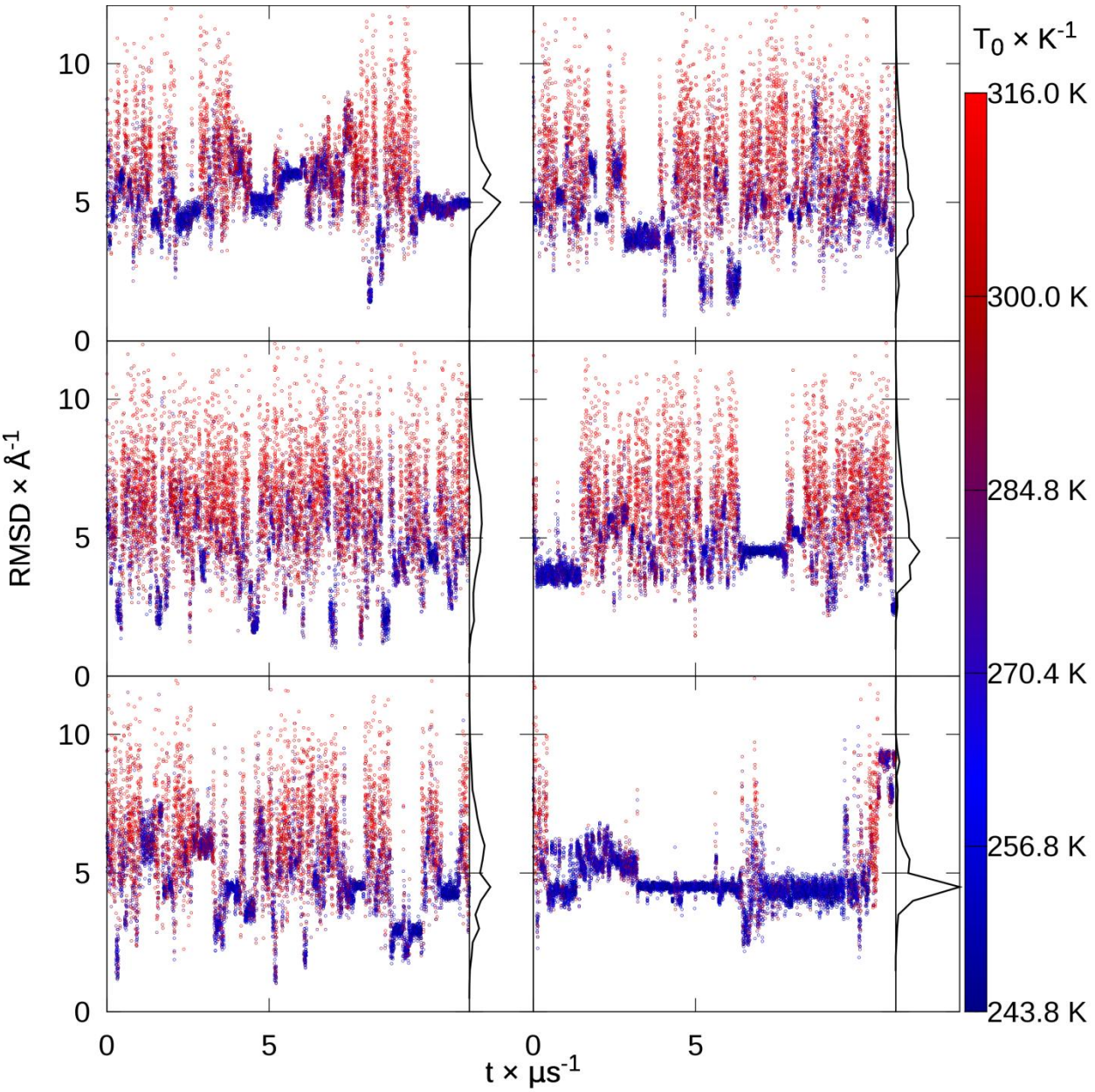


**Figure 3.S11.** BBA replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 34.8 | 11.0 | 7.4 | 4.8 | 4.3 |
| Centroid Cα | 4.6 | 3.4 | 4.1 | 4.4 | 5.9 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S7.**BBA top 5 extended REMD cluster populations and centroid Cα RMSDs.

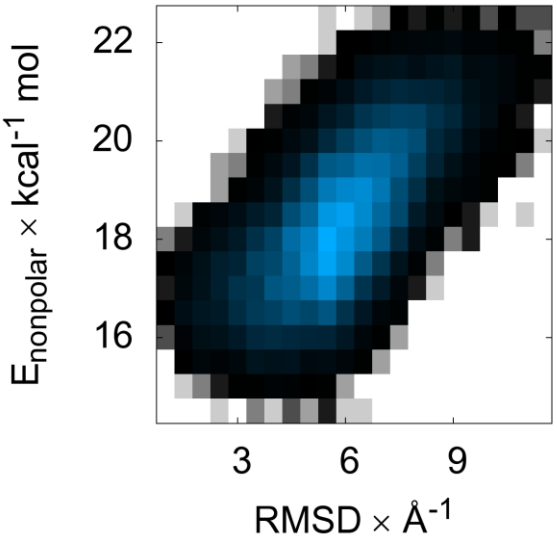


**Figure 3.S12.** BBA surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal $mol^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is similarly favorable at low (2 Å) to mid (6 Å) RMSDs, with no strong bias favoring low RMSD structures.

**Fip35**

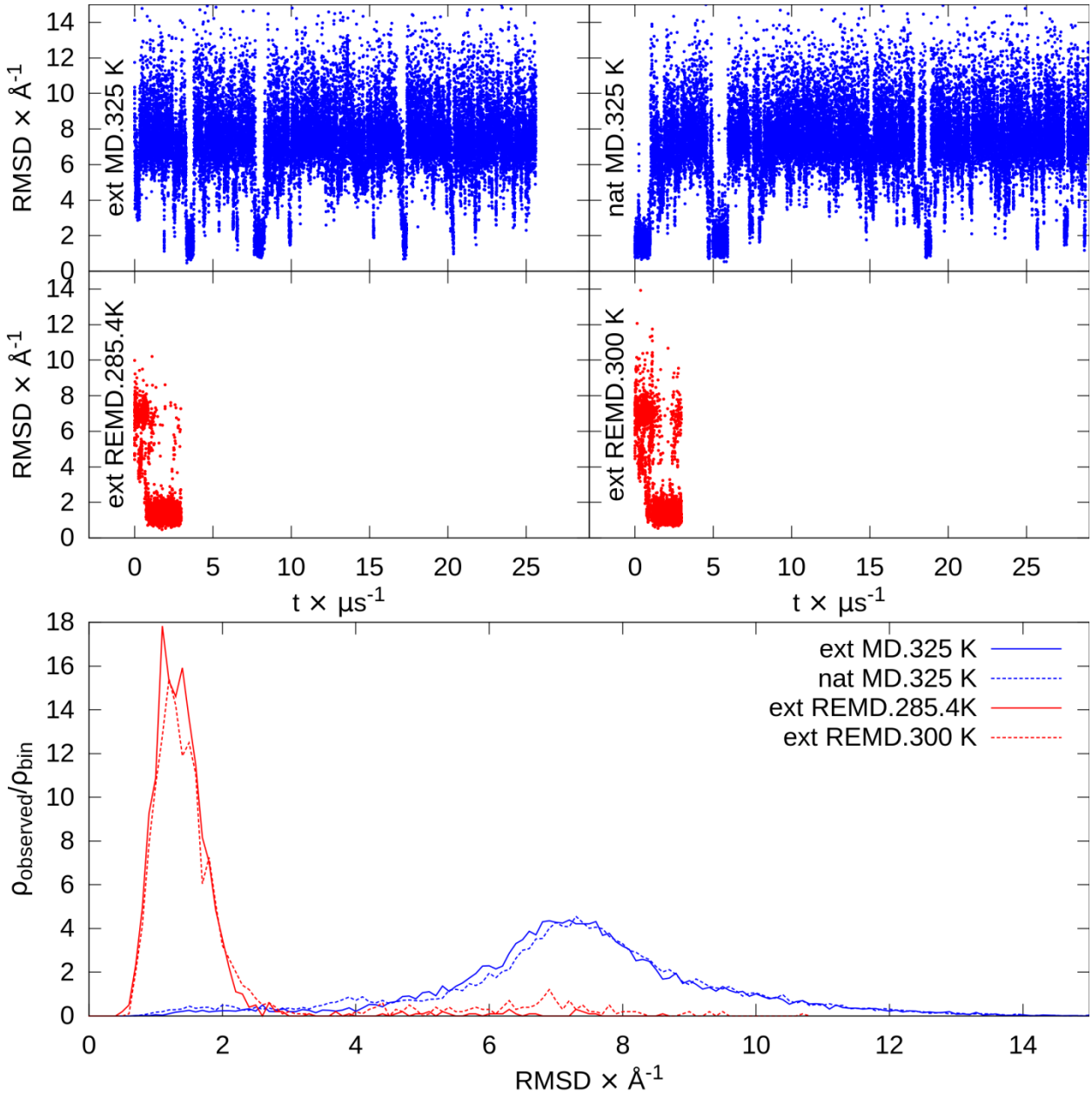


**Figure 3.S13**. Fip35 RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

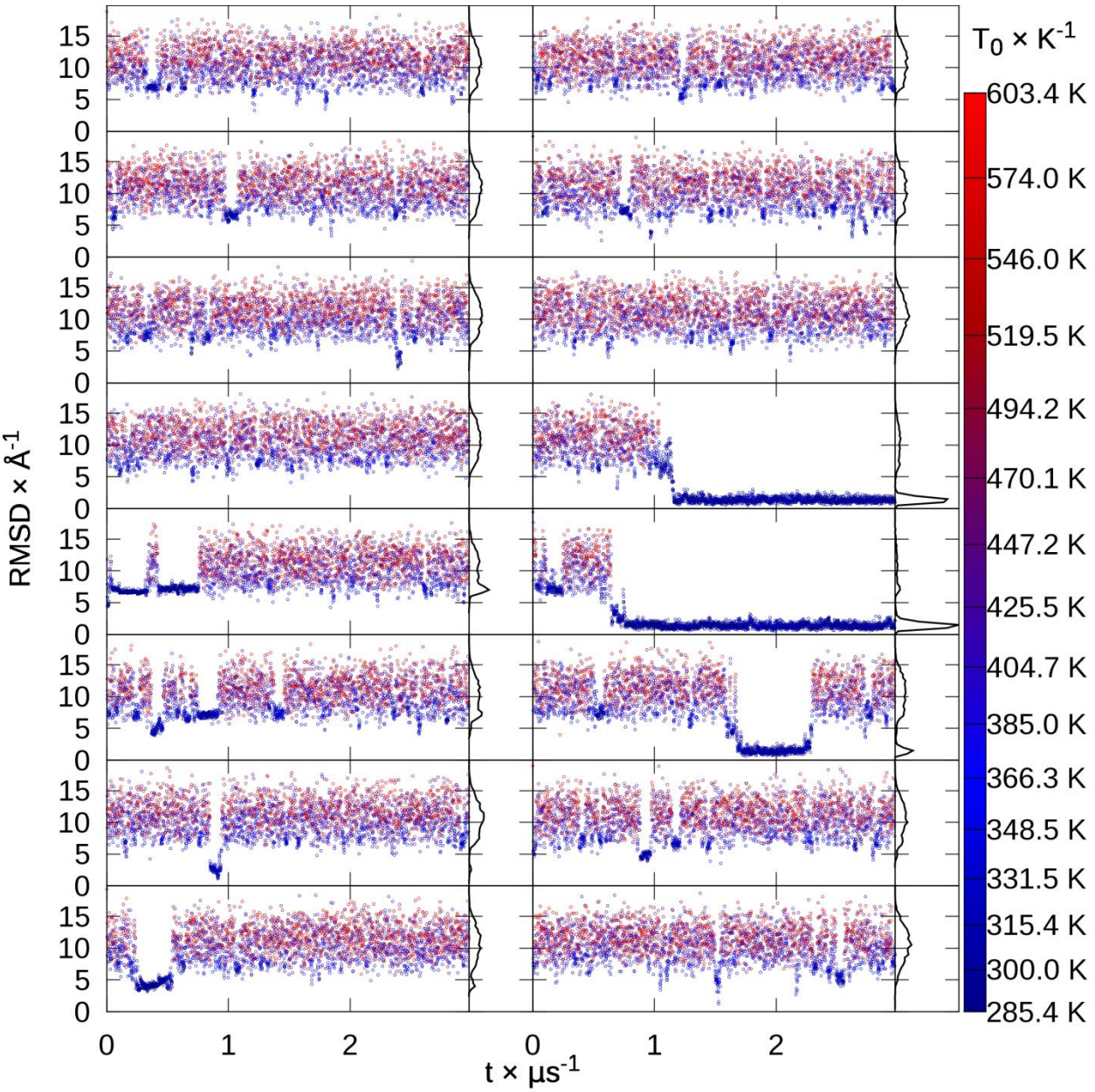


**Figure 3.S14.** Fip35 replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 70.7 | 7.5 | 4.6 | 4.3 | 2.4 |
| Centroid Cα | 1.3 | 7.1 | 4.2 | 6.6 | 7.1 |

| | | | | | |
|---|---|---|---|---|---|
| RMSD (Å) | | | | | |

**Table 3.S8.** Fip35 top 5 extended REMD cluster populations and centroid Cα RMSDs.

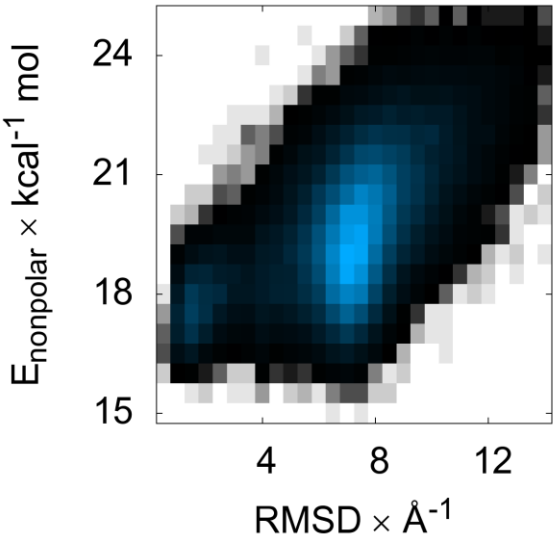


**Figure 3.S15.** Fip35 surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal $mol^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is similarly favorable from low (1 Å) to medium (8 Å) RMSDs, indicating no strong driving force toward low RMSD values.

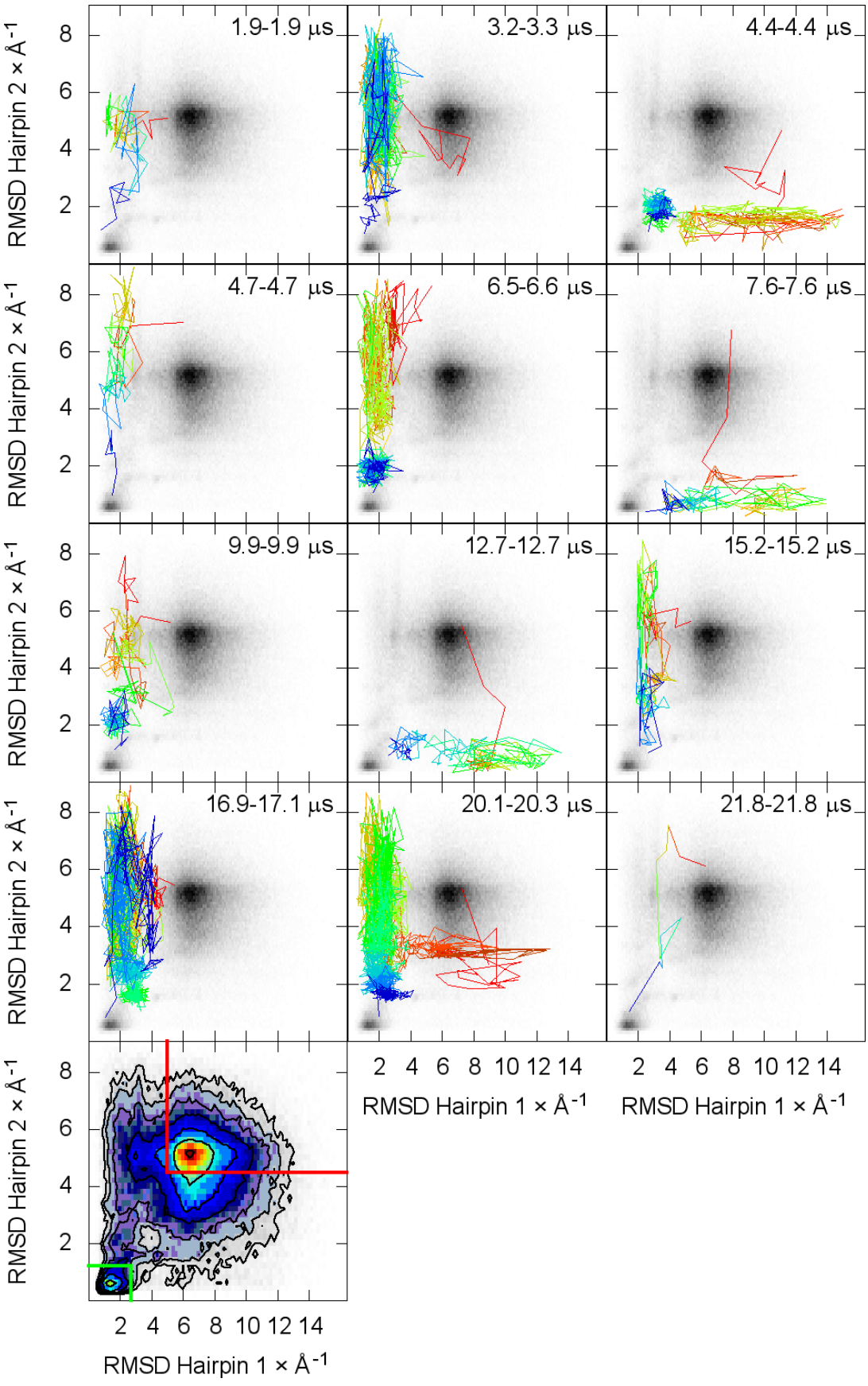
1.9-1.9 μs
3.2-3.3 μs
4.4-4.4 μs
4.7-4.7 μs
6.5-6.6 μs
7.6-7.6 μs
9.9-9.9 μs
12.7-12.7 μs
15.2-15.2 μs
16.9-17.1 μs
20.1-20.3 μs
21.8-21.8 μs
RMSD Hairpin 2 × Å-1
RMSD Hairpin 1 × Å-1
2 4 6 8 10 12 14
2 4 6 8

**Figure 3.S16.** Fip35 folding pathways and population histogram. The twelve unique folding pathways from fully unfolded to fully folded, as defined in Methods, are colored from red to yellow to green to cyan to blue. Eight proceed by folding of hairpin 1 first, two by folding of hairpin 2 first, and two by both simultaneously. At bottom, histogram with contours defining exponents of 2 shows two states in hairpin 1-hairpin 2 RMSD space, with unfolded boxed in red and folded in green. Trajectories are from the extended MD run shown in the top left corner of Figure 3.S13.

**GTT**

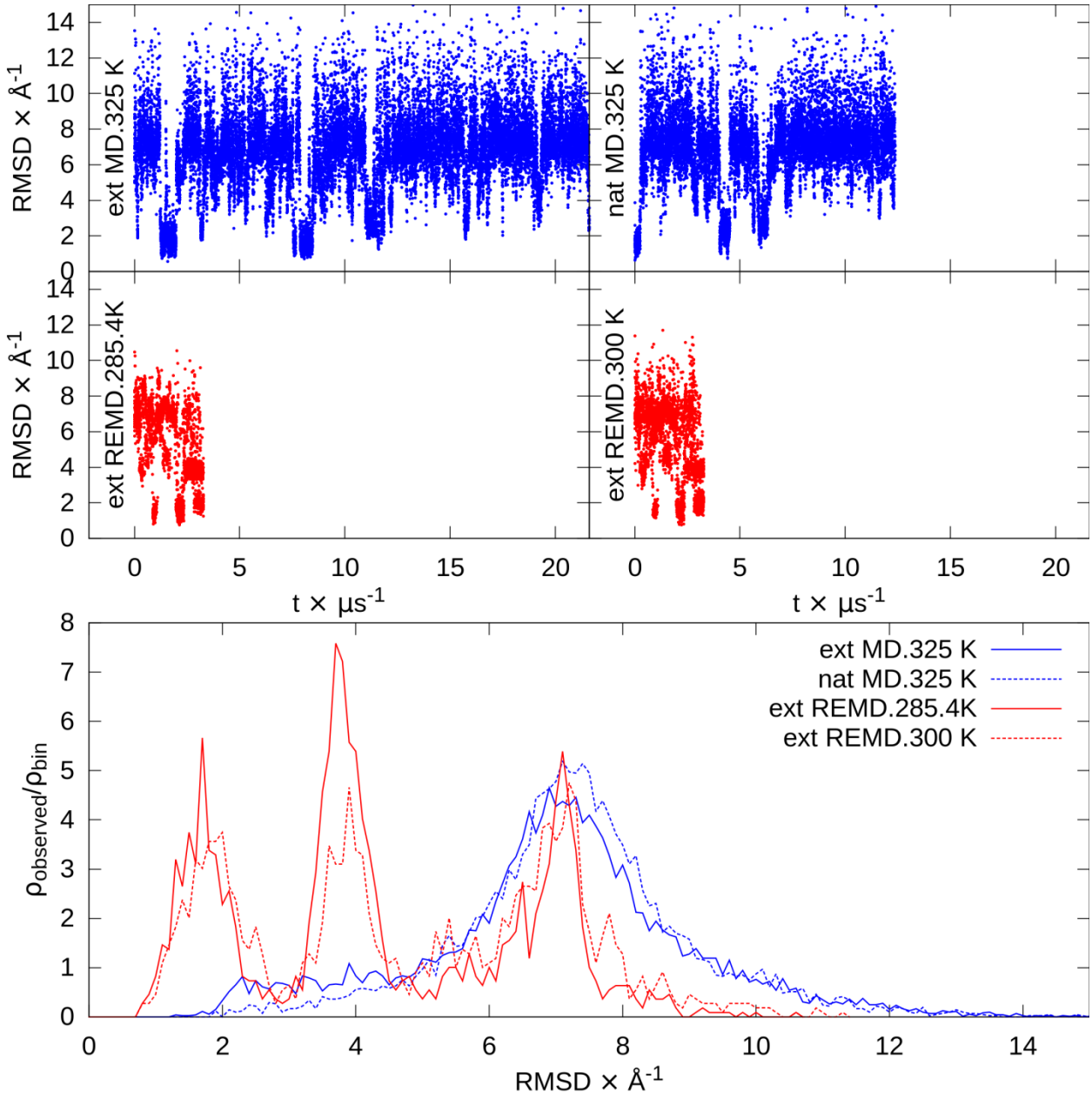


**Figure 3.S17**. GTT RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half

of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

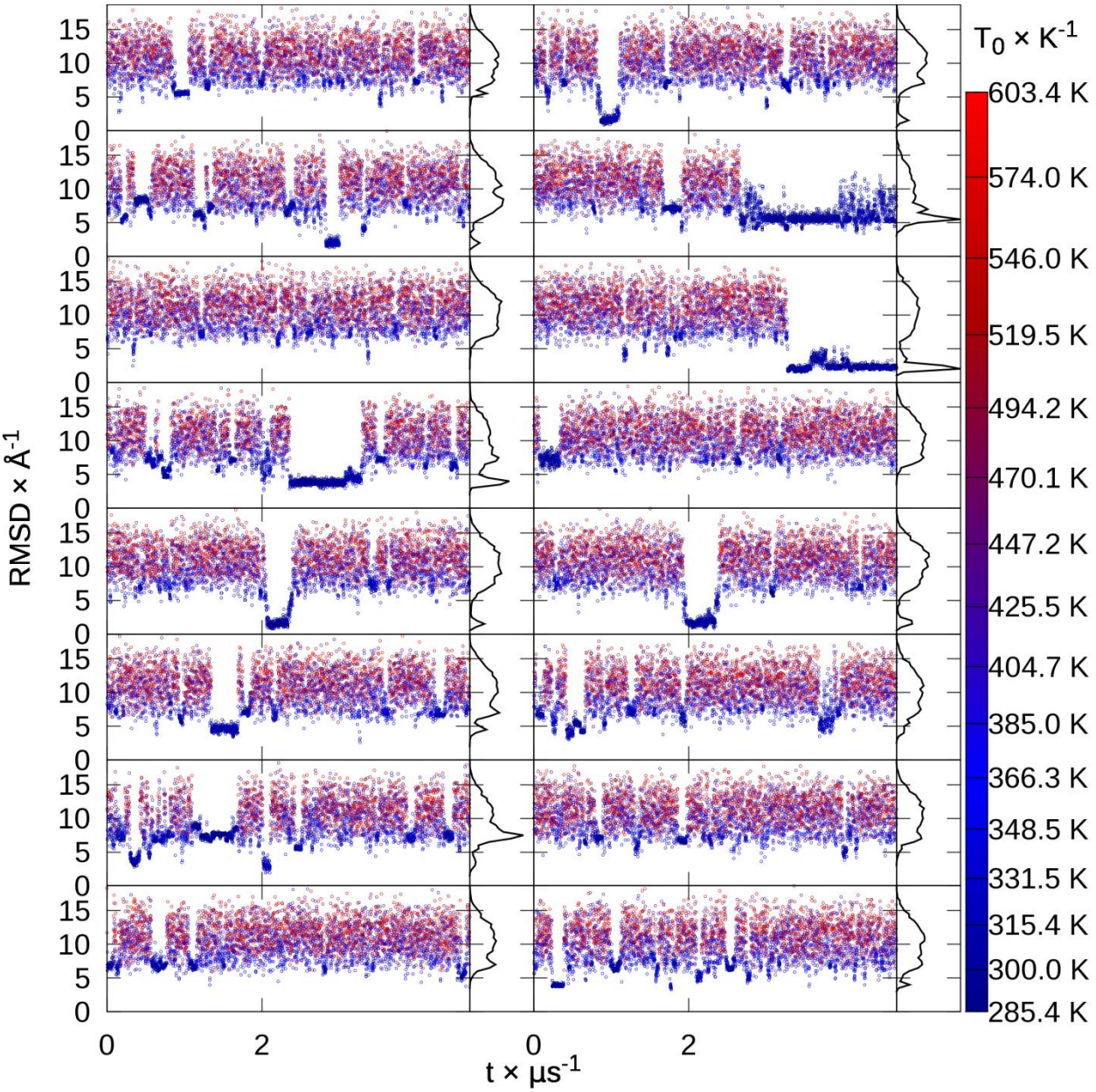


**Figure 3.S18.** GTT replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 14.7 | 12.9 | 9.3 | 5.1 | 4.1 |

| | | | | | |
|---|---|---|---|---|---|
| Centroid Cα RMSD (Å) | 1.4 | 3.8 | 7.3 | 7.2 | 7.2 |

**Table 3.S9.** GTT top 5 extended REMD cluster populations and centroid Cα RMSDs.

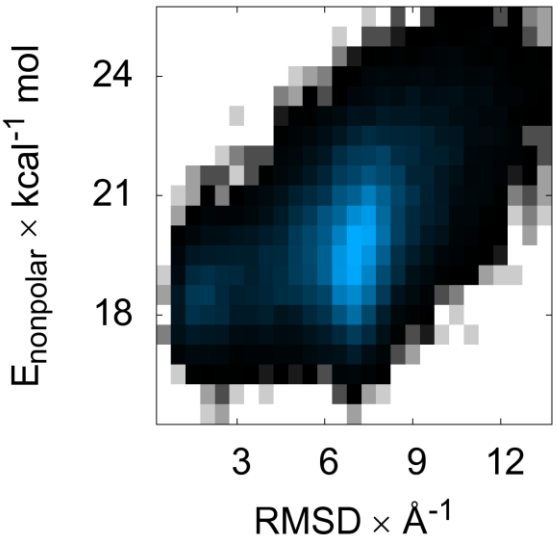


**Figure 3.S19.** GTT surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is flat from mid (6–8 Å) to low (2-3 Å) RMSD.

**Villin HP36**

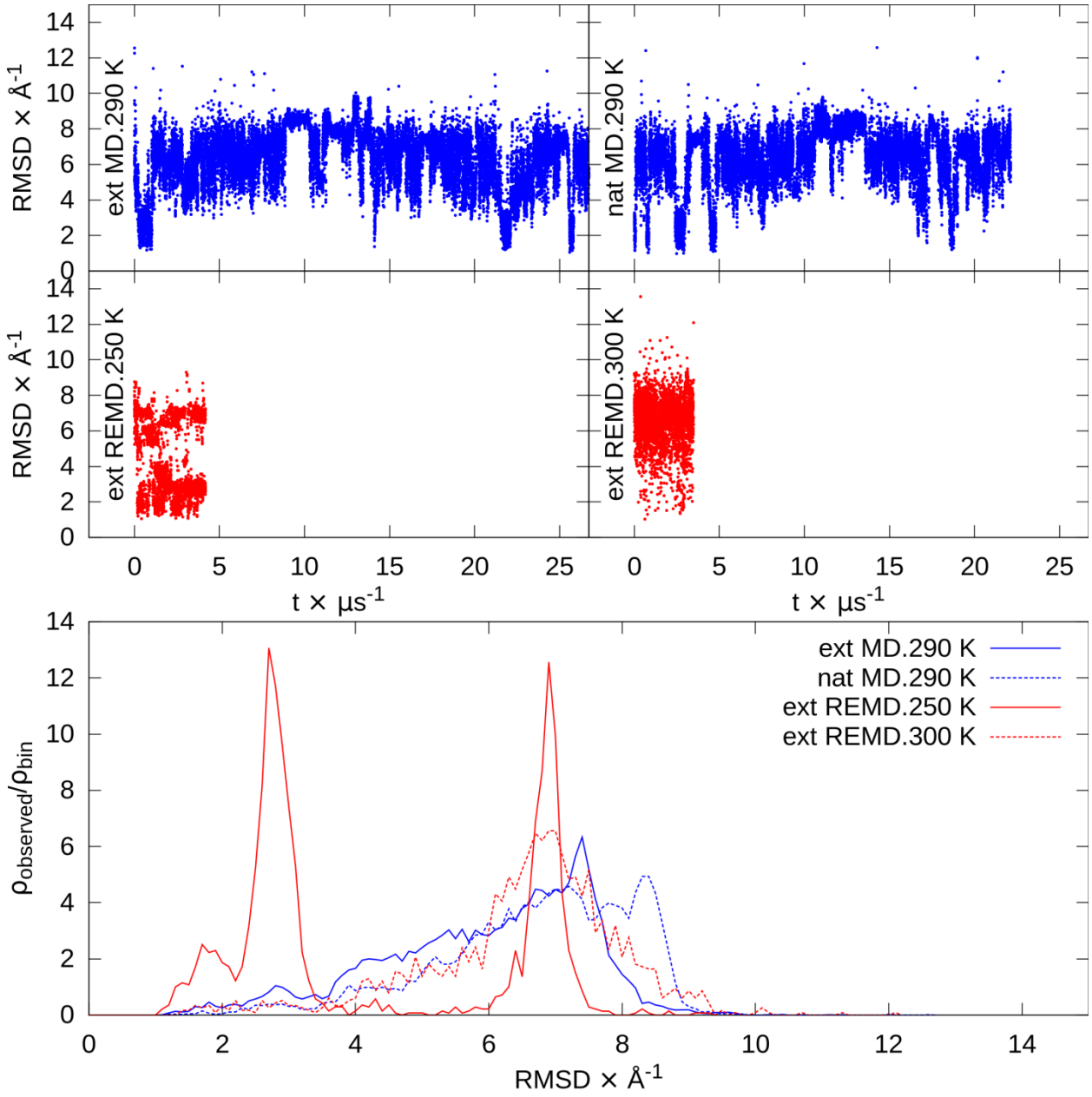


**Figure 3.S20.** HP36 RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

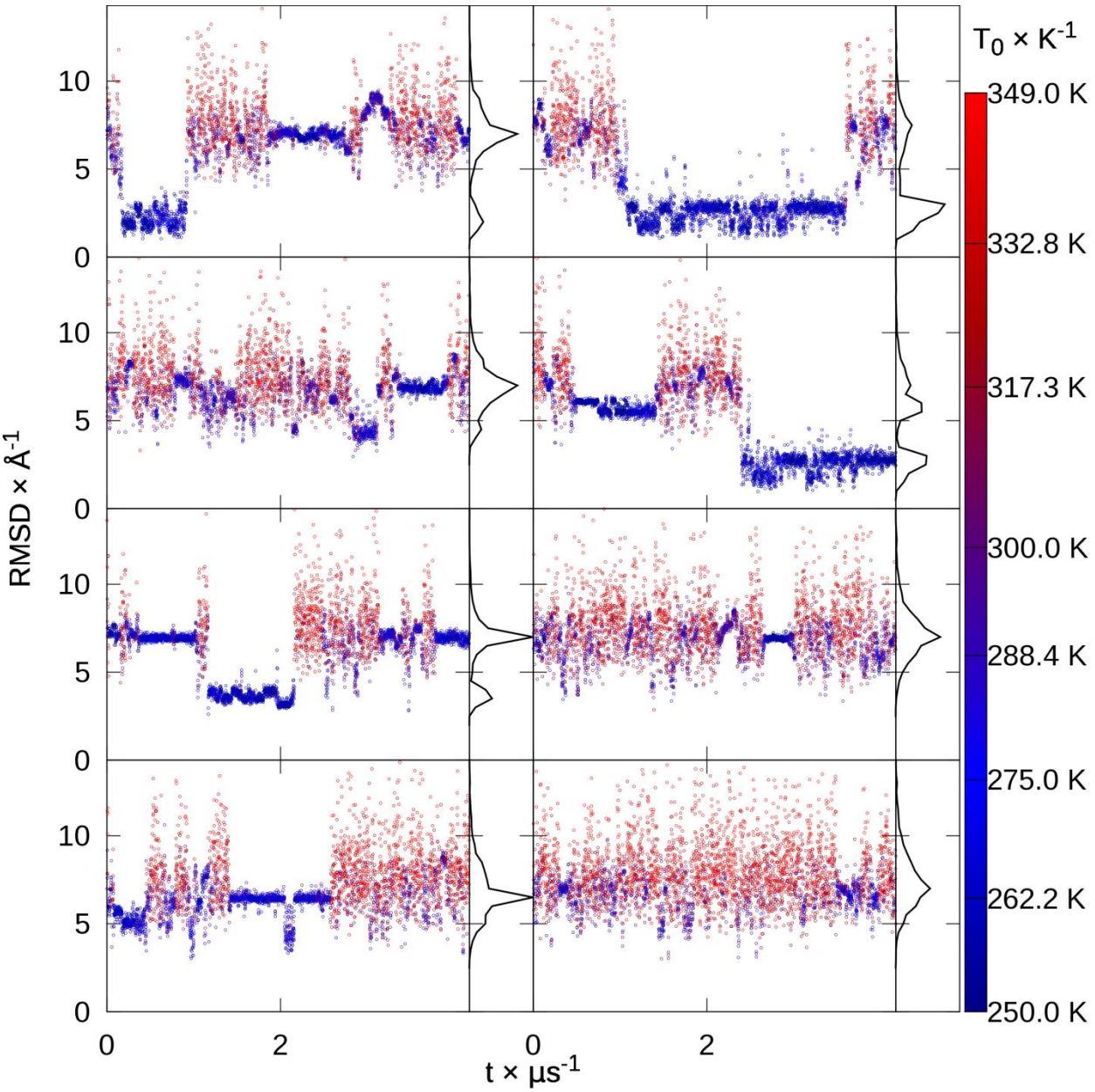


**Figure 3.S21.** Villin HP36 replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 43.3 | 10.8 | 10.3 | 10.2 | 4.5 |
| Centroid Cα | 2.3 | 5.5 | 3.5 | 6.9 | 6.9 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S10.** Villin HP36 top 5 extended REMD cluster populations and centroid Cα RMSDs.

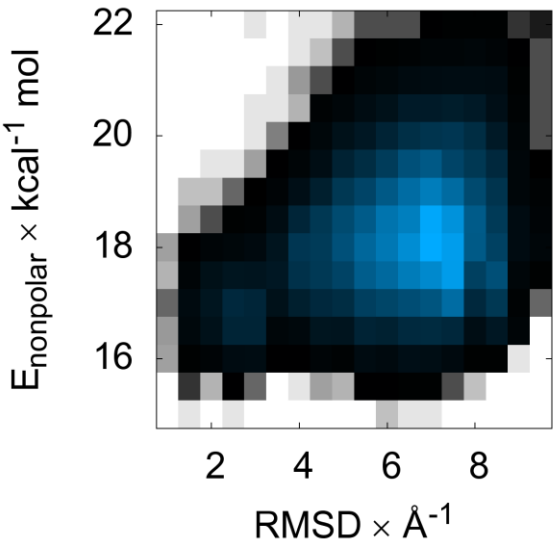


**Figure 3.S22.** Villin HP36 surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal $mol^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is flat from mid (6–8 Å) to low (1–3 Å) RMSD.

**NTL9 (39 AA)**

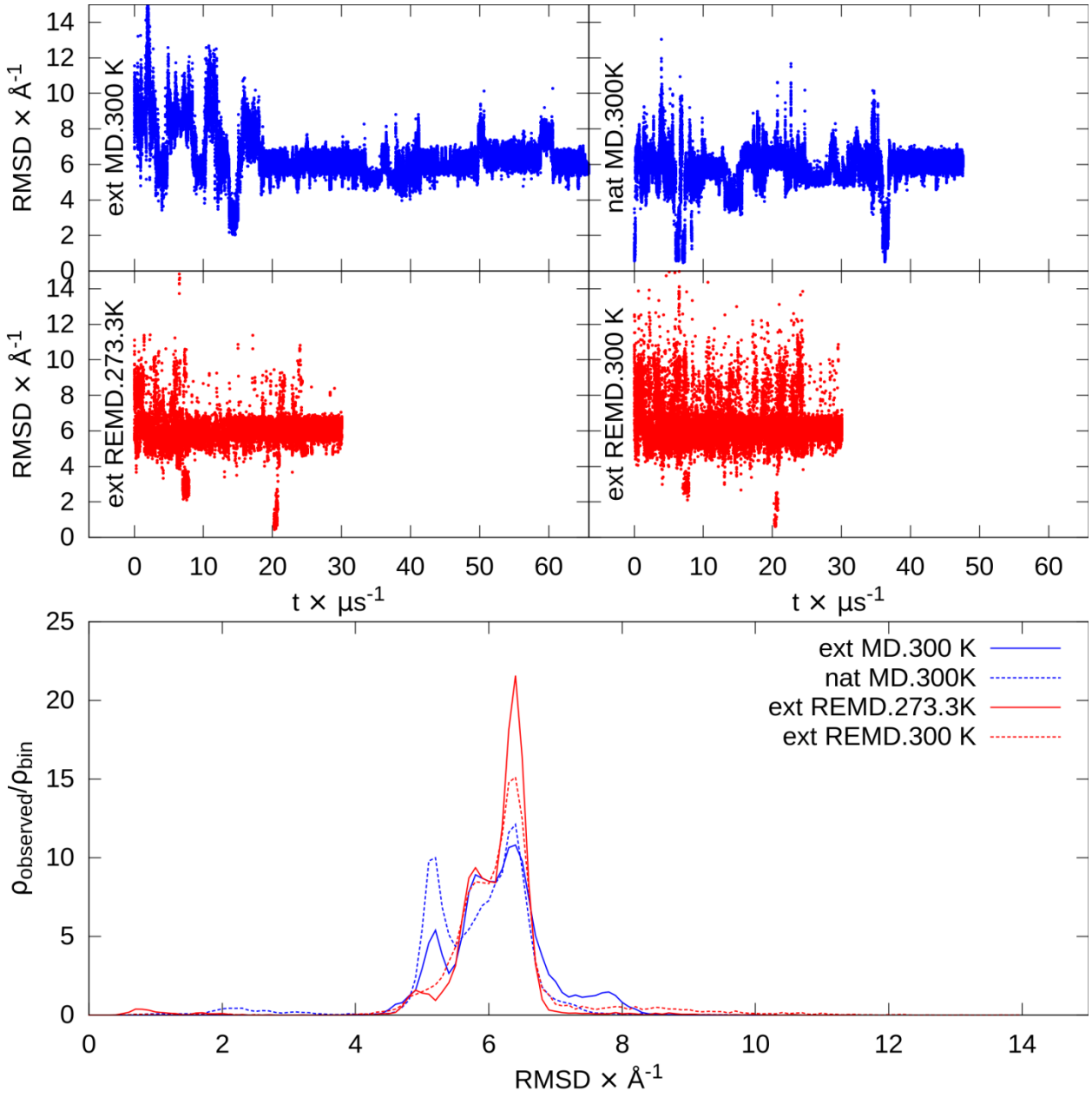


**Figure 3.S23.** NTL (39 AA) RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

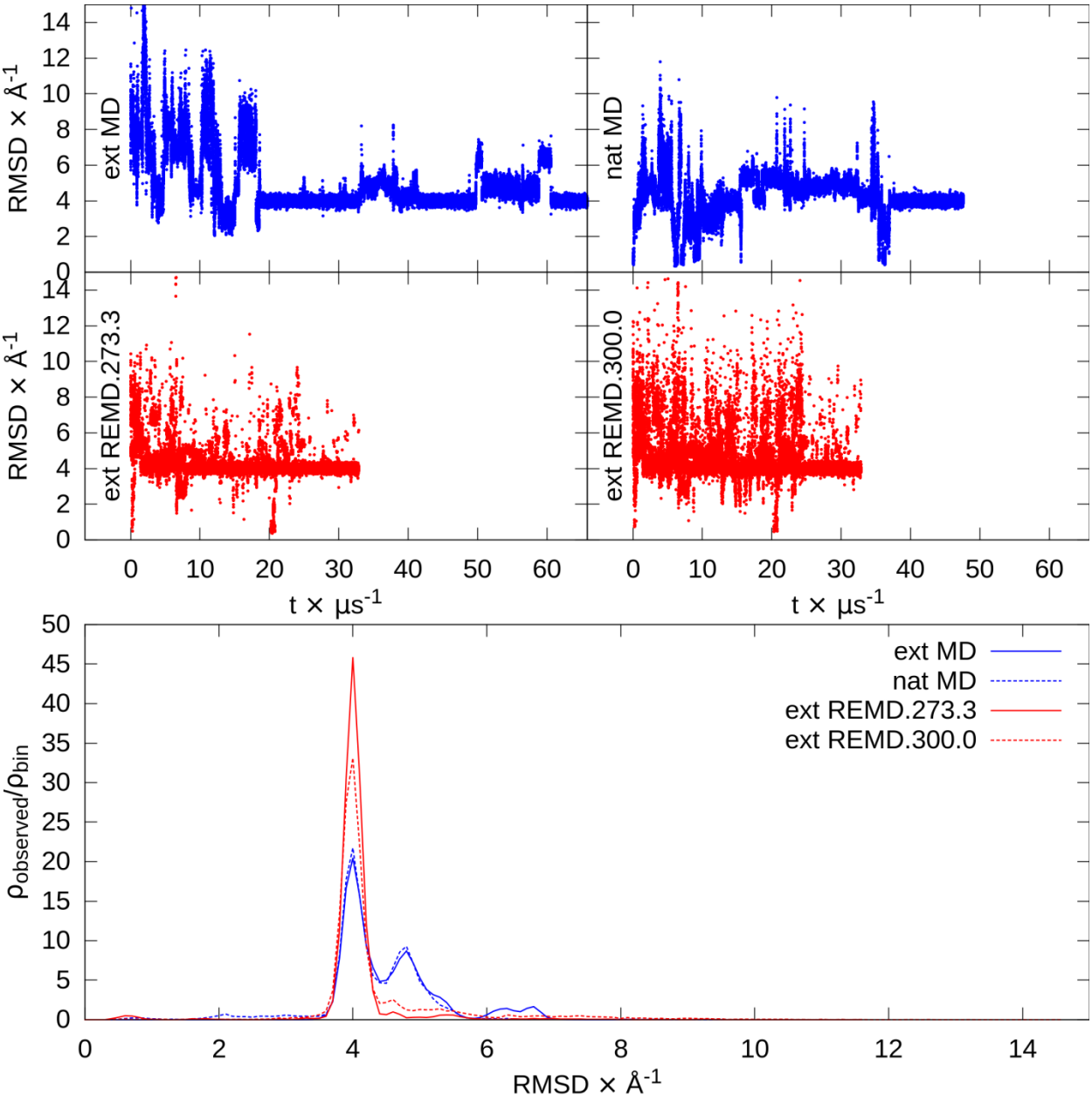


**Figure 3.S24**. NTL (39 AA) RMSDs, excluding the 7-16 loop described in the main text. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

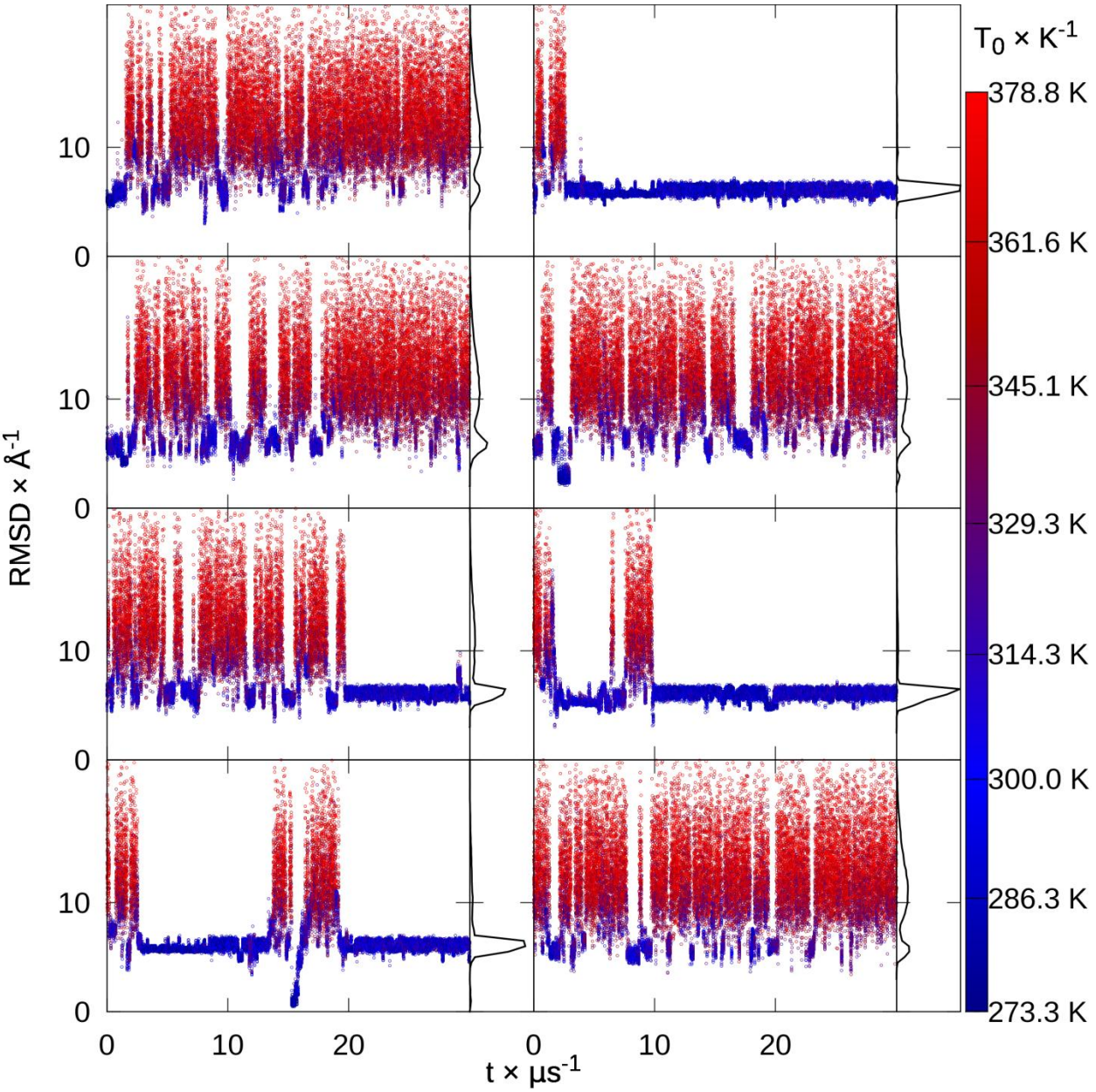


**Figure 3.S25**. NTL9 (39 AA) replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 68.1 | 8.9 | 6.8 | 5.5 | 2.2 |
| Centroid Cα | 6.1 | 5.9 | 6.0 | 4.6 | 5.4 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S11.** NTL9 (39 AA) top 5 extended REMD cluster populations and centroid Cα RMSDs.

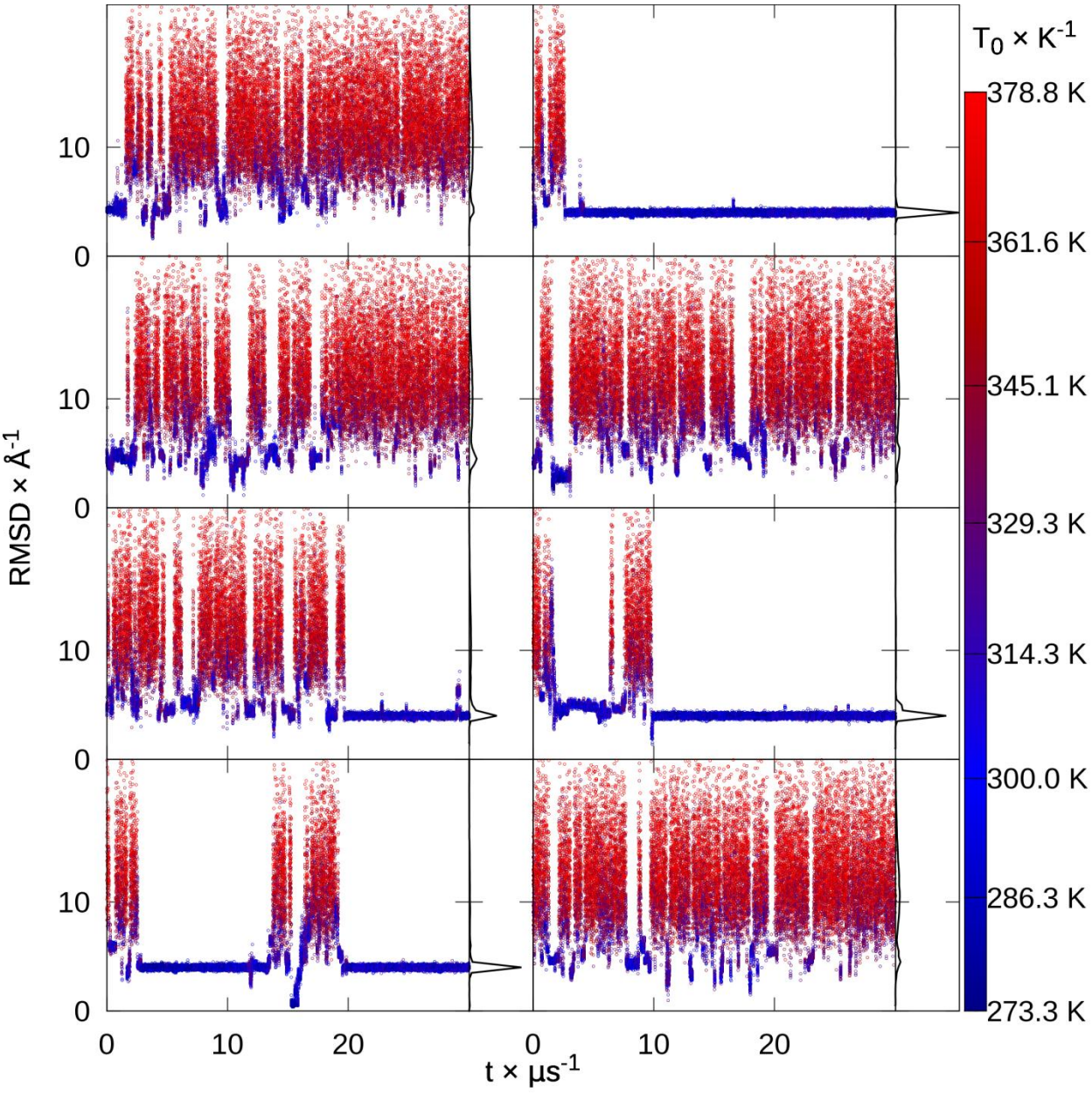


**Figure 3.S26.** NTL9 (39 AA) replica RMSDs, excluding the 7-16 loop. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

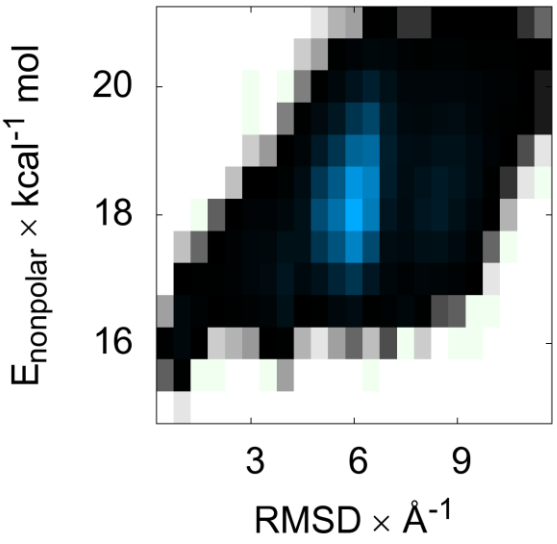


**Figure 3.S27.** NTL9 (39 AA) surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is more favorable at low (~1 Å) RMSD.

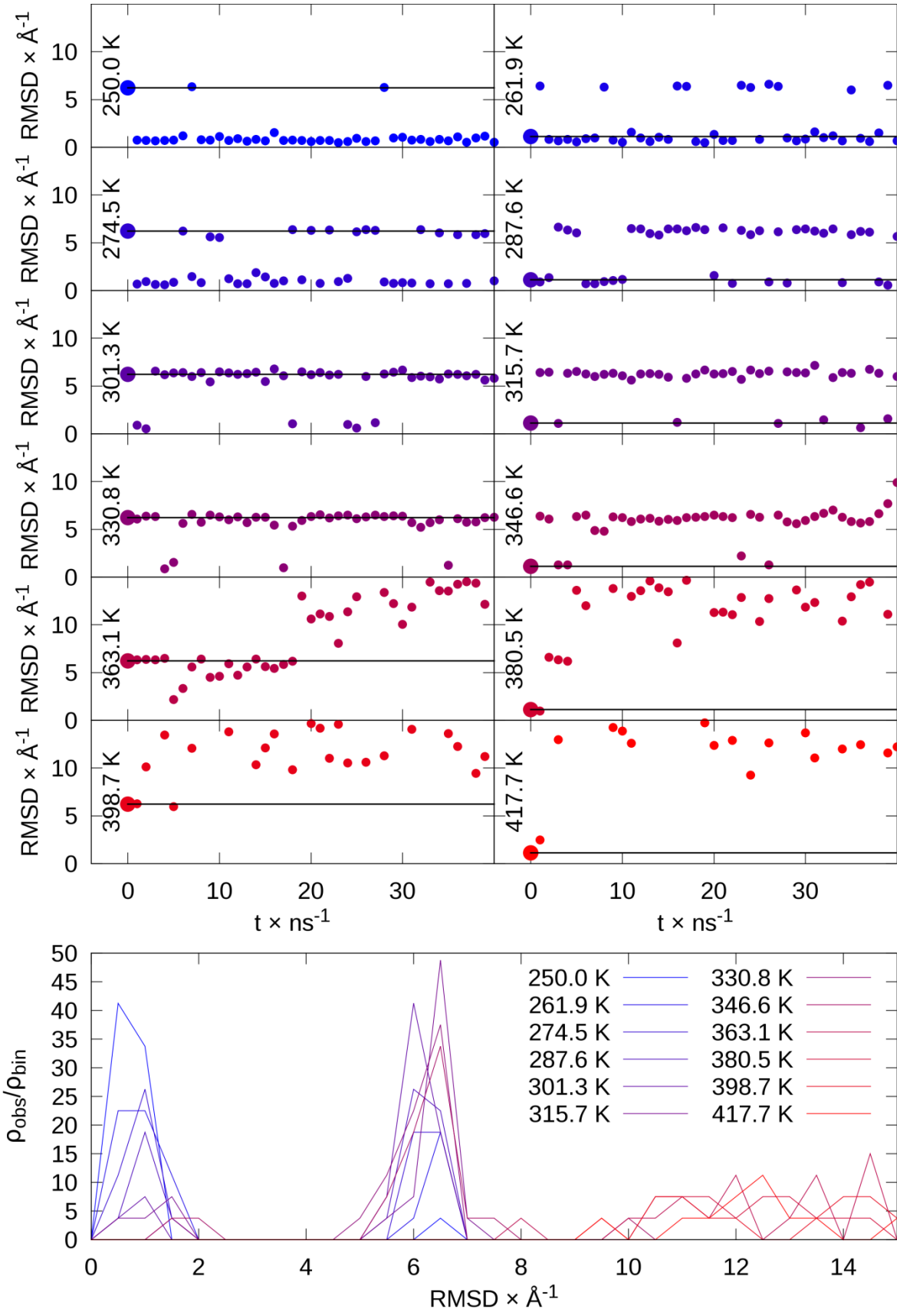
250.0 K
261.9 K
274.5 K
287.6 K
301.3 K
315.7 K
330.8 K
346.6 K
363.1 K
380.5 K
398.7 K
417.7 K
RMSD × Å-1
t × ns-1
ρobs/ρbin
RMSD × Å-1

**Table 3.S28**. Seeded REMD sorting of NTL9 (39 AA) conformations: partly unfolded (6.2 Å) and native-like (1.1 Å), repeated for 12 replicas. At top, RMSD vs time for each temperature shows sorting of native-like conformations to the lowest temperatures, with partly unfolded structures mixing in by 287.6 K. The line indicates the initial rmsd sampled at that temperature. At bottom, histograms show preference of low RMSD conformations at the lowest temperatures, with preference of the partly unfolded conformation beginning between 287.6 and 301.3 K.

**BBL**

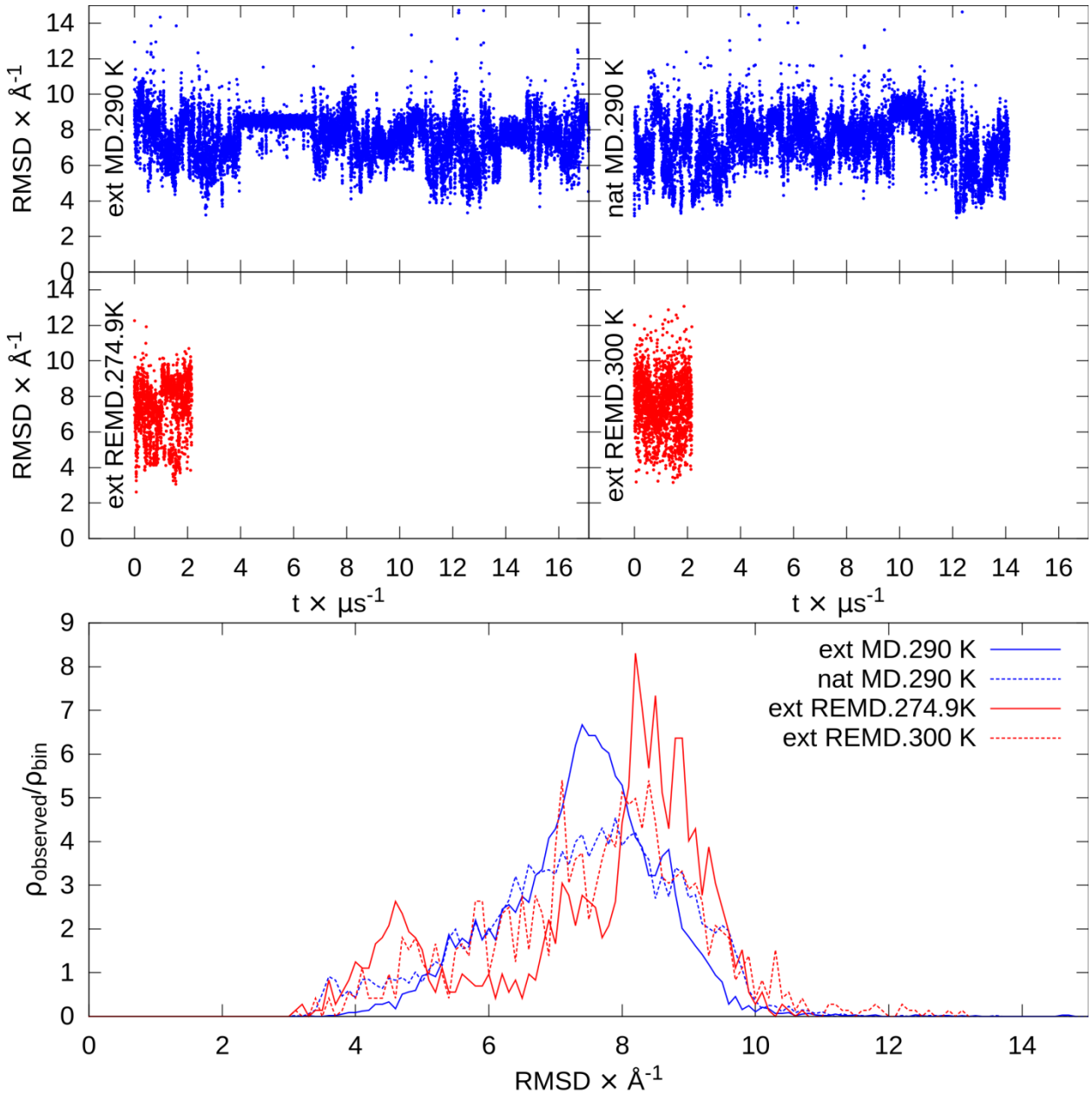


**Figure 3.S26.** BBL RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half

of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

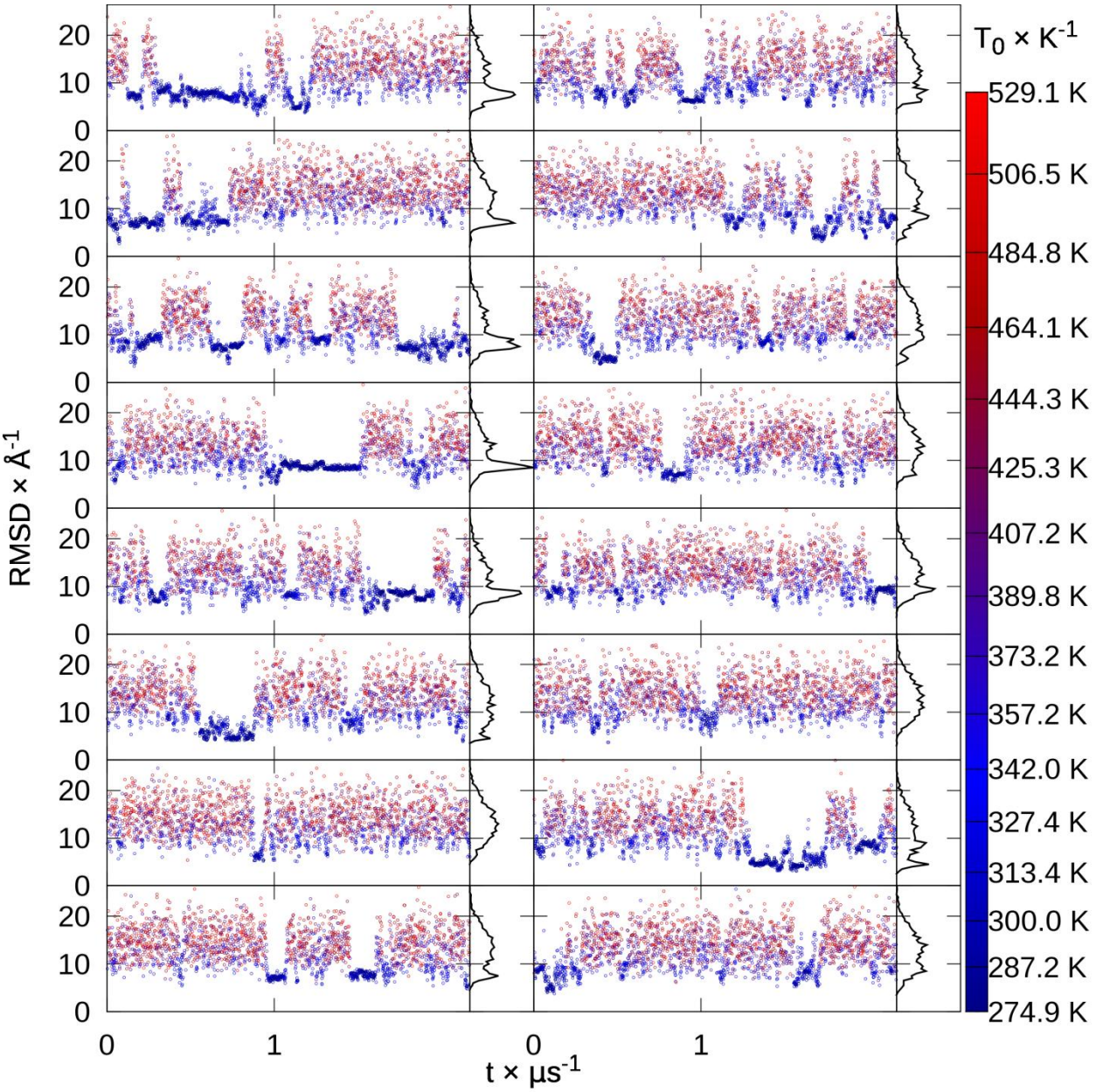


**Figure 3.S27.** BBL replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 8.4 | 6.9 | 4.9 | 4.8 | 4.4 |

| | | | | | |
|---|---|---|---|---|---|
| Centroid Cα RMSD (Å) | 8.3 | 4.3 | 4.8 | 8.2 | 9.3 |

**Table 3.S12.** BBL top 5 extended REMD cluster populations and centroid Cα RMSDs.

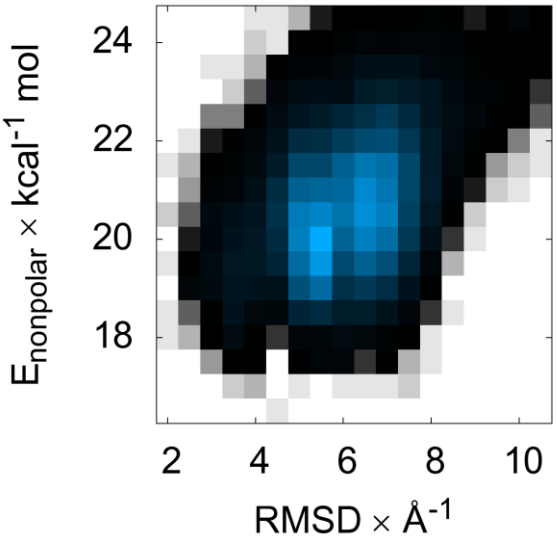


**Figure 3.S28.** BBL surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is flat from mid (6–7 Å) to mid-low (3–4 Å) RMSD.

**Protein B**

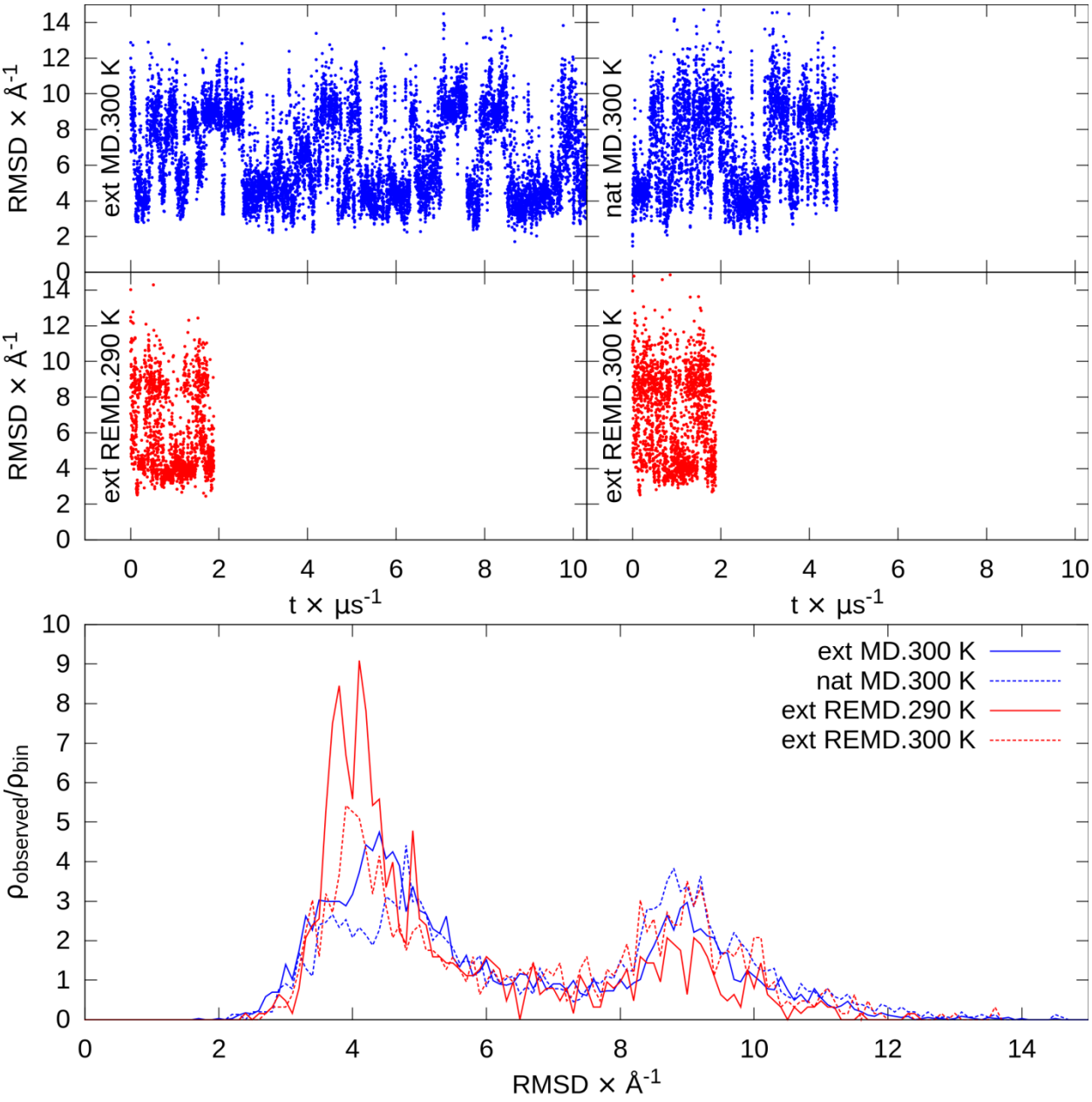


**Figure 3.S29.** Protein B RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

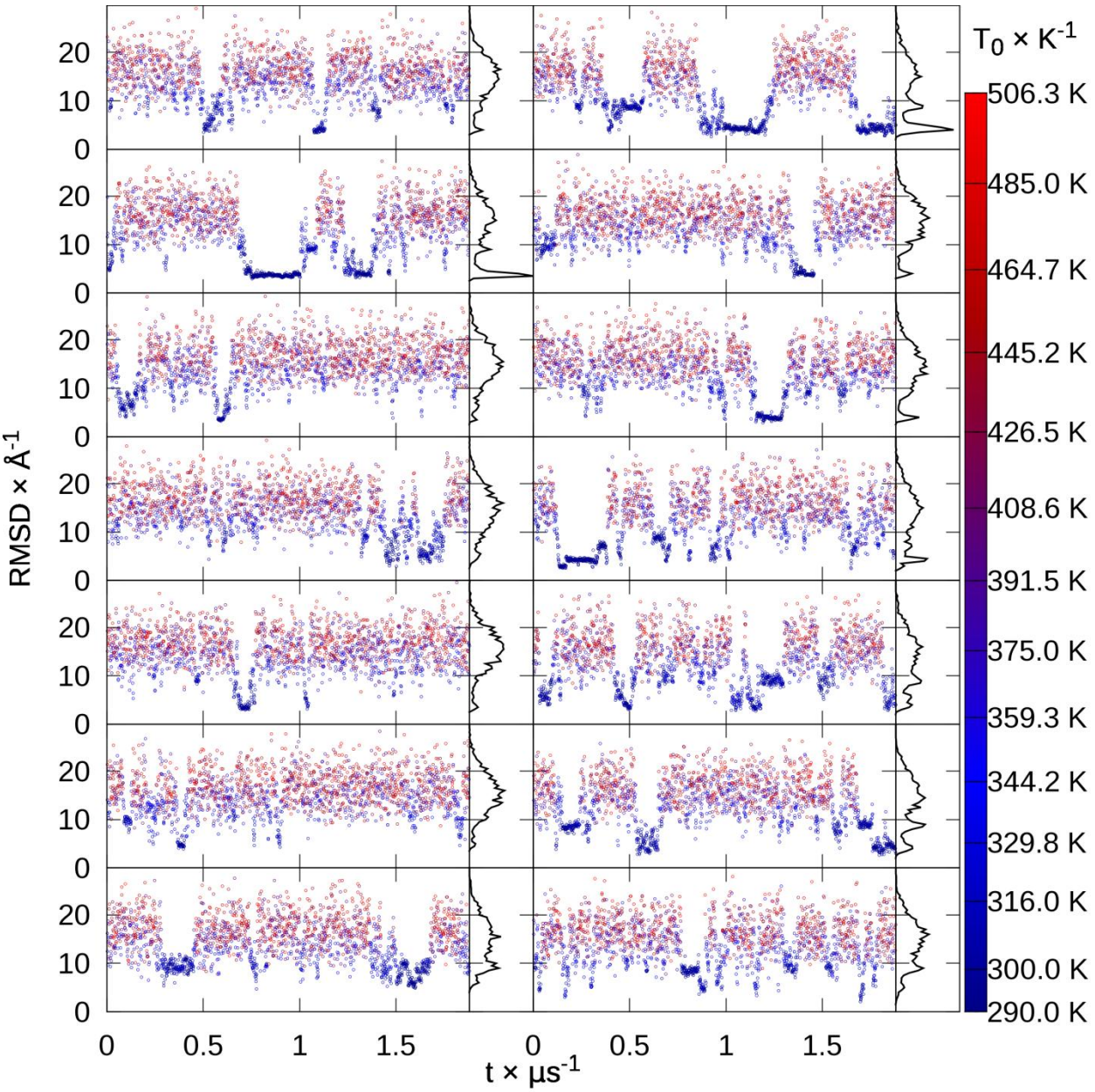


**Figure 3.S30.** Protein B replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 18.6 | 13.9 | 9.1 | 4.6 | 4.6 |
| Centroid Cα | 4.2 | 3.4 | 2.7 | 3.8 | 3.4 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S13.** Protein B top 5 extended REMD cluster populations and centroid Cα RMSDs.

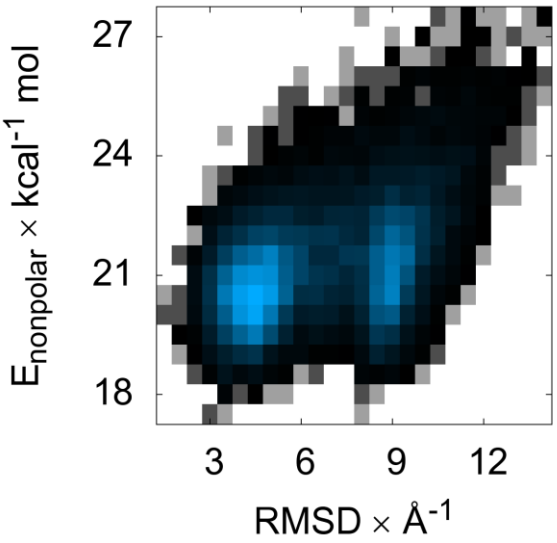


**Figure 3.S31.** Protein B surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is similarly favorable at low (2–4 Å) and mid-high (8–9 Å) RMSD.

**Engrailed homeodomain**

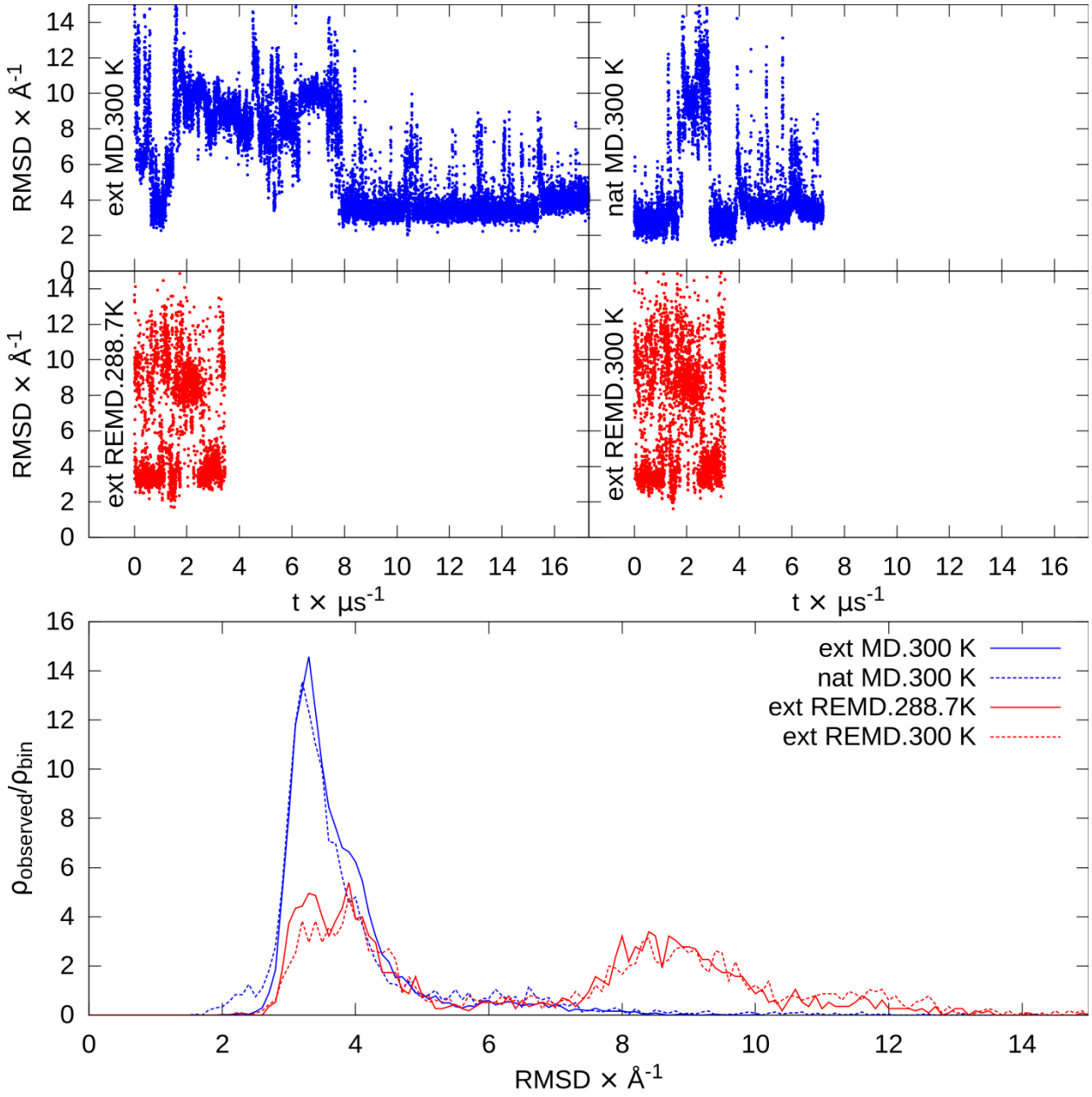


**Figure 3.S32.** Homeodomain RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

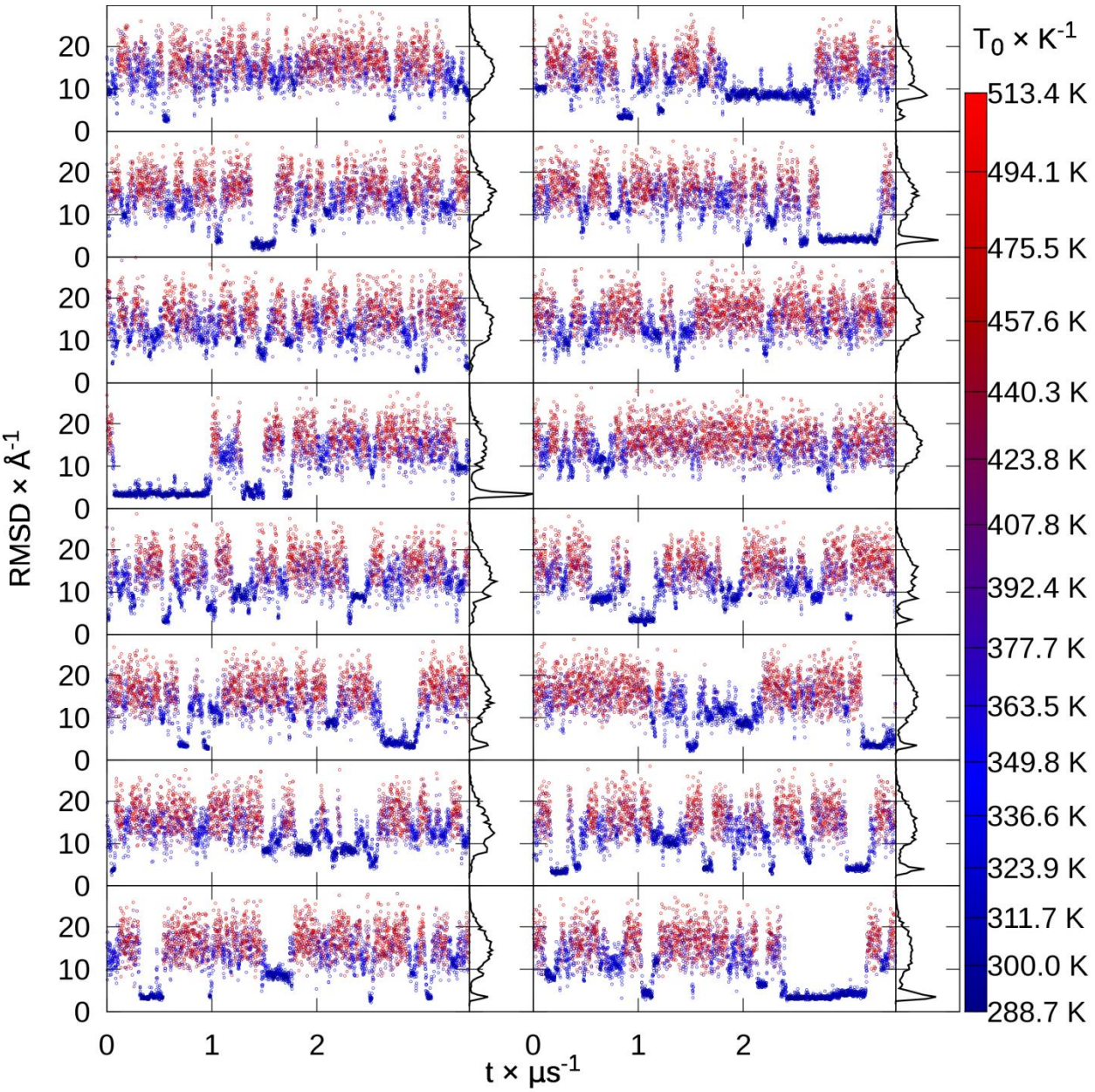


**Figure 3.S33.** Engrailed homedomain replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 22.5 | 7.9 | 7.7 | 6.7 | 5.6 |
| Centroid Cα | 3.2 | 2.3 | 3.9 | 3.1 | 7.8 |

| | | | | | |
|---|---|---|---|---|---|
| RMSD (Å) | | | | | |

**Table 3.S14.** Engrailed homeodomain top 5 extended REMD cluster populations and centroid Cα RMSDs.

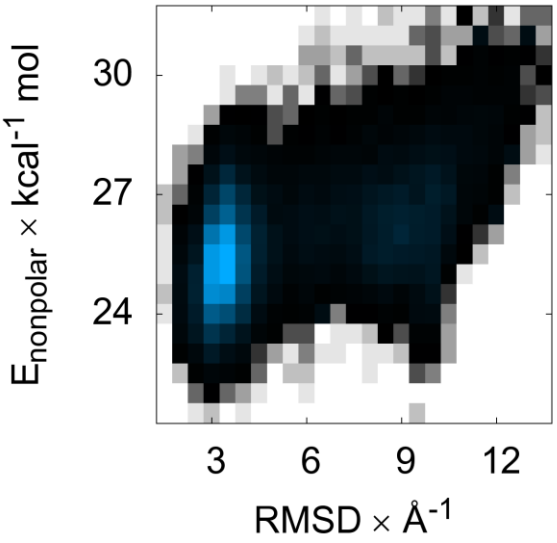


**Figure 3.S34.** Engrailed homeodomain surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is similarly favorable at low (2–4 Å) and high (9–11 Å) RMSDs.

**NTL9 (52 AA)**

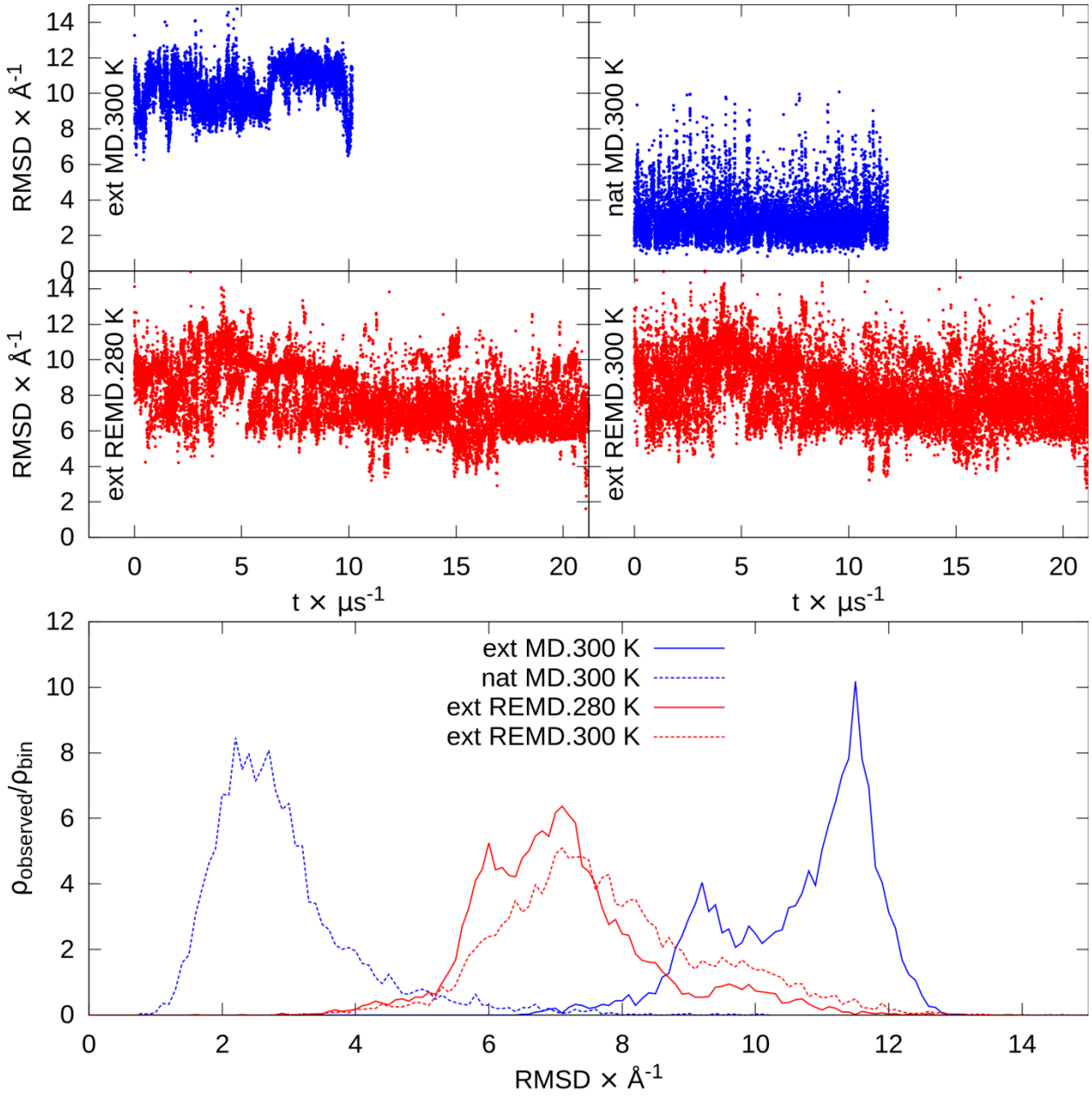


**Figure 3.S35.** NTL9 (52 AA) RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

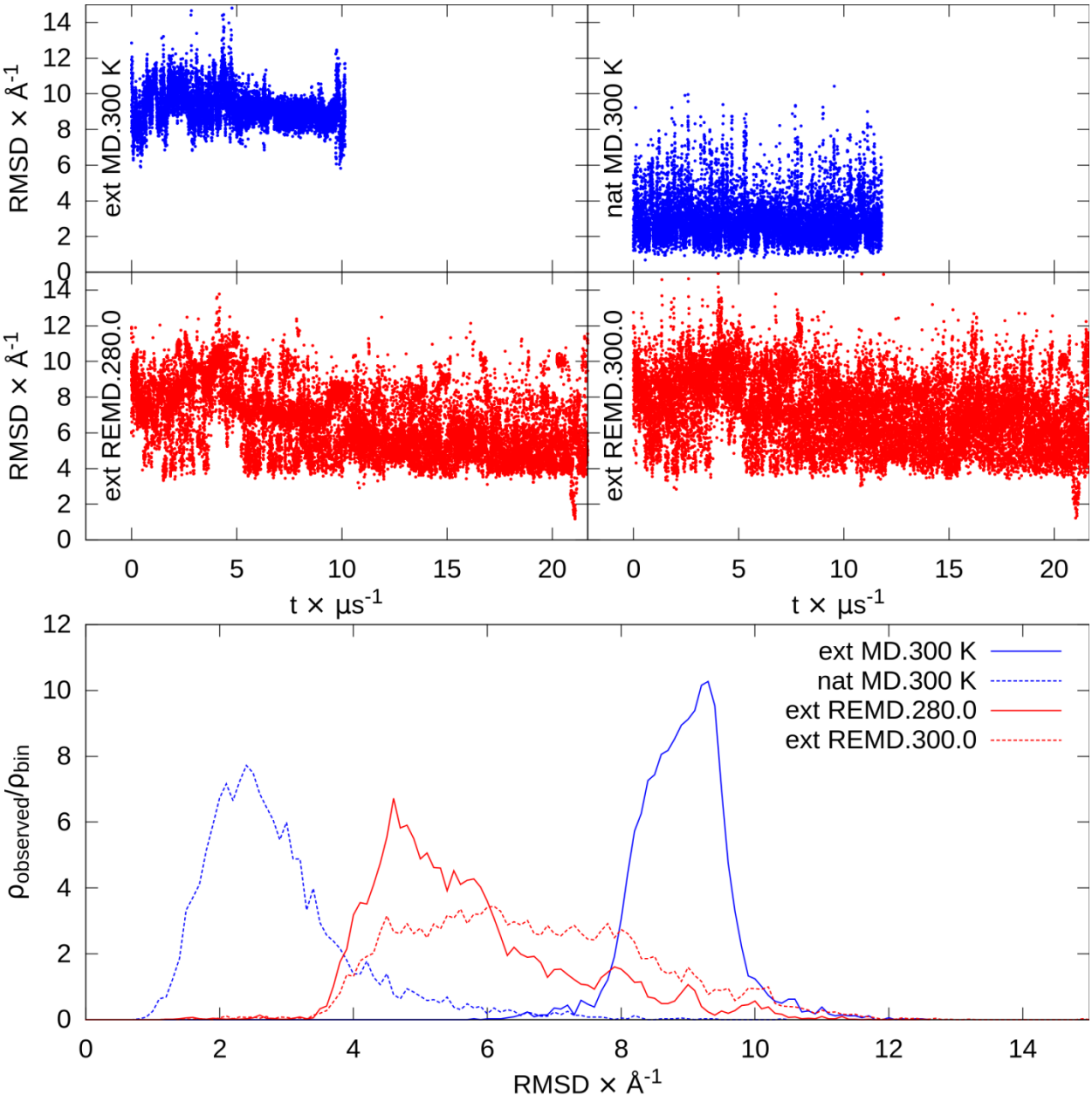


**Figure 3.S36.** NTL9 (52 AA) RMSDs, excluding the 7-16 loop. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

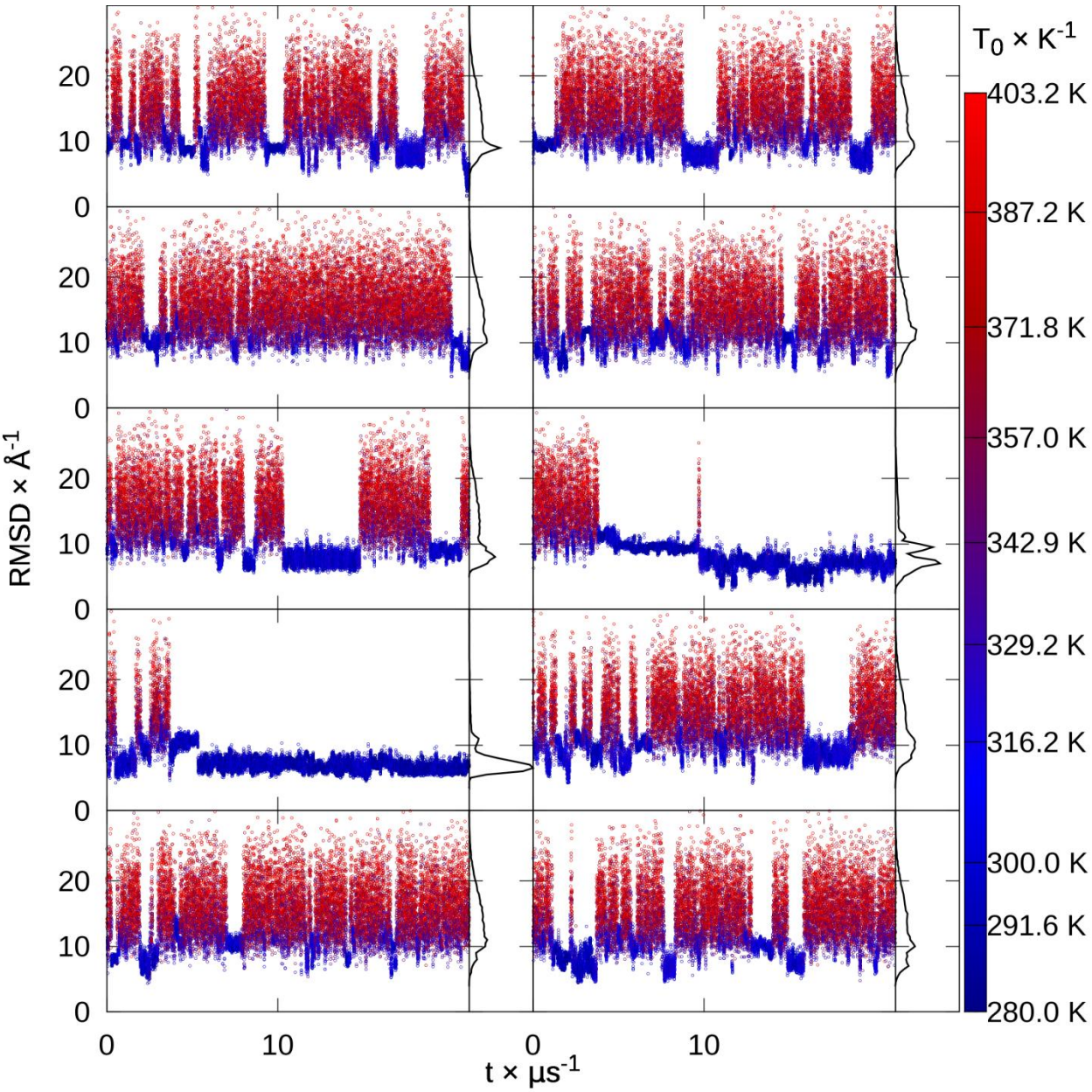


**Figure 3.S37.** NTL9 (52 AA) replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 25.0 | 9.5 | 8.1 | 6.6 | 6.0 |
| Centroid Cα | 6.0 | 9.6 | 6.7 | 6.1 | 7.2 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S15.** NTL9 (52 AA) top 5 extended REMD cluster populations and centroid Cα RMSDs.

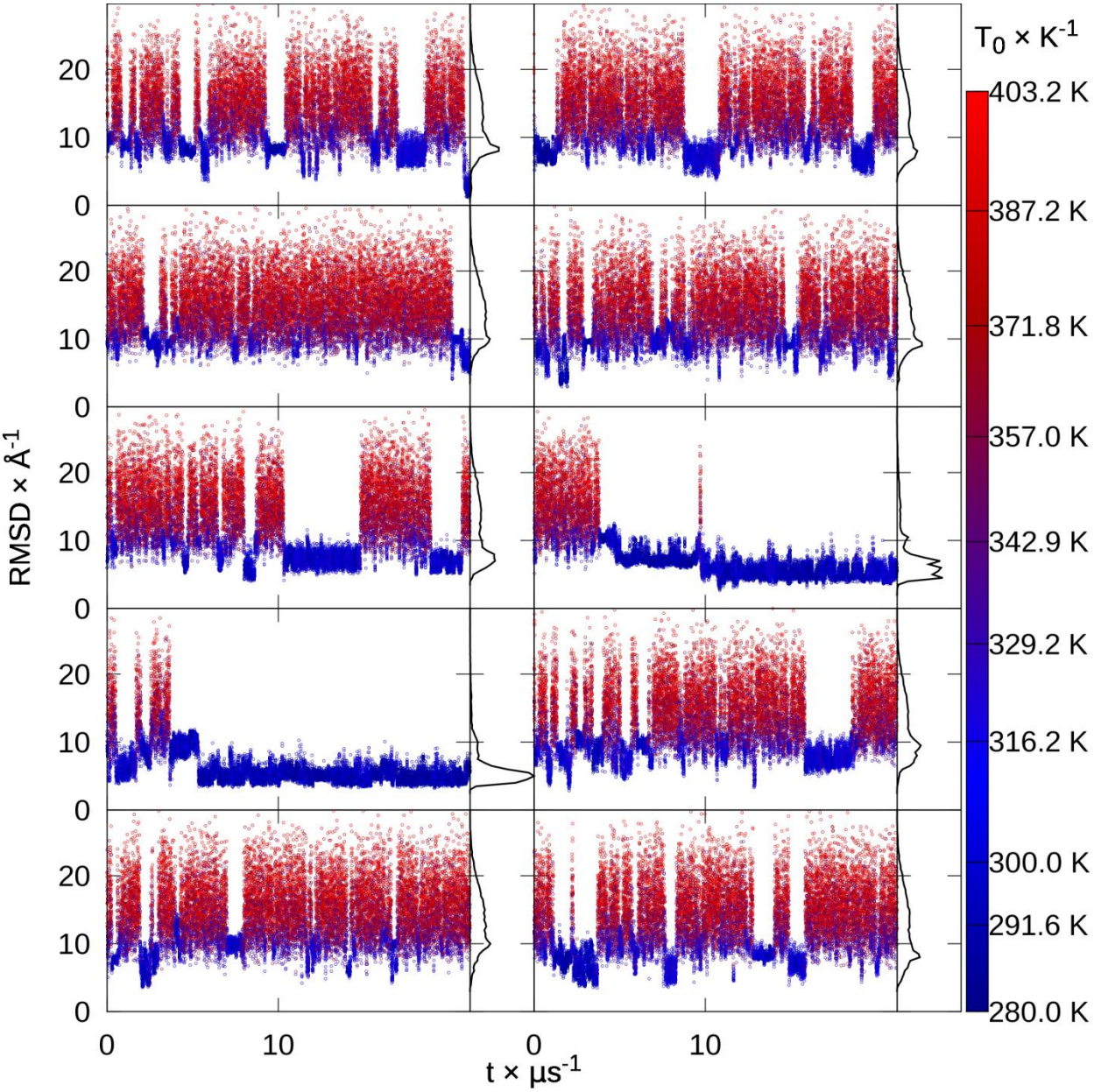


**Figure 3.S38.** NTL9 (52 AA) replica RMSDs, excluding loop. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

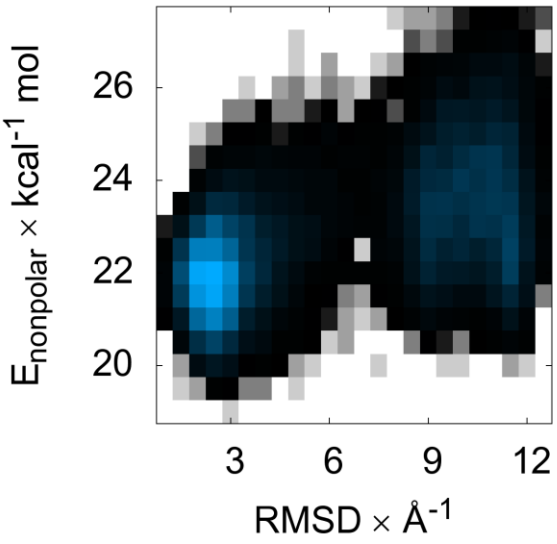


**Figure 3.S39.** NTL9 (52 AA) surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is more favorable at low (1–3 Å) than high (9–12 Å) RMSDs.

**NuG2 variant**

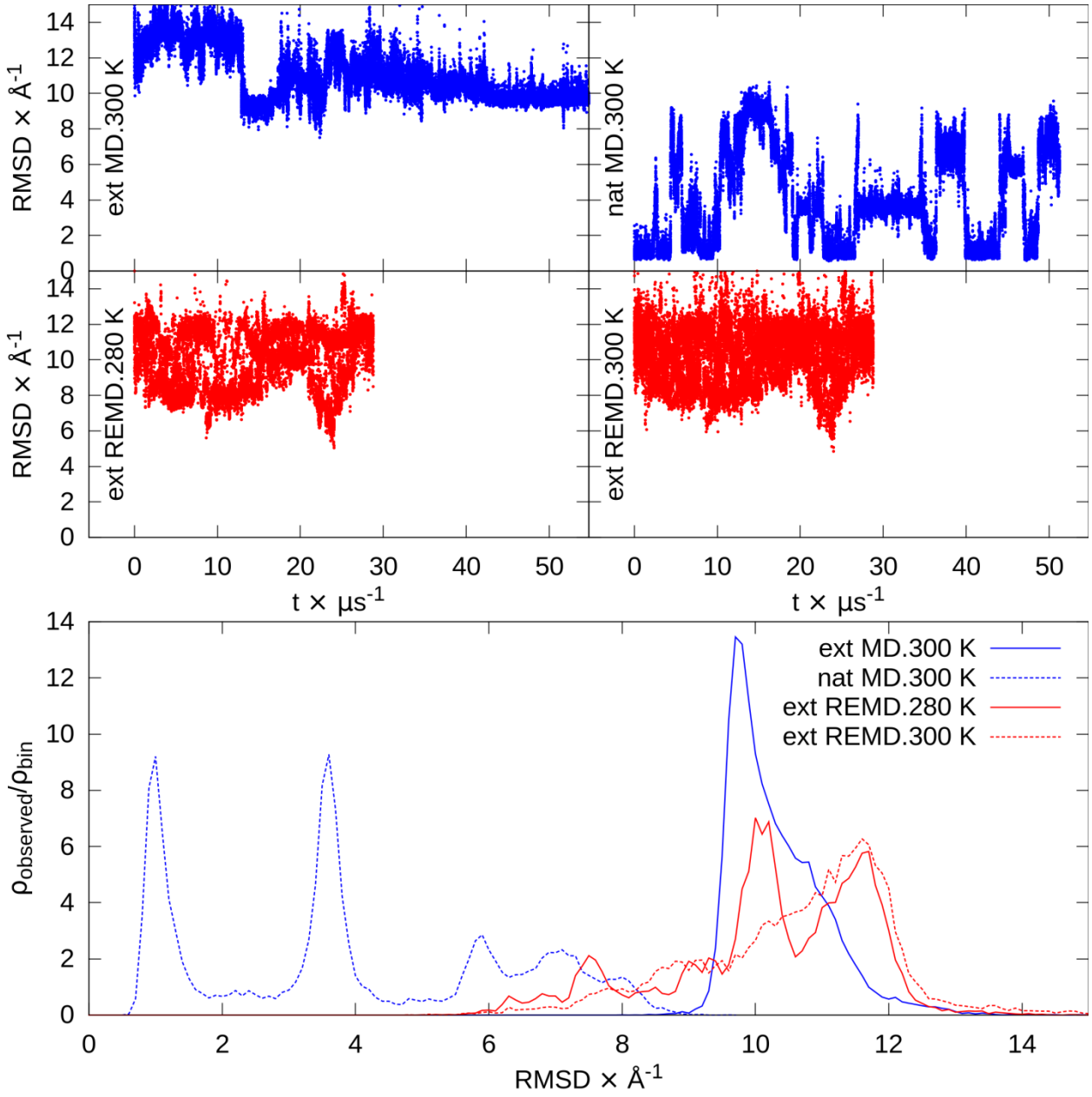


**Figure 3.S40.** NuG2 variant RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

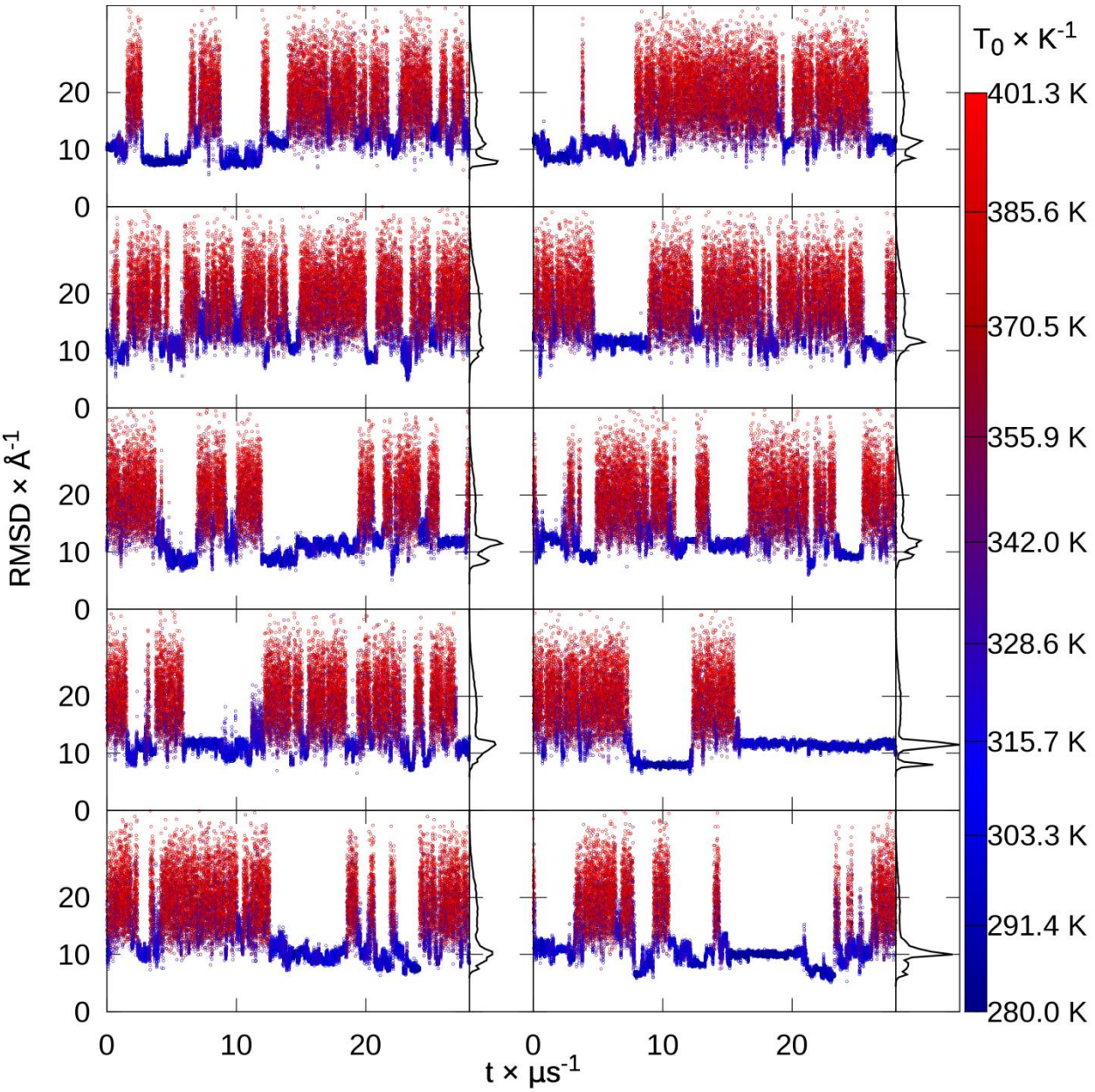


**Figure 3.S41.** NuG2 variant replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 23.3 | 18.1 | 13.8 | 6.9 | 6.6 |
| Centroid Cα | 11.4 | 7.9 | 9.8 | 7.5 | 8.1 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S16.** NuG2 variant top 5 extended REMD cluster populations and centroid Cα RMSDs.

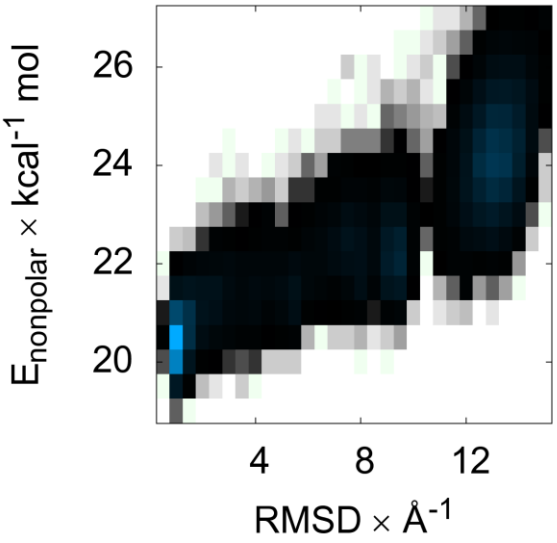


**Figure 3.S42.** NuG2 variant surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal $mol^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is more favorable at low (0–2 Å) RMSD.

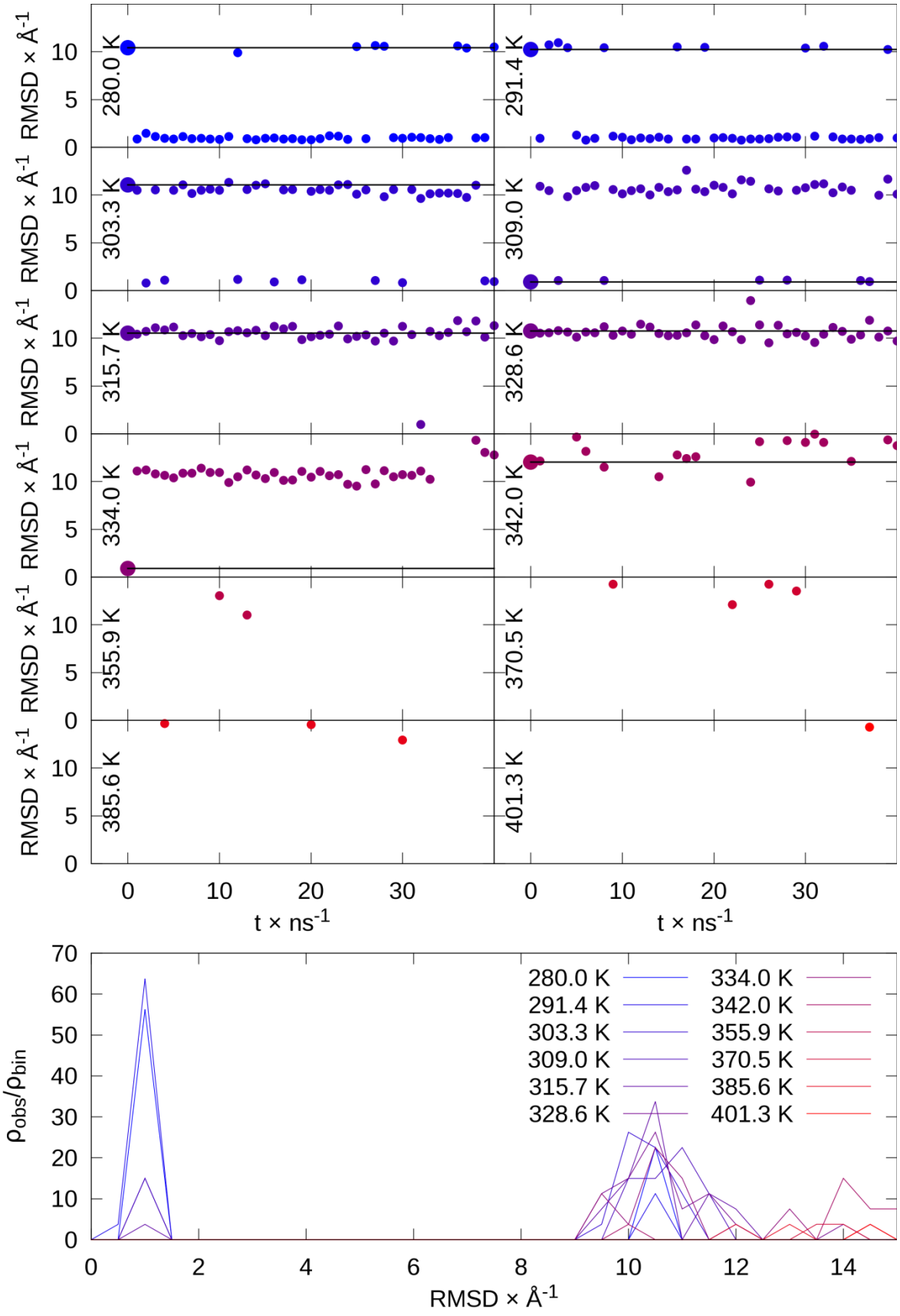
280.0 K
291.4 K
303.3 K
309.0 K
315.7 K
328.6 K
334.0 K
342.0 K
355.9 K
370.5 K
385.6 K
401.3 K
RMSD × Å-1
t × ns-1
ρobs/ρbin
RMSD × Å-1

**Figure 3.S43.** Seeded REMD sorting of NuG2 conformations: 2 × native-like (0.9 Å) added to 10 conformations from extended REMD from 10.4 to 30.3 Å RMSD. Lines indicate initial RMSD value that each temperature. At top, RMSD vs time for each temperature shows sorting of low RMSD conformations to low temperatures. At bottom, histogram shows preference of native-like conformations at low temperatures.

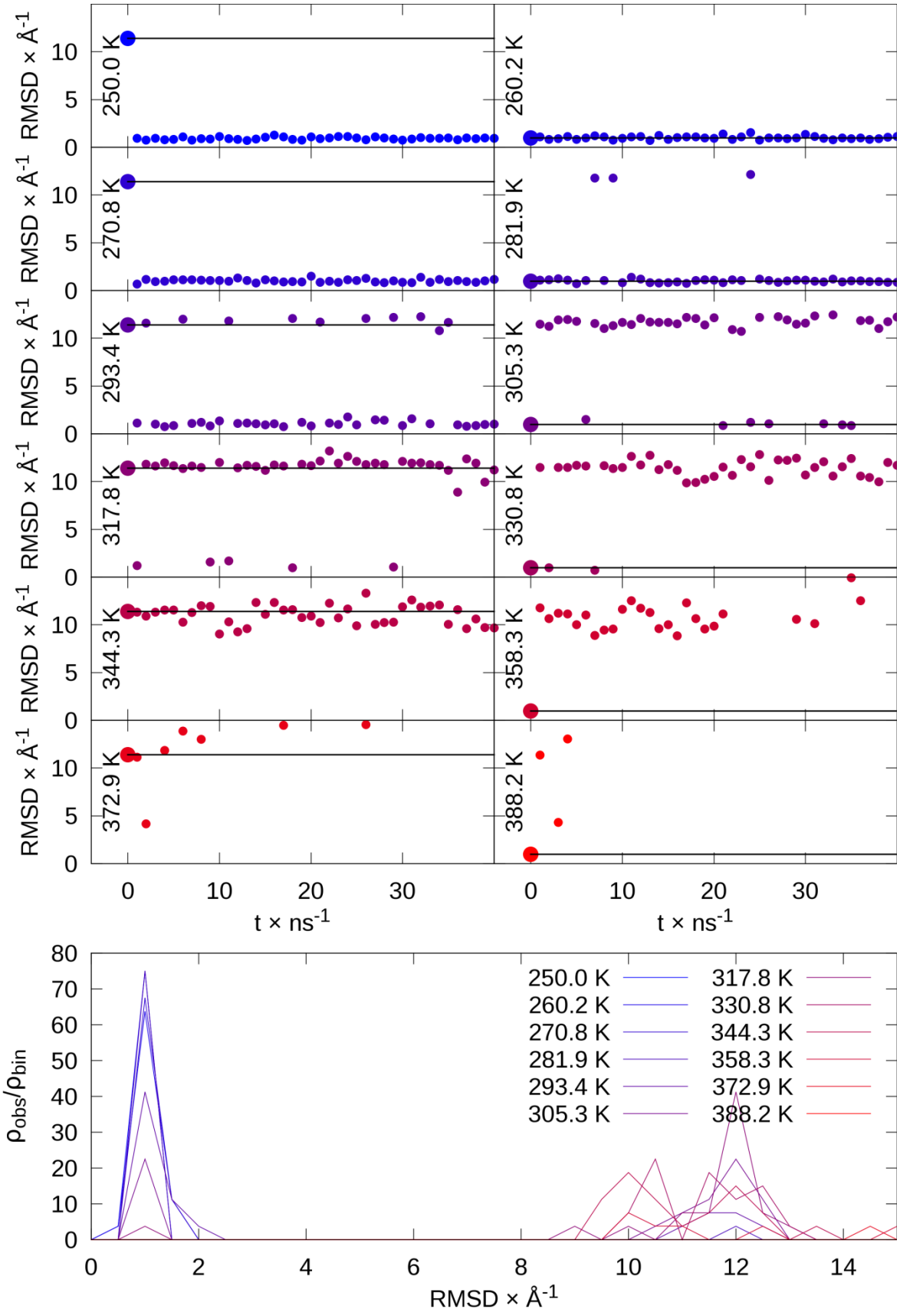
250.0 K
260.2 K
270.8 K
281.9 K
293.4 K
305.3 K
317.8 K
330.8 K
344.3 K
358.3 K
372.9 K
388.2 K
RMSD × Å-1
t × ns-1
ρobs/ρbin
RMSD × Å-1

**Figure 3.S44.** REMD sorting of NuG2 conformations: unfolded (11.4 Å) and native-like (1.0 Å), repeated for 12 replicas. Lines indicate initial RMSD value that each temperature. At top, RMSD vs time for each temperature shows sorting of low RMSD conformations to low temperatures, except for the native conformation at 388.2 K that unfolded. At bottom, histogram shows preference of native-like conformations at low temperatures.

**CspA**

For the β-barrel CspA, the most populated conformation forms a barrel with correct strands 1-3, but the long flexible loop from positions 35-47 adopts a β-hairpin and displaces strand 5, which moves to where strand 4 should be, and the displaced strand 4 adopts a helical conformation (Figure 3.S2). The population is comparable to that of another cluster with near-native fold at 4.8 Å (Table 3.S17).

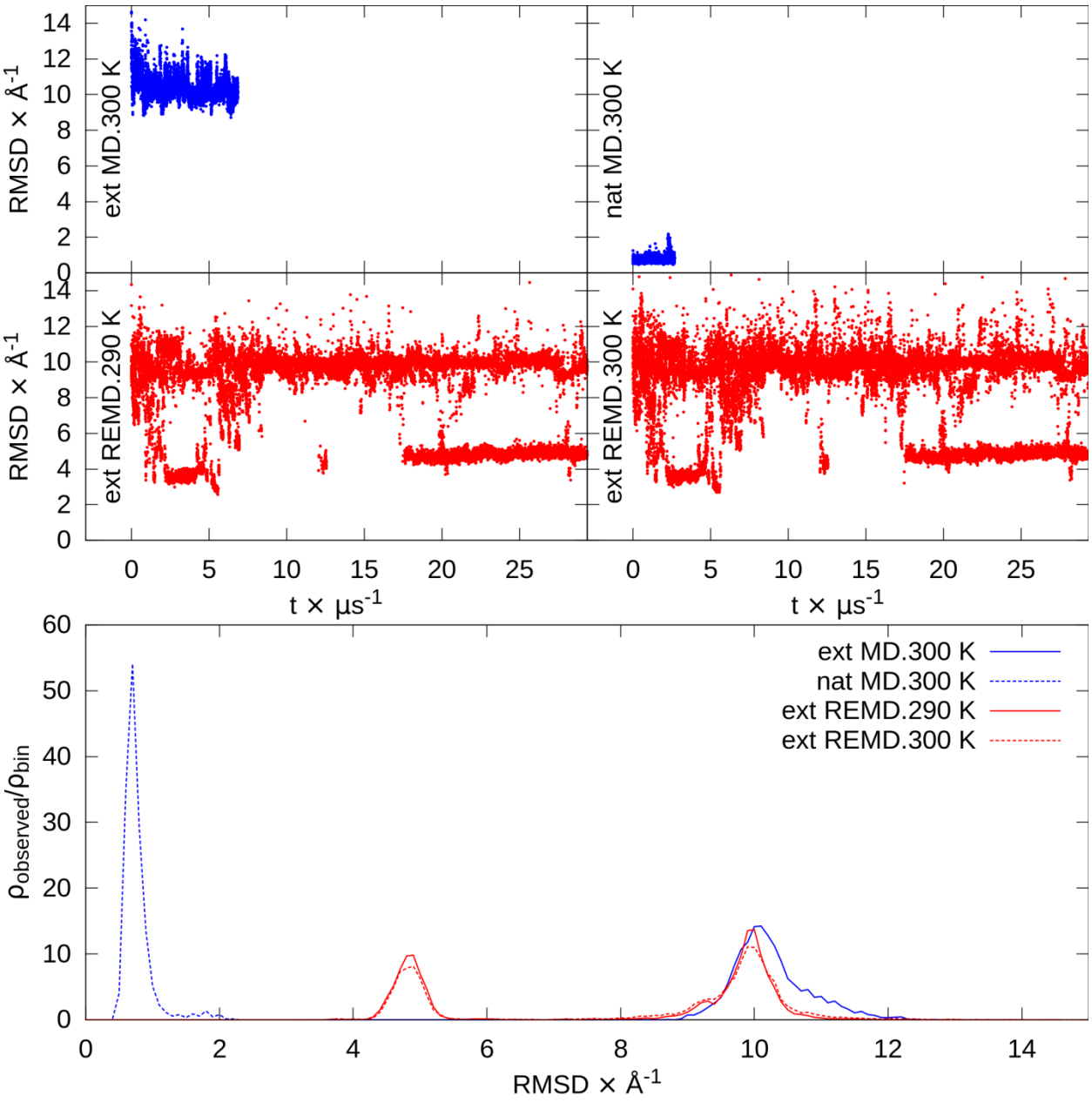


**Figure 3.S45.** CspA RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

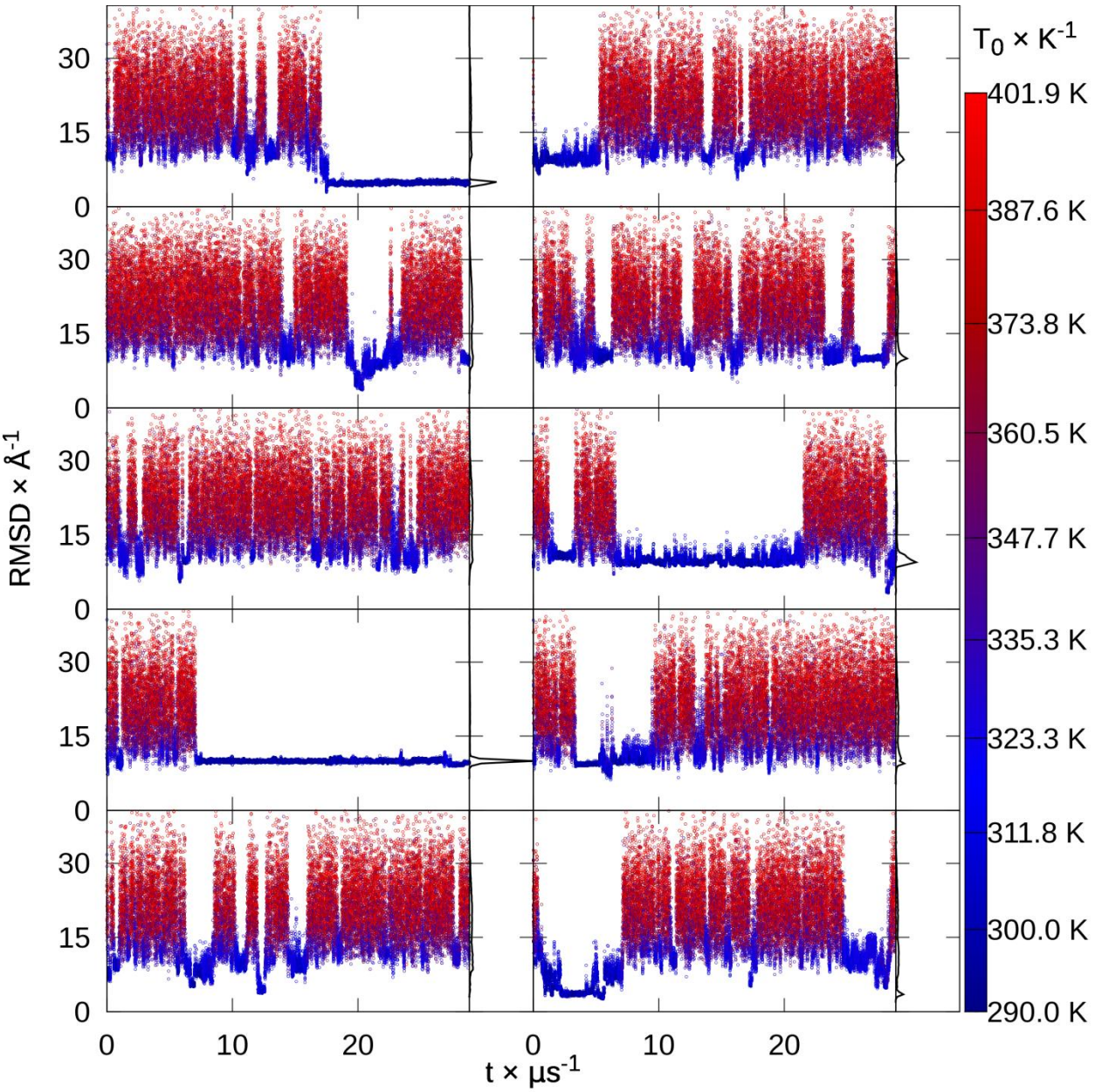


**Figure 3.S46.** CspA replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 33.0 | 17.6 | 11.5 | 6.0 | 5.2 |
| Centroid Cα | 9.9 | 4.8 | 9.5 | 3.3 | 10.0 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S17**. CspA top 5 extended REMD cluster populations and centroid Cα RMSDs.

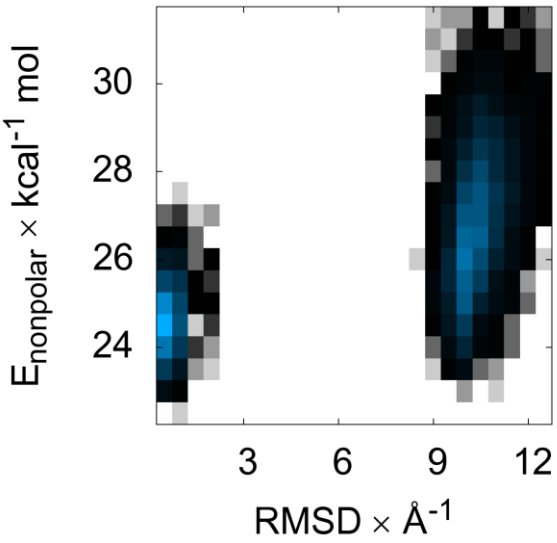


**Figure 3.S47.** CspA surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is more favorable at low (0–2 Å) than high (9–12 Å) RMSDs.

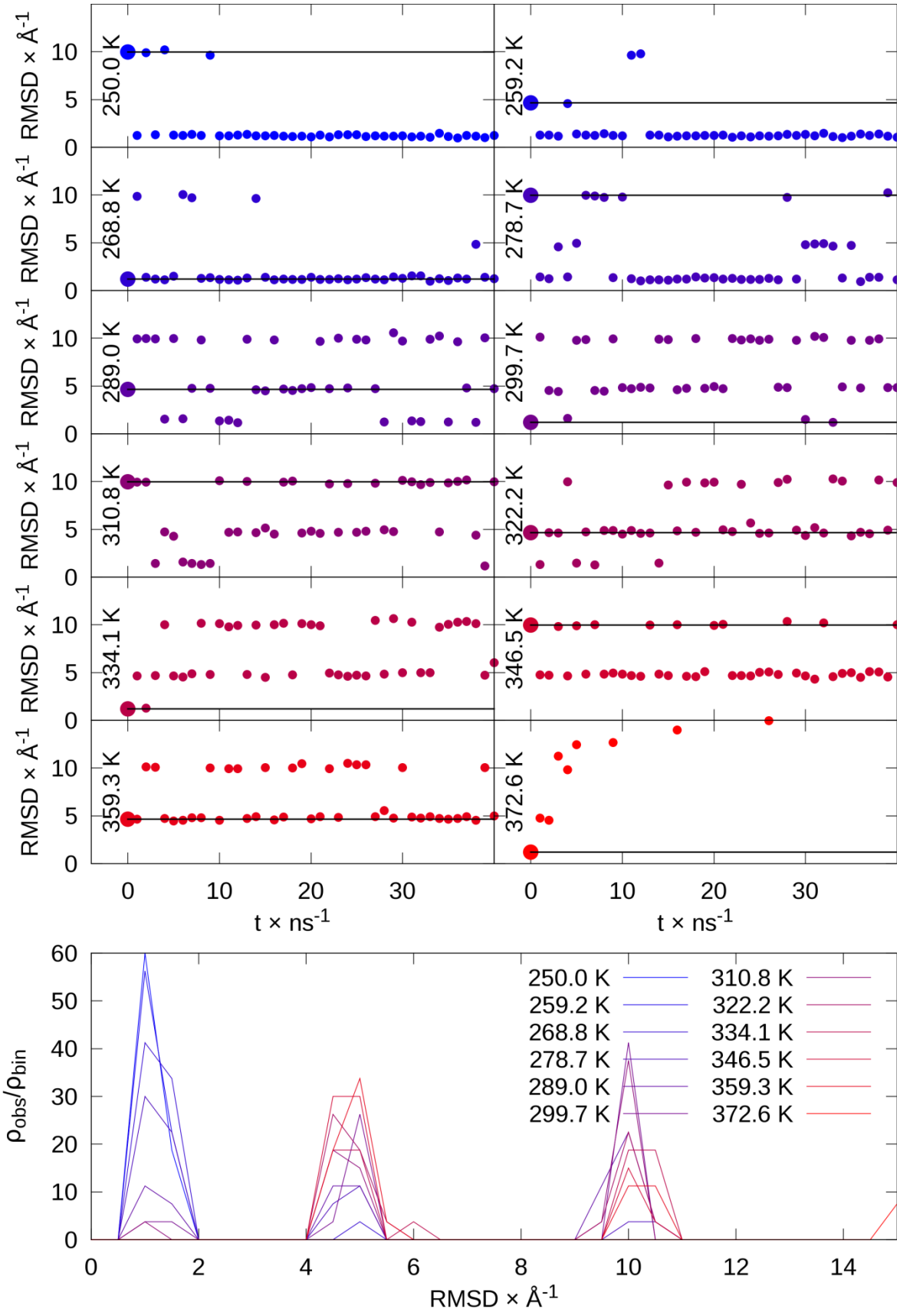
250.0 K
259.2 K
268.8 K
278.7 K
289.0 K
299.7 K
310.8 K
322.2 K
334.1 K
346.5 K
359.3 K
372.6 K
RMSD × Å-1
t × ns-1
ρobs/ρbin
RMSD × Å-1

**Figure 3.S48.** REMD sorting of CspA conformations: unfolded (10.0 Å), partly unfolded (4.7 Å), and native-like (1.2 Å), repeated for 12 replicas. At top, RMSD vs time for each temperature shows sorting of low RMSD conformations to low temperatures. Lines indicate initial RMSD value that each temperature. At bottom, histogram shows preference of low RMSD conformations at low temperatures.

**Hyp protein 1WHZ**

The 70 amino acid hypothetical protein 1WHZ folds to the correct structure with a 3-stranded β-sheet and 3 helices, but the most populated structure replaces the first β-strand with a helix and the last two helices with two β-strands. Otherwise, the RMSDs of the first helix and N-terminus (residues 1 to 18) and the second and third β-strands (residues 28 to 44) are both 1.8 Å. Examining the RMSD evolution of individual replicas in the ~10 μsec REMD run indicates that multiple misfolded structures are sampled, typically stable for several μsec, adopting a variety of mixed α /β topologies (Figure 3.S50–51).

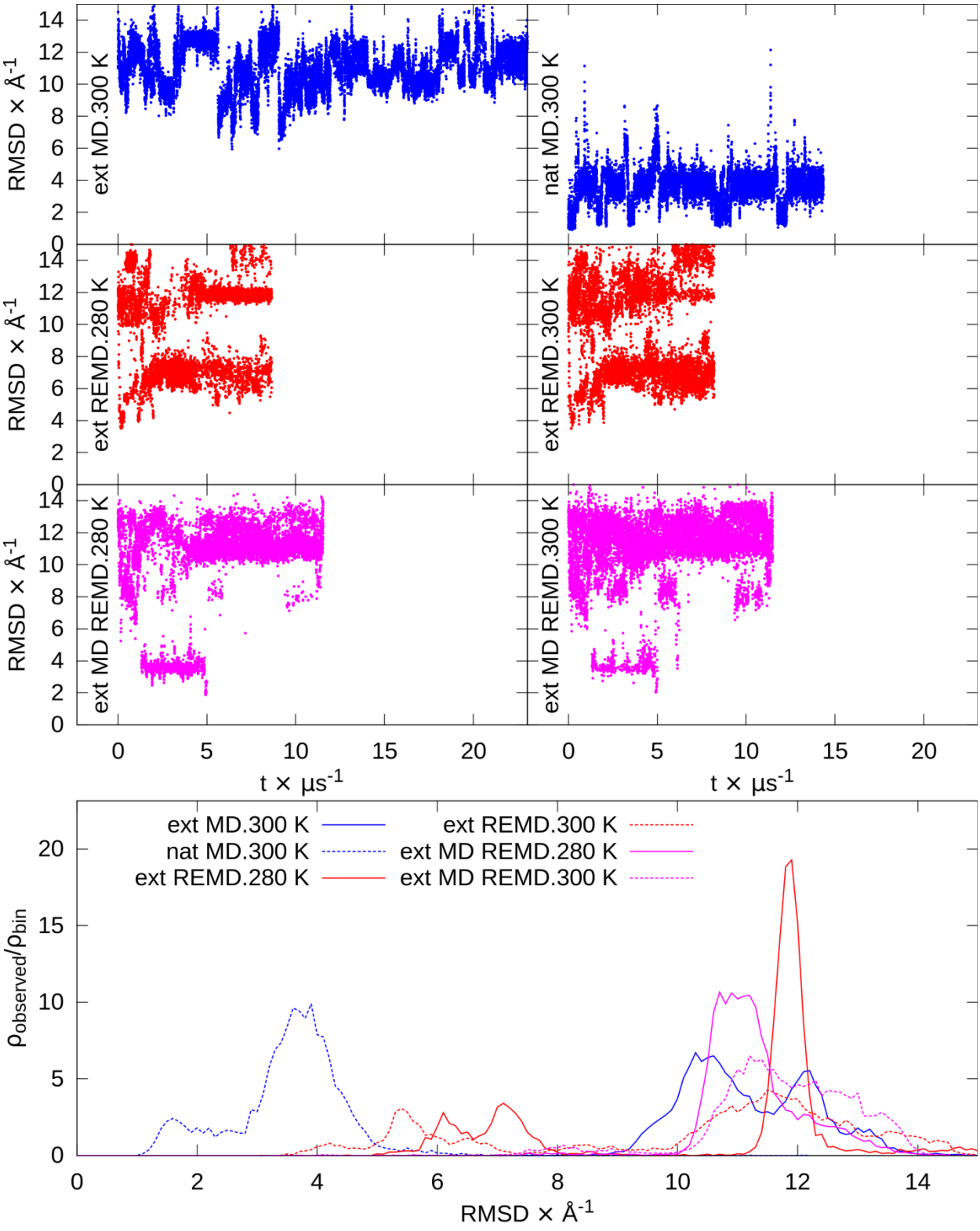


**Figure 3.S49.** Hypothetical protein 1WHZ RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD

histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

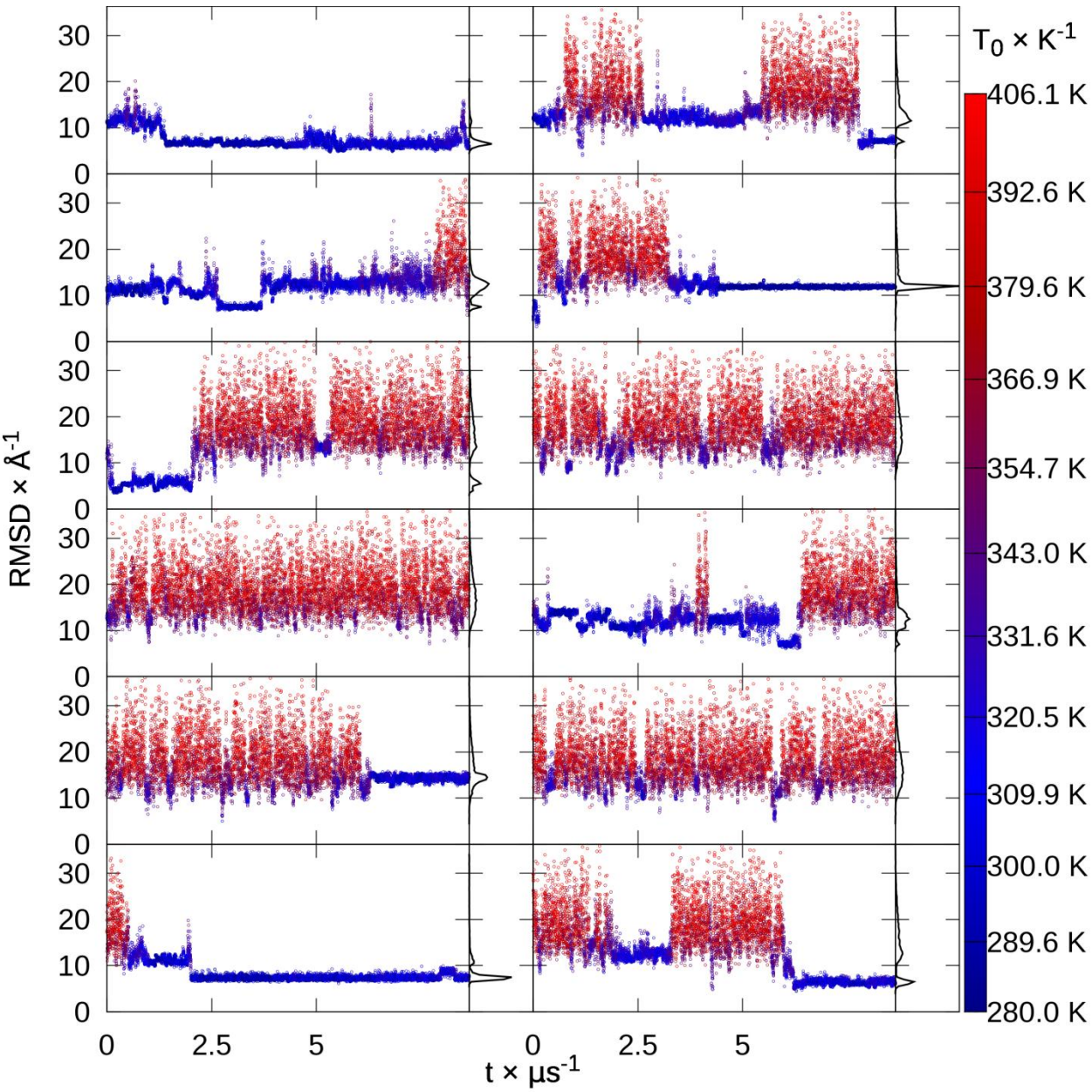


**Figure 3.S50.** Hypothetical protein 1WHZ replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 36.1 | 16.7 | 14.8 | 5.5 | 4.6 |
| Centroid Cα RMSD (Å) | 11.8 | 7.2 | 6.6 | 11.8 | 13.9 |

**Table 3.S18.** Hypothetical protein 1WHZ top 5 extended REMD cluster populations and centroid Cα RMSDs.

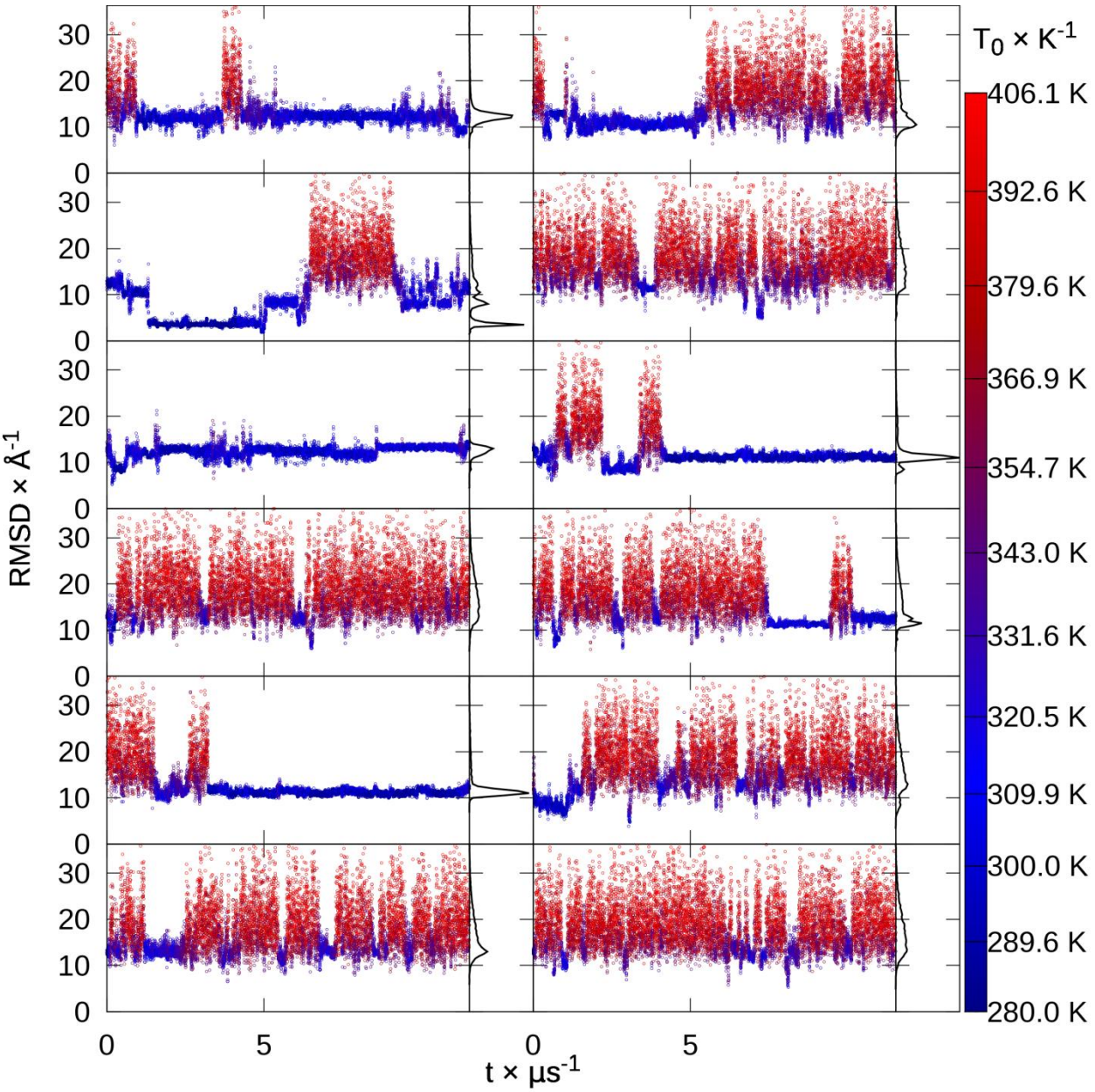


**Figure 3.S51.** Hypothetical protein 1WHZ replica RMSDs. RMSD to native of each replica from replica exchange initiated with extended MD structures versus time, colored by snapshot temperature from blue to red, with histograms. This differs from the former hypothetical protein replica RMSDs by the starting structures of the REMD.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 43.9 | 18.1 | 9.9 | 5.1 | 3.3 |

| | | | | | |
|---|---|---|---|---|---|
| Centroid Cα RMSD (Å) | 11.0 | 3.4 | 12.5 | 12.7 | 11.8 |

Table 3.S1. Hypothetical protein 1WHZ top 5 extended MD REMD cluster populations and centroid Cα RMSDs.

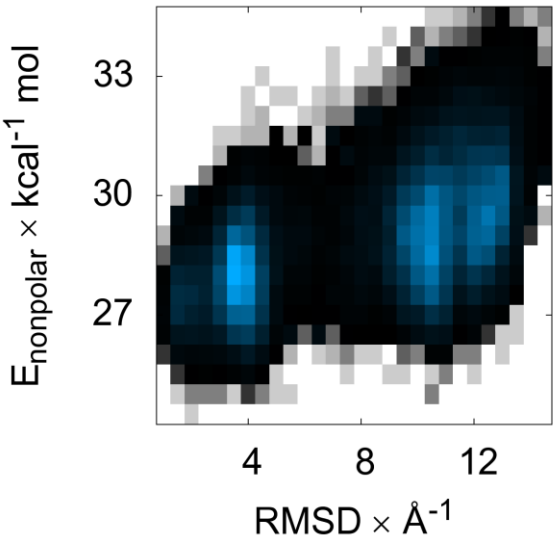


**Figure 3.S52**. Hyp protein 1WHZ surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal $mol^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is slightly more favorable at low (1–3 Å) than high (9–12 Å) RMSDs.

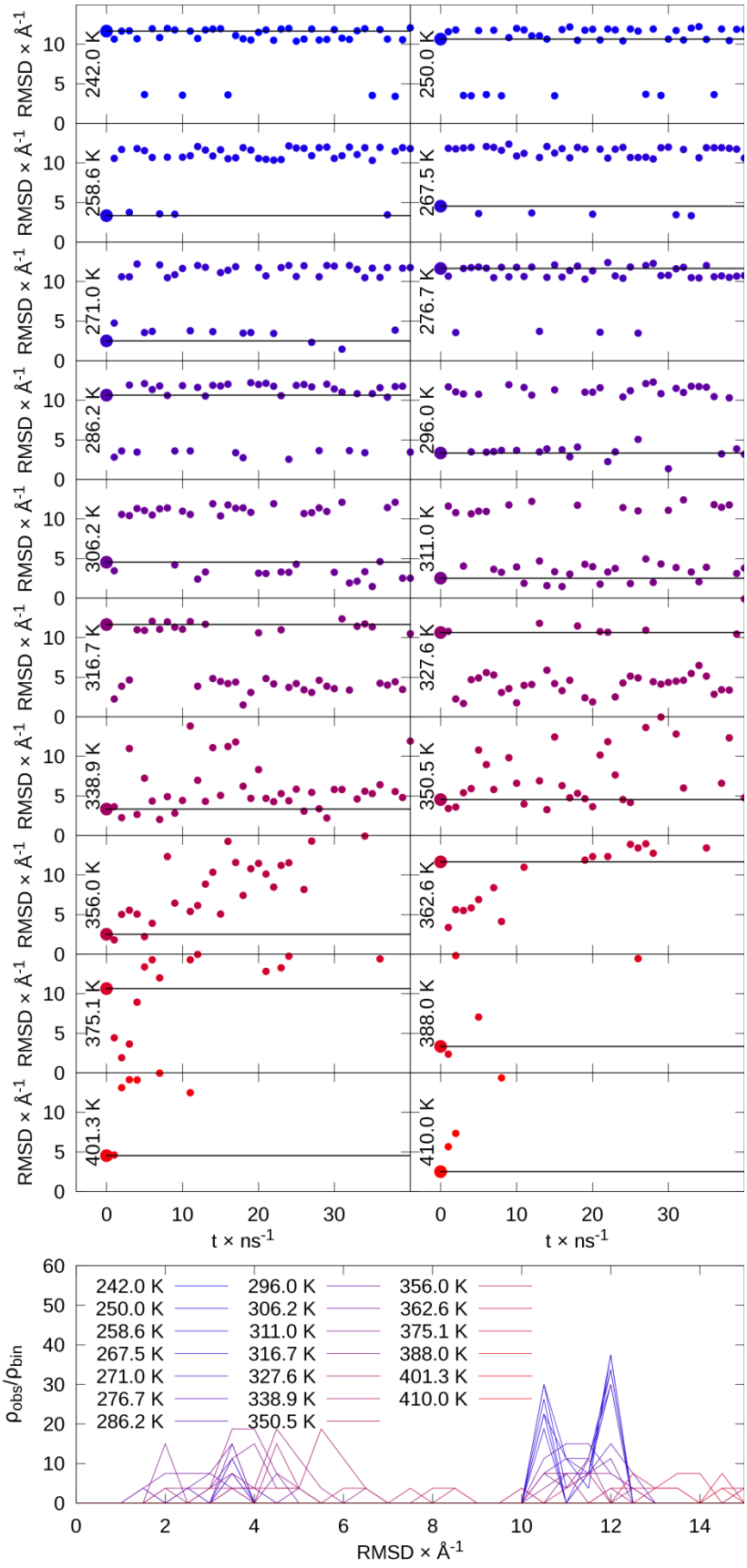
242.0 K
250.0 K
258.6 K
267.5 K
271.0 K
276.7 K
286.2 K
296.0 K
306.2 K
311.0 K
316.7 K
327.6 K
338.9 K
350.5 K
356.0 K
362.6 K
375.1 K
388.0 K
401.3 K
410.0 K
RMSD × Å-1
t × ns-1
ρobs/ρbin
RMSD × Å-1

**Figure 3.S53.** REMD sorting of hypothetical protein 1WHZ conformations: 2 unfolded (11.7 Å, 10.7 Å), 2 partly folded (3.3 Å, 4.5 Å), and 1 native-like (2.5 Å), repeated for 20 replicas. At top, RMSD vs time for each temperature shows sorting of high RMSD conformations to low temperatures, followed by partly followed conformations. Lines indicate initial RMSD value that each temperature. At bottom, histogram shows preference of high and then intermediate RMSD conformations at low temperatures.

**α3D**

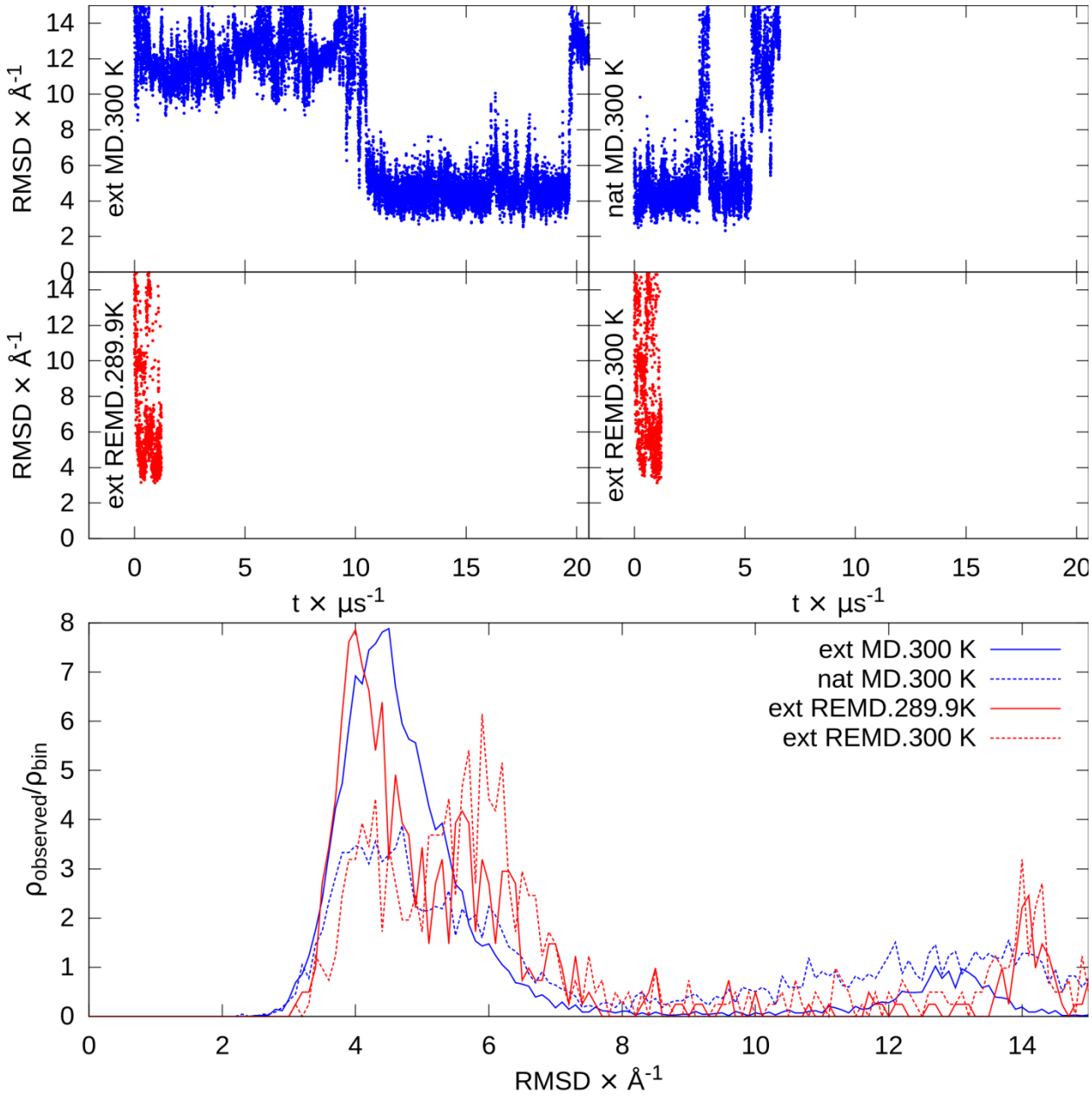


**Figure 3.S54.** α3D RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half

of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

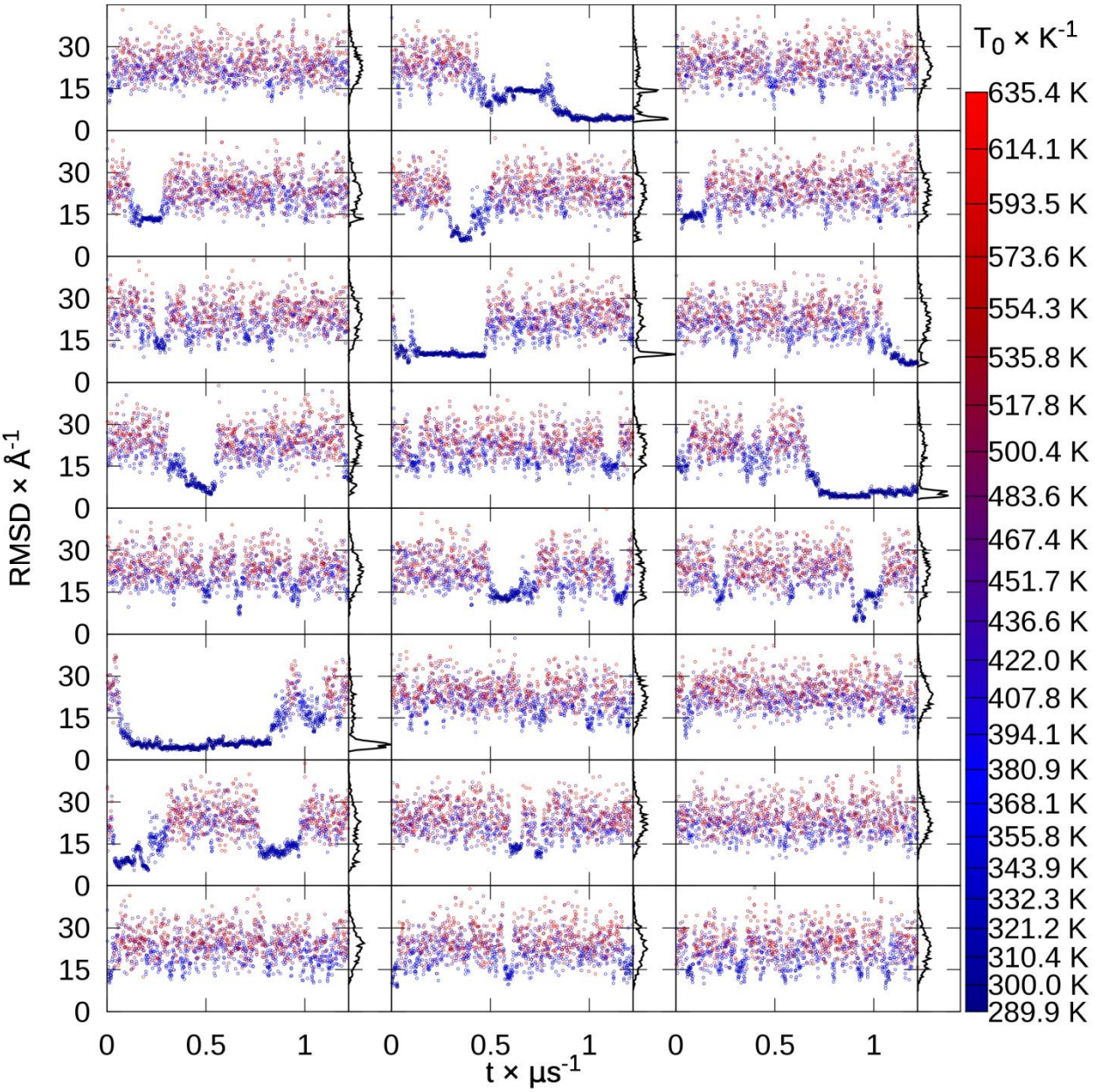


**Figure 3.S55.** α3D replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 32.7 | 18.4 | 9.2 | 7.1 | 6.8 |

| | | | | | |
|---|---|---|---|---|---|
| Centroid Cα RMSD (Å) | 4.0 | 4.1 | 5.5 | 10.1 | 6.0 |

**Table 3.S19.** α3D top 5 extended REMD cluster populations and centroid Cα RMSDs.

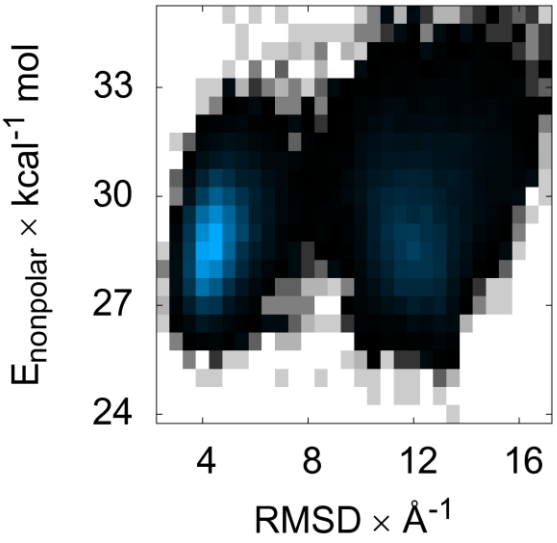


**Figure 3.S56.** α3D surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is similarly to slightly more favorable at high (10–13 Å) than low (2-4 Å) RMSDs.

**λ-repressor**

The 80 amino acid λ-repressor shows transient folding to the native structure in REMD, but the majority of the population adopts a misfolded structure with RMSD of 12 Å (Figure 3.S57). In this case, the 5 α-helices are largely present, but they pack against the first helix in a clockwise fashion, rather than counterclockwise as seen in the native fold. Using the coordinate-seeded REMD approach, we combined structures of the misfolded topology, the lowest RMSD from REMD, and native structures. Similar to 1WHZ, the results indicated our model prefers the structure with the helices formed but incorrectly arranged around the first helix (S60).

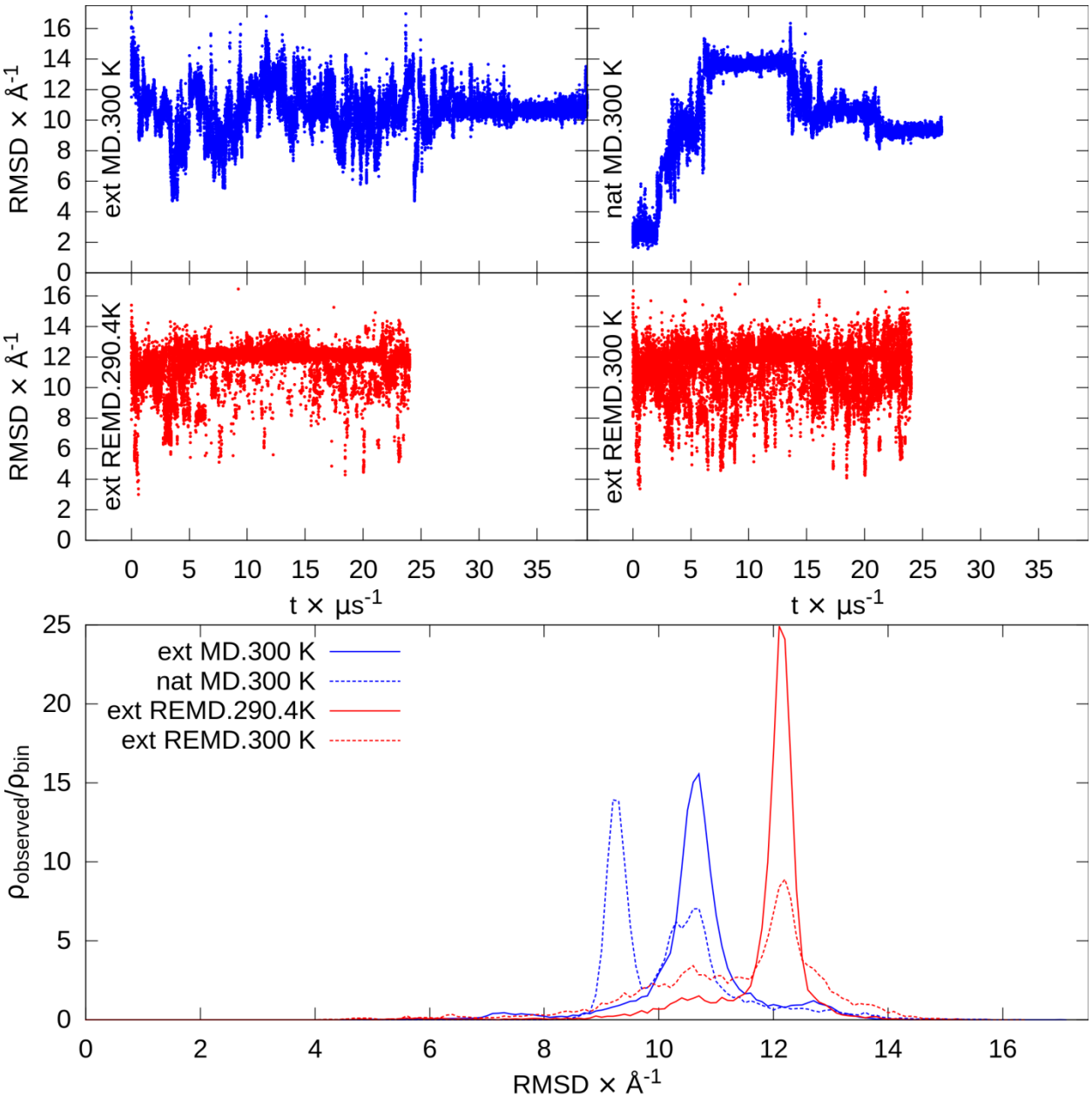


**Figure 3.S57.** λ-repressor RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

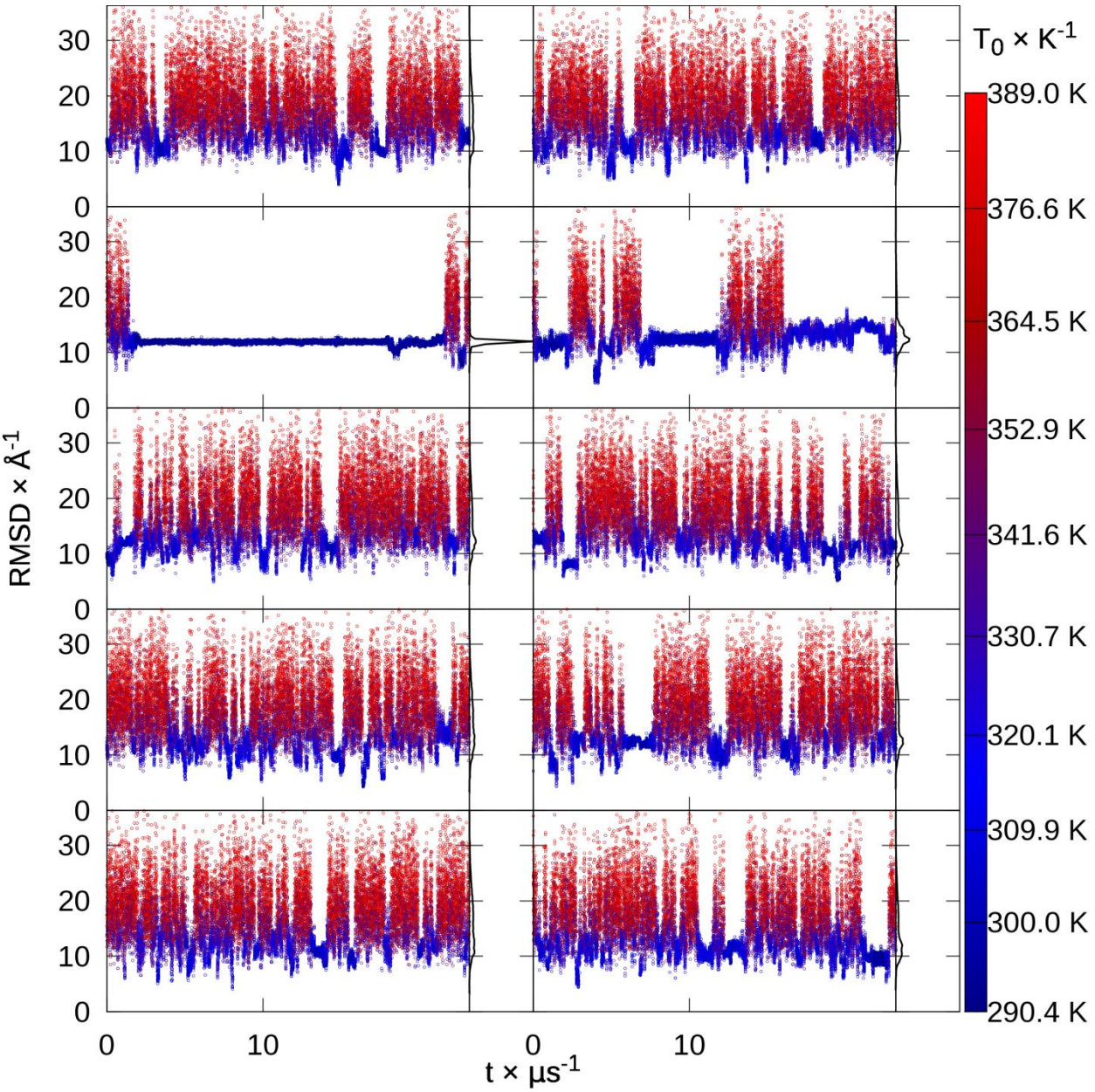


**Figure 3.S58.** λ-repressor replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 53.9 | 4.4 | 3.8 | 3.4 | 3.3 |
| Centroid Cα | 11.9 | 10.9 | 11.2 | 12.1 | 9.8 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S20**. λ-repressor top 5 extended REMD cluster populations and centroid Cα RMSDs.

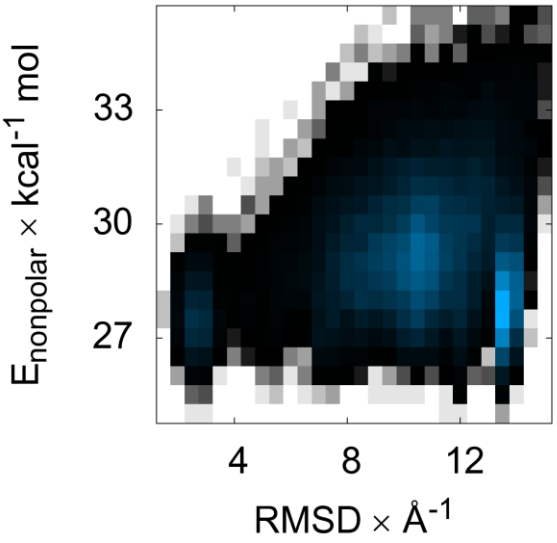


**Figure 3.S59.** λ-repressor surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal mol$^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is flat across low (2-4 Å) to high (12-14 Å) RMSDs

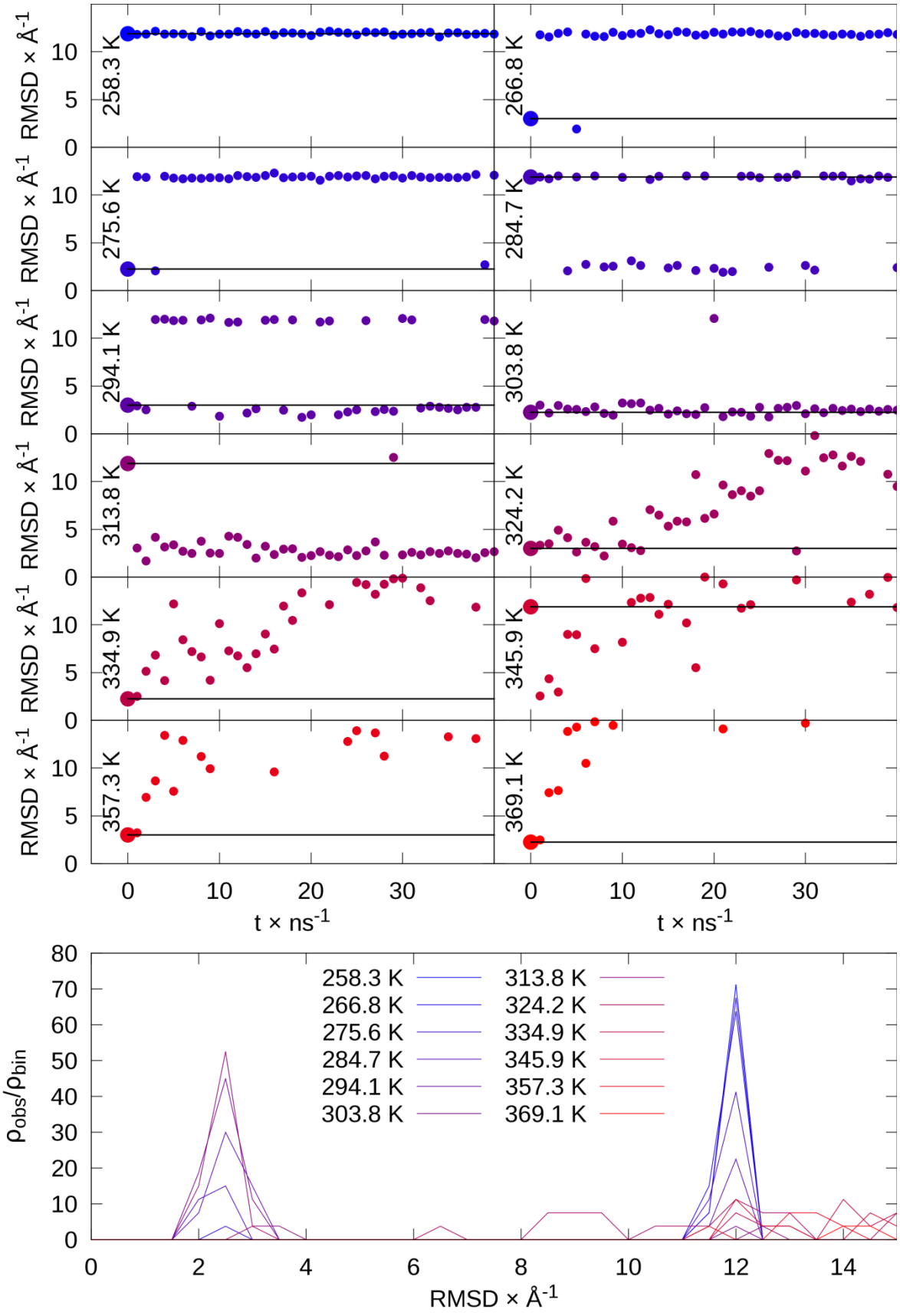
258.3 K
266.8 K
275.6 K
284.7 K
294.1 K
303.8 K
313.8 K
324.2 K
334.9 K
345.9 K
357.3 K
369.1 K
RMSD × Å-1
t × ns-1
ρobs/ρbin
RMSD × Å-1

**Figure 3.S60.** REMD sorting of λ-repressor conformations: unfolded (12.1 Å), partly unfolded (3.0 Å), and native-like (2.3 Å), repeated for 12 replicas. At top, RMSD vs time for each temperature shows sorting of high RMSD conformations to low temperatures. Lines indicate initial RMSD value that each temperature. At bottom, histogram shows preference of high RMSD conformations at low temperatures.

**Top7**

The largest system we studied is the designed protein Top7 (92 amino acids). Conformational sampling is observed to be very slow in this system, with the extended conformation and folded structure both stable for the entire ~5 μsec MD runs (Figure 3.S61). Most replicas spend the majority of the simulation trapped in different local minima, suggesting that the data are poorly converged at 20 μsec of REMD (Figure 3.S64). The native topology for Top7 resembles 2 zinc finger domains with the β-hairpins connected through an additional β-strand, forming a 5-stranded sheet in the protein. The misfolded structure with highest population is only sampled by 1 replica. It shows correct placement of strands 3, 4 and 5, as well as the helix between strands 3 and 4, with an RMSD of 3.3 Å for the region 42-92 (Figures S2 and S66). β-strand 1 is also folded, but the subsequent strand 2 and helix are not yet well formed. Seeded REMD combining the misfolded and correctly folded structures showed a strong preference for the correct fold (Figure 3.S3), moving all misfolded structures to higher temperatures (Figure 3.S68). Interestingly, two of the misfolded replicas refolded to the correct structure during this run.

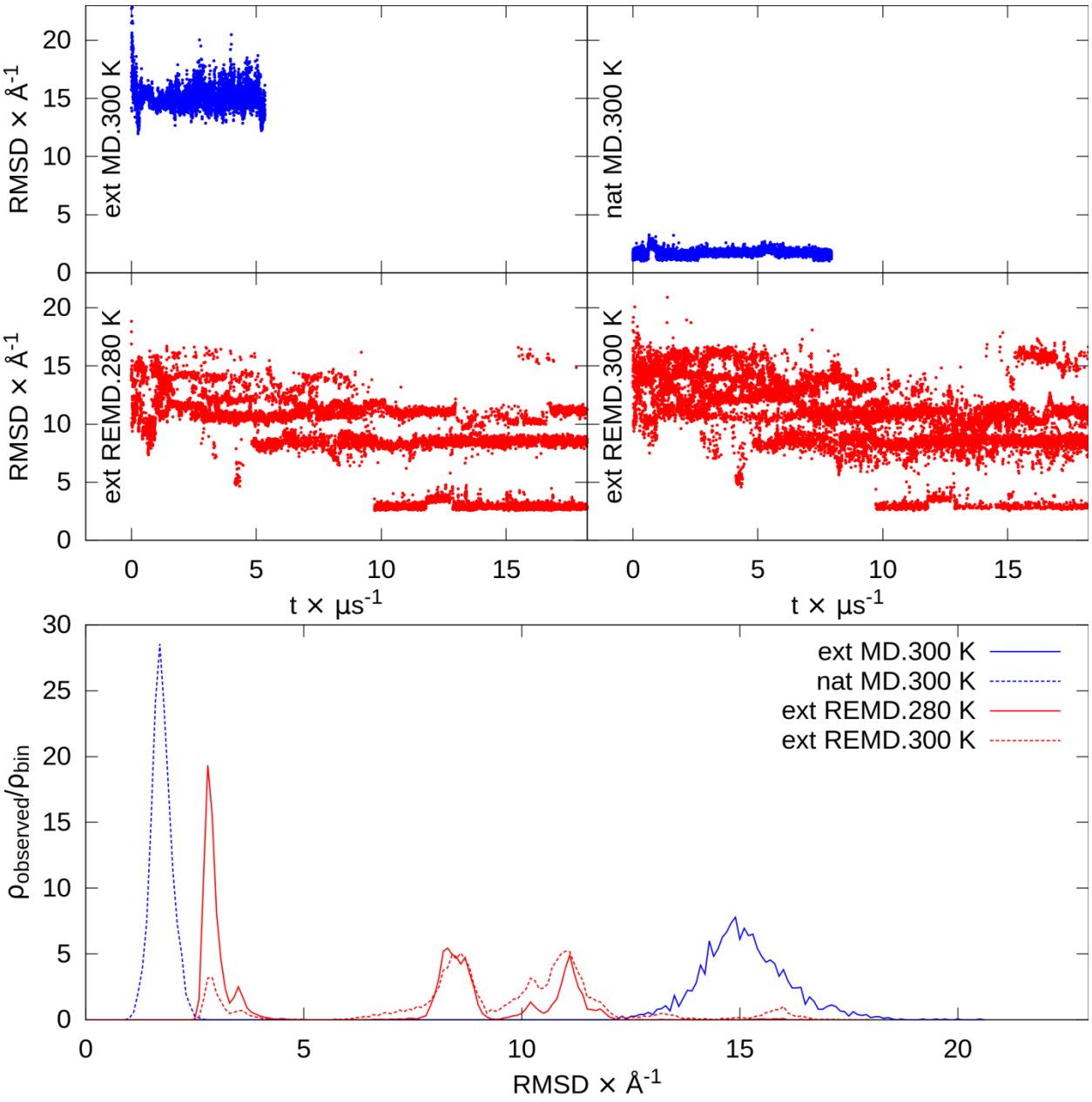


**Figure 3.S61.** Top7 RMSDs. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

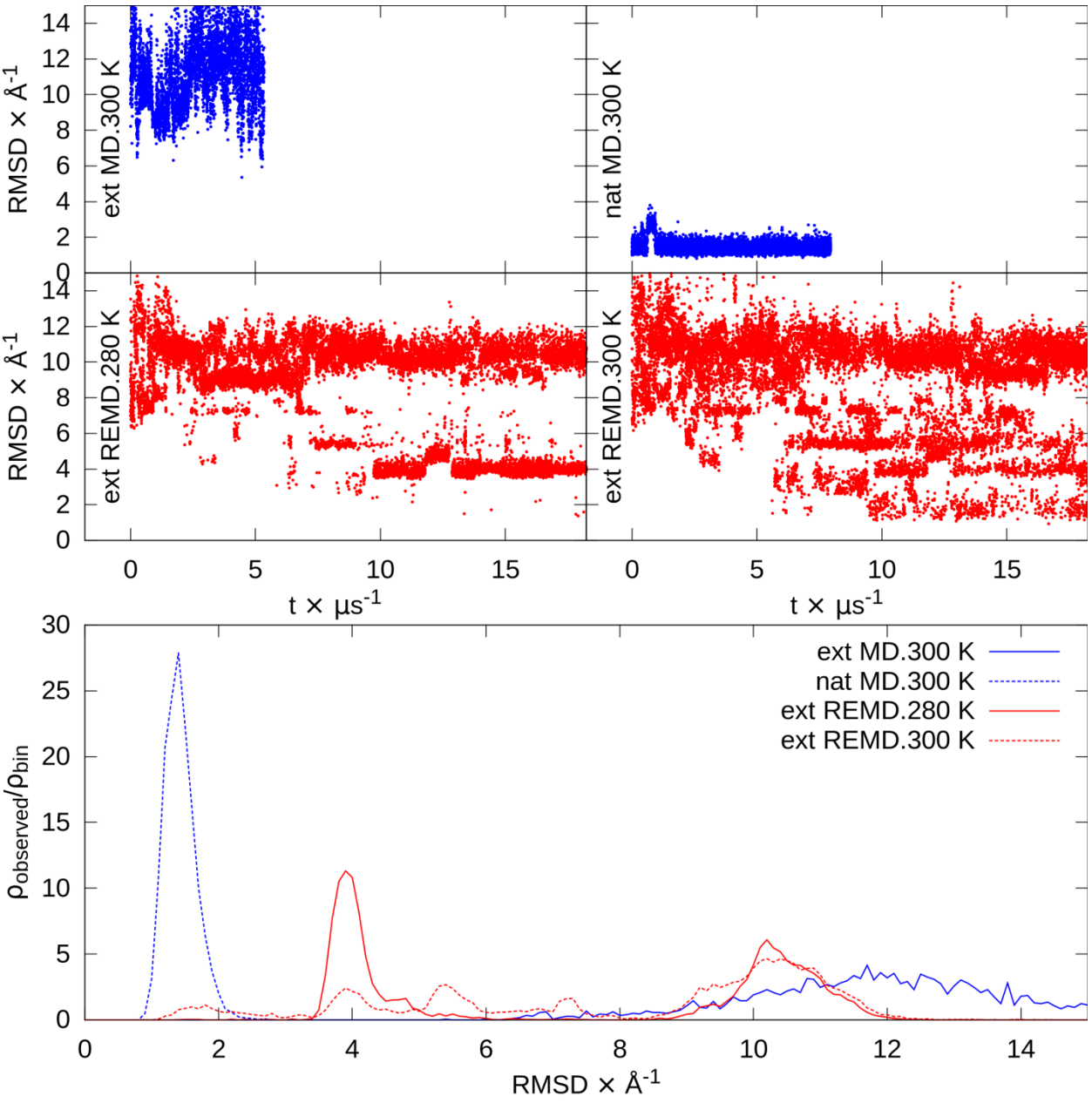


**Figure 3.S62.** Top7 RMSDs, residues 1 to 40. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

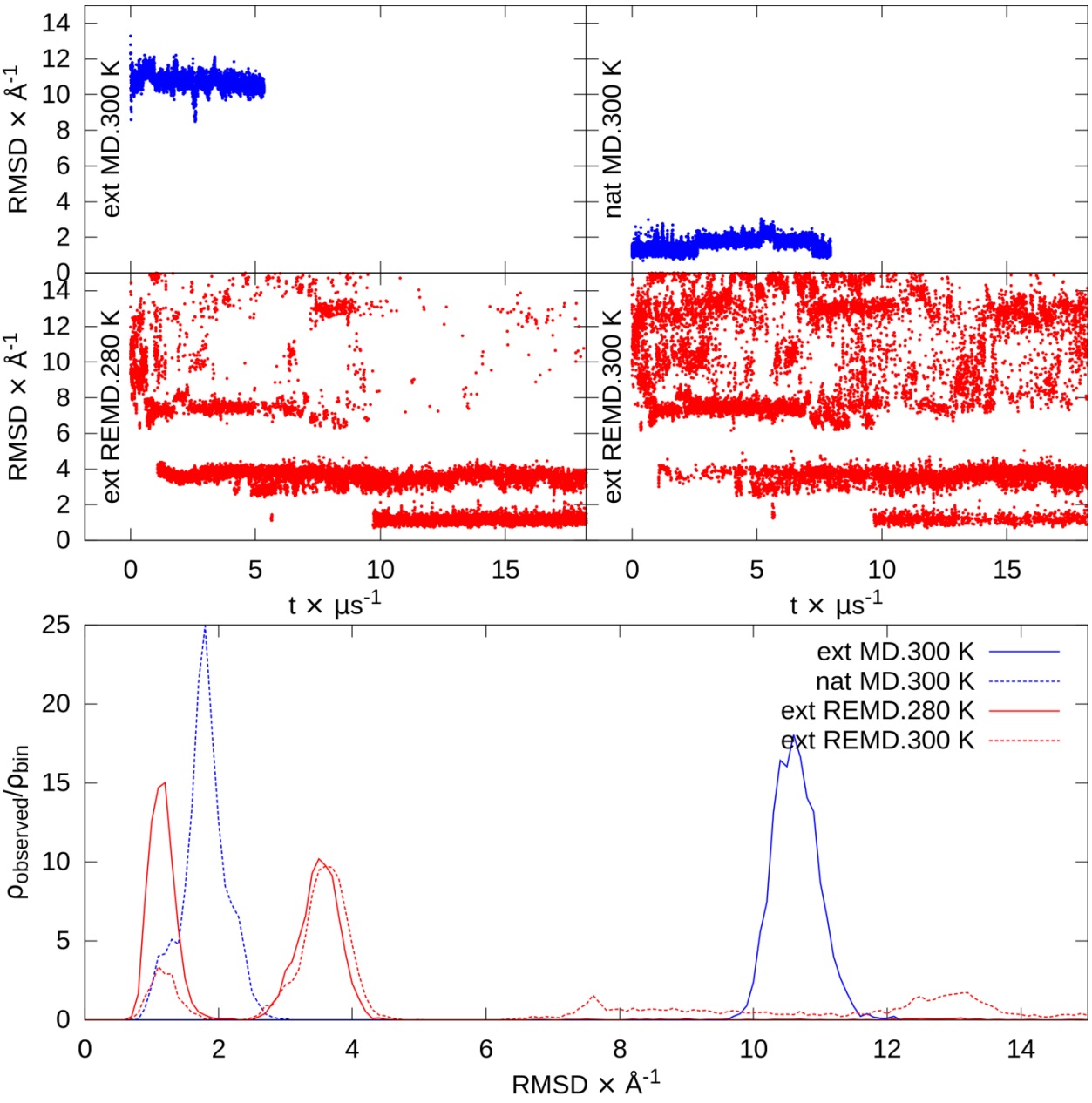


**Figure 3.S63.** Top7 RMSDs, residues 42 to 92. At top are RMSD versus time for extended and native MD and the lowest temperatures from extended REMD. At bottom are RMSD histograms of the second half of each simulation. $\rho_{observed}/\rho_{bin}$ represents the fraction observed ($\rho_{observed}$) relative to the fraction of the binsize ($\rho_{bin}$).

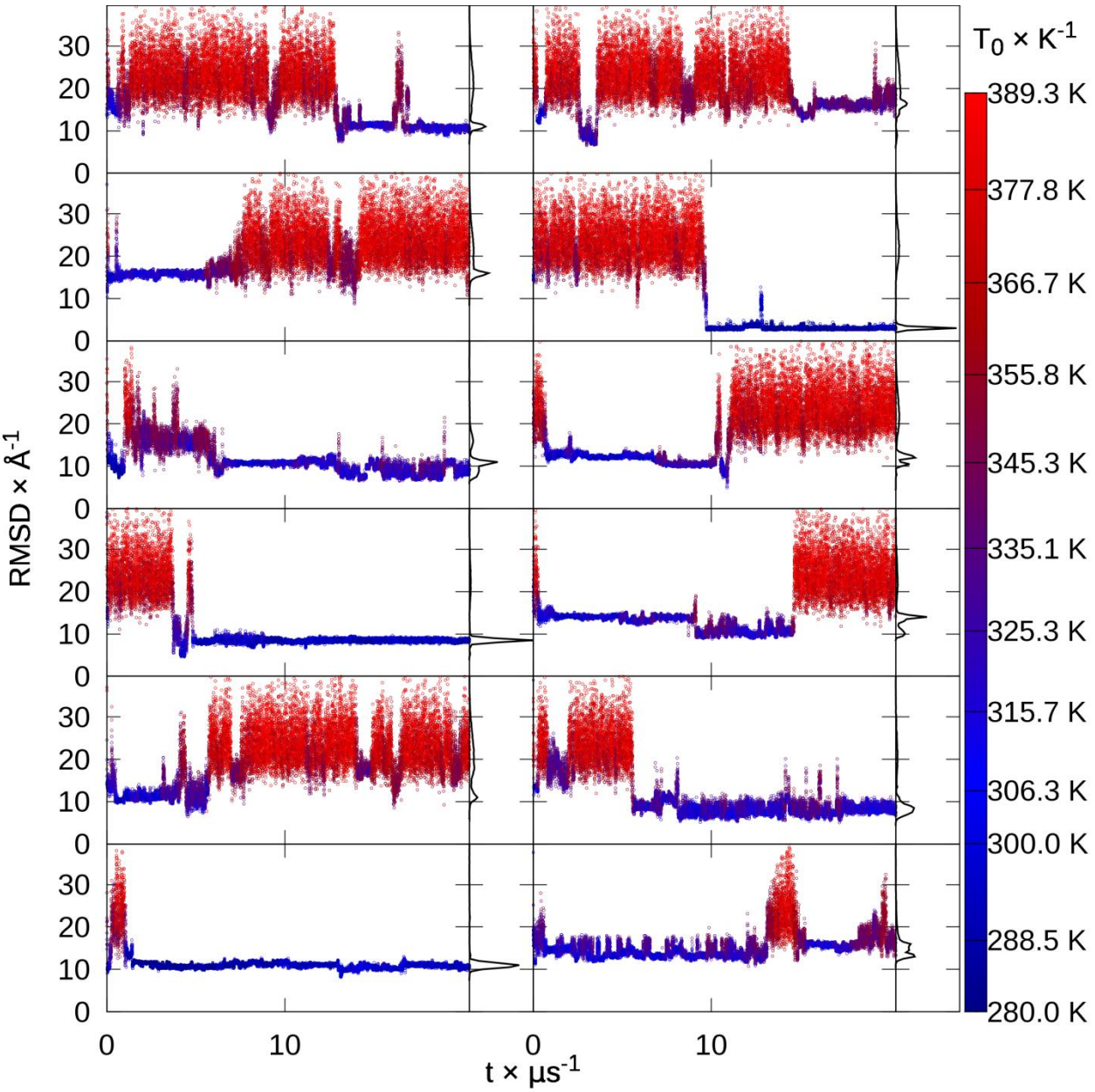


**Figure 3.S64.** Top7 replica RMSDs. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

| | | | | | |
|---|---|---|---|---|---|
| Cluster population (%) | 35.9 | 24.1 | 19.0 | 3.2 | 2.2 |
| Centroid Cα | 11.2 | 2.7 | 8.3 | 13.9 | 8.3 |

| RMSD (Å) | | | | | |
|---|---|---|---|---|---|

**Table 3.S21.** Top7top 5 extended REMD cluster populations and centroid Cα RMSDs.

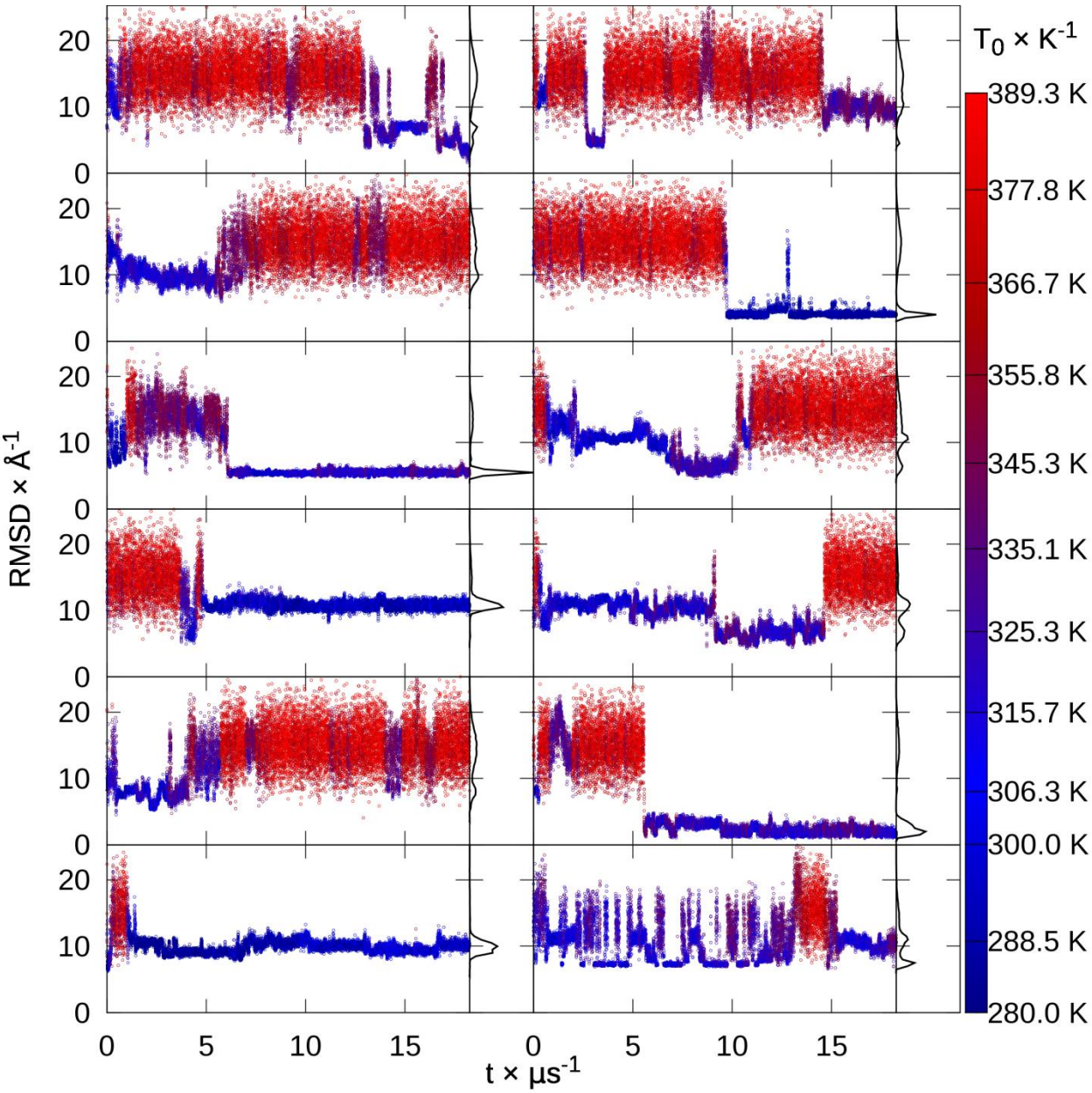


**Figure 3.S65.** Top7 replica RMSDs, residues 1 to 40. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

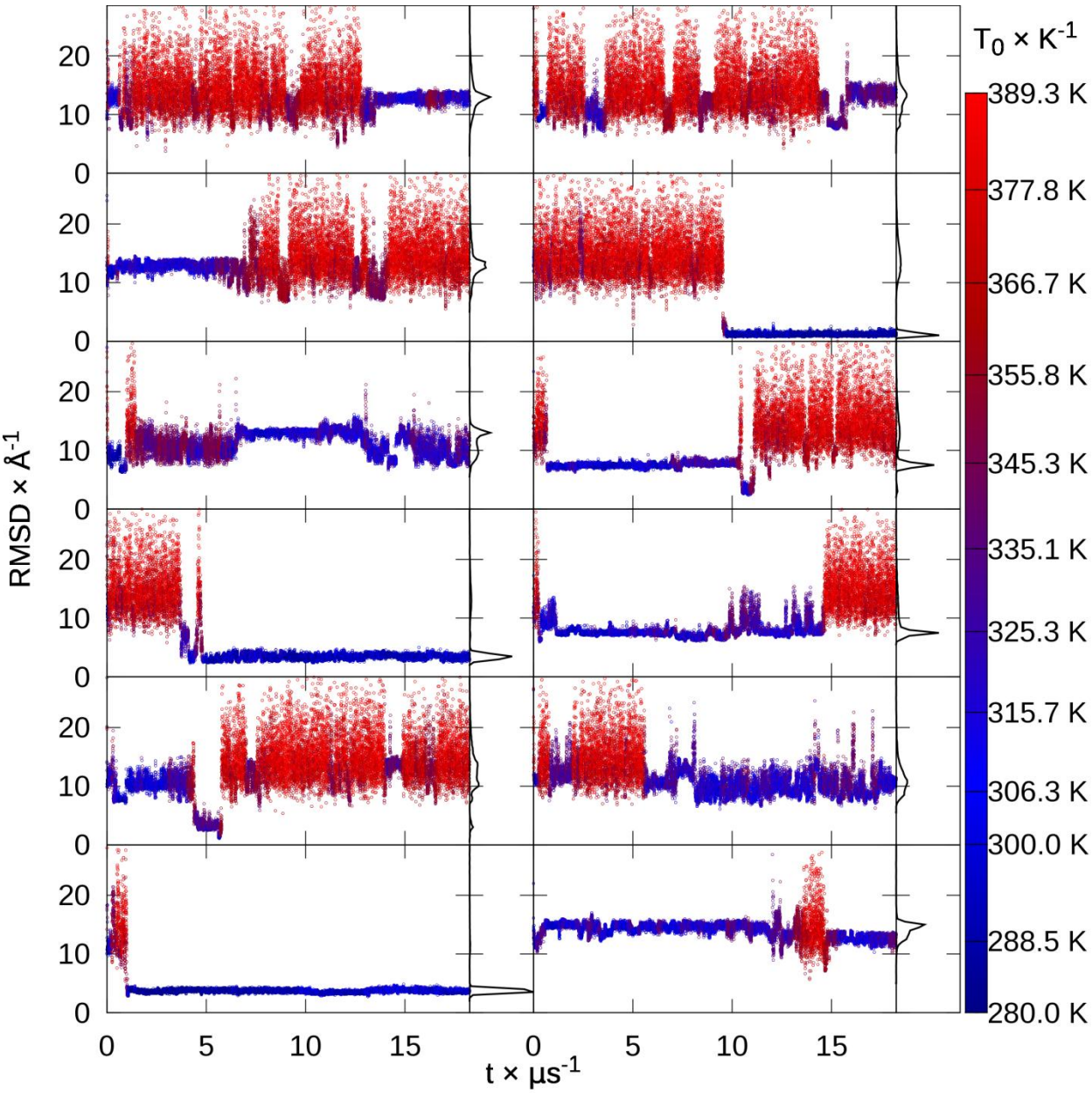


**Figure 3.S66.** Top7 replica RMSDs, residues 42 to 92. RMSD to native of each replica from extended replica exchange versus time, colored by snapshot temperature from blue to red, with histograms.

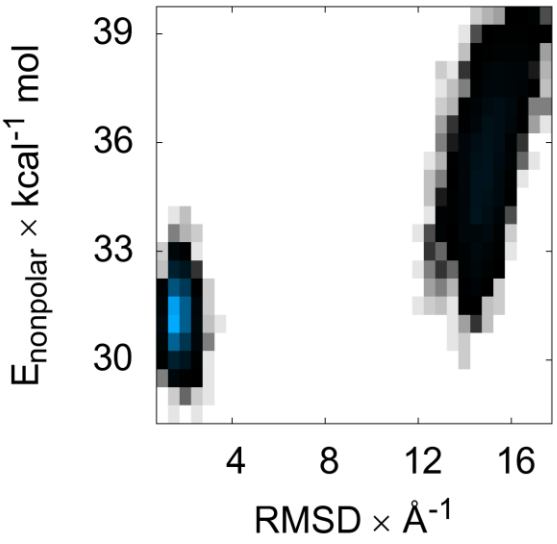


**Figure 3.S67.** Top7 surface area energy versus RMSD. Color indicates the histogrammed population in each 0.5 Å by 0.5 kcal $mol^{-1}$ bin, going from white (no population) to black (1% of maximum bin population) and then to blue (maximum bin population). The correction for the solvent-accessible surface area, determined by recursively optimizing spheres around each atom starting from icosahedra, is more favorable at low (1–3 Å) RMSD.

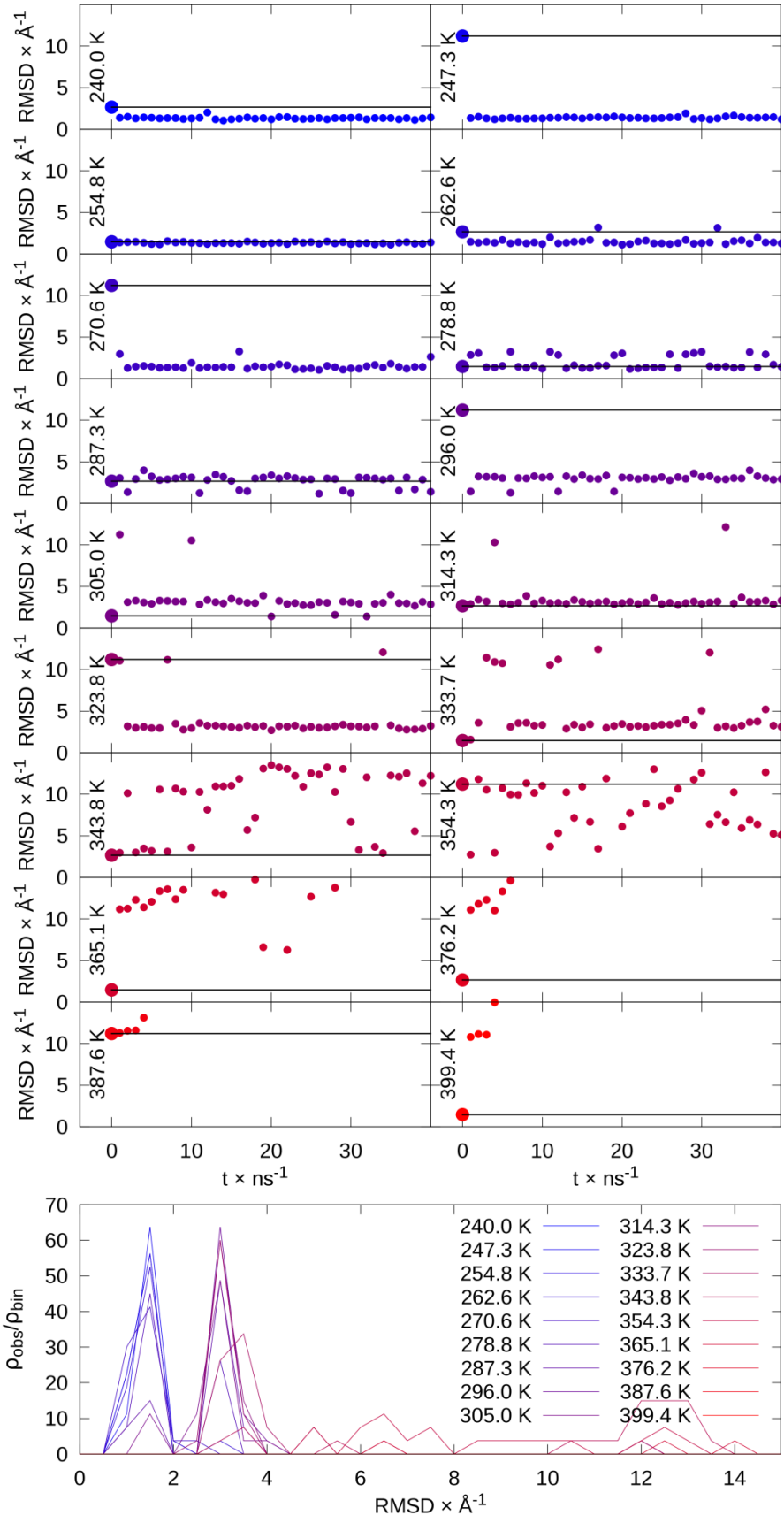
RMSD × Å-1
240.0 K
247.3 K
254.8 K
262.6 K
270.6 K
278.8 K
287.3 K
296.0 K
305.0 K
314.3 K
323.8 K
333.7 K
343.8 K
354.3 K
365.1 K
376.2 K
387.6 K
399.4 K
t × ns-1
ρobs/ρbin
RMSD × Å-1

**Figure 3.S68.** REMD sorting of Top7 conformations: partly folded (2.7 Å), unfolded (11.2 Å), and native-like (1.5 Å), repeated for 18 replicas. At top, RMSD vs time for each temperature shows sorting of native-like conformations to low temperatures, partly folded to intermediate temperatures, and unfolded to high temperatures. Lines indicate initial RMSD value that each temperature. At bottom, histogram shows preference of native-like conformations at low temperatures and partly folded conformations at intermediate temperatures.

# Chapter 4. Refinement of Generalized-Born Neck Parameters for Nucleic Acid and Their Complex with Protein

**Acknowledgments.** This chapter is direct excerpt from "Nguyen, H.; Pérez, A.; Simmerling, C., Refinement of Generalized-Born Neck Parameters for Nucleic Acid and Their Complex with Protein". Pérez and Simmerling revised the manuscript. Nguyen thanks He Huang, Sherry Bermeo for critical reading of this manuscript.

**Abstract**: Although the Generalized–Born (GB) model is widely used for protein Molecular Dynamics (MD) simulations, there is limited usage for nucleic acid simulations. Most GB models are not able to keep stable nucleic structure and others introduce structural bias in protein simulation that leads to artifact in nucleic acid and protein complex simulation. In this study, we propose a refitting procedure for a recently developed GB-Neck model by designing broad training sets and extending the number of empirical parameters. This new parameter set, named GB-Neck2nu, significantly reduces energy error to Poisson Boltzmann calculation for both absolute and relative energy compared to its ancestor GB-Neck model for both nucleic acids and their complex with protein. The improvement in solvation energy calculation translates to increased structural stability in DNA and RNA duplexes, quadruplex simulations and in protein-nucleic acid complexes. The GB-Neck2nu model also successfully folds small DNA and RNA hairpins to near native structures as determined from experiment. The robustness of GB-Neck2nu model is also shown by producing the correct ligand binding site in DNA minor groove for tested system.

## 4.1 Introduction

The interest in computational models of nucleic acids has spiked recently thanks to genomics and epigenomics projects and the implications in health (e.g. cancer). This has been reflected by increased number of publications describing nucleic acid simulations.[35] Nucleic acids have been traditionally challenging in simulations due to their highly charged backbones and the importance of bound ions.[164] The incorporation of Particle Mesh Ewald[89] in explicit solvent simulations allowed for the first time stable simulations of nucleic acids[165]. Since then, simulations in explicit solvent have been quite standard.

Explicit solvent simulations are the state of the art in protein and nucleic acid simulations, however there are many reasons why one would like to use the more approximate implicit solvent simulations, especially pairwise Generalized-Born (GB) solvent model[21, 23-24]: (i.) lower number of particles resulting in faster simulation times and more overlap[166] in replica exchange molecular dynamics[2] (REMD), (ii.) much higher performance on standard GPU-based computer architectures[3, 48] (breaching the microsecond/day barrier[167]), (iii.) much more sampling thanks to low solvent viscosity,[1a] (iv) better scaling with the number of CPU.[4] For proteins implicit solvent simulations have become quite standard.[1a, 5b, 50] For nucleic acids implicit solvent is even more important: the linear geometry of nucleic acids (as opposed to the globular one of proteins) makes for very big water boxes,[35] with a ratio of biomolecule to water much lower than in the case of proteins. In essence this means that most of the computer time in explicit solvent simulations of nucleic acids involves water-water interactions.

To the best of our knowledge, there are only few GB models that can maintain stable nucleic acid simulations.[36] Three of them are widely in CHARMM program[25] (GBMV,[18] GBMV2[20] and GBSW[26]) while two others (GB-HCT,[21] GB-OBC[23]) are widely used in AMBER program.[22b, 111, 168] Examples of their applications for DNA and RNA simulations can be found cited articles.[4, 35-36, 169] GBMV and GBMV2 are arguably among the most accurate GB models (in term of reproducing solvation energy of higher theory level such as Poisson Boltzmann method[77]) but the accuracy comes with slow speed in MD simulation since those models use computationally expensive molecular surface to define solute/solvent boundary.[14] Additionally GBMV and GBMV2 use a sharp molecular surface boundary between solute and solvent; this introduces unstable force calculation in long time scale simulation.[170] It is suggested to use small time step of 1 fs to achieve stable nucleic duplex simulation[170] while the general time step is

2fs[29, 56, 170] or even 4 fs.[167, 171] This even makes GBMV and GBMV2 much slower in MD simulation. GBSW[26] is an analytical version of GBMV and GBMV2 using Van der Waals (VDW) surface to define the solute/solvent boundary. GBSW's fast speed in solvation energy calculation comes with much less accuracy as compared to GBMV models.[14]

Two other GB models that can work with nucleic acid are implemented in AMBER program: GB-HCT[21] and GB-OBC[23]. Those models are based on pairwise approximation approach introduced by Hawkins et al.[21] Latter model (GB-OBC) introduced correction parameters to reduce the overestimation of solvation energy of former model (GB-HCT). Those two pairwise models use Van der Waals (VDW) surface to define solute/solvent making them much faster than GBMV (using molecular surface) with the trade-off lower accuracy.

Among all tested GB models, GBMV models and GB-OBC are the most accurate models in solvation energy calculation for proteins if using higher theory level Poisson-Boltzmann calculation as benchmark.[14] GB-OBC is in practice more suited for long MD simulation thanks to its fast speed and its suitability for parallel calculations.[4, 48]

Recently Gaillard et al.[36] compared different GB models and they concluded that GBMV2 and GB-HCT models are better in reproducing DNA parameters (such as major and minor groove width) from experimental data than GB-OBC model. However, GB-HCT (and even GB-OBC) introduced strong helical structural bias in protein simulation,[31] preventing its application from protein/DNA or RNA binding complex simulation. Another pairwise GB model such as GB-Neck[24] model was developed by introducing correction to GB-OBC to mimic the molecular surface (which is more realistic) while keeping similar speed in calculating solvation force. Theoretically GB-Neck should be promising approach in achieving both reasonably good accuracy and fast speed. However GB-Neck was shown to disfavor native structure of either protein or nucleic acid.[24, 36, 49] Overall it is clear that there is currently no fast and numerically stable GB model that works well with protein and nucleic acid at the same time. We address this issue in the current work.

Our work on a nucleic acid-compatible solvent model closely follows our recent work at extending and refitting the GB-Neck model for peptides and proteins.[29] The new model for proteins, named GB-Neck2[29], has better agreement to PB method in solvation energy calculation or to explicit solvent in reproducing secondary structure profiles and salt bridge strength as compared to the older models in Amber (GB-HCT, GB-OBC, GB-Neck)[29] or to experiment in

quantitatively reproducing thermodynamic profile for different small peptide motifs such as hairpin or mix of alpha helix, 3-10 helix, PP2. GB-Neck2 was also shown to successfully fold a series of μs to millisecond time scale folding proteins[167] which are considered to be a hard problem for current computer power using explicit solvent MD simulation.[172] This model is based on the functional form of GB-Neck model[24] but adding flexibility to the model in the form of per atom parameters. The process of training and testing new parameter was also iteratively designed. Continuing this work, we introduce the GB-Neck2 parameters to work with nucleic acids while they can be combined with protein GB-Neck2 parameters. This sets the stage for stable simulations of protein-nucleic acid complexes.

We have tested several options to fit nucleic acid parameters for GB. First, we fitted only Phosphate (P) parameters while for other elements, using the same atom parameters from the GB-Neck2 model for both protein and nucleic acid .[29] Secondly, we also tried to refit GB-Neck2 parameters by using the same parameter set for both protein and nucleic acid and ignoring all previous GB-Neck2 parameters. Thirdly, we introduced nucleic atoms their own parameters rather than reusing the protein ones and then refitted using an extensive training set, comprising various DNA and RNA motifs such as duplex, pseudoknot, ribosomal RNA etc. In all three versions, we unfortunately did not get sufficiently low energy error between the GB and PB calculations for the training sets and this made the DNA/RNA duplex strands quickly unfold in MD simulations (data not shown). Therefore we decided to focus on improving parameters for duplex structures; its derivation such as hairpin structure, quadruplex and their complex with protein.

In the results section we show how in addition to improving the behavior for isolated nucleic acids, we are also able to carry out stable simulations of protein-nucleic acid complexes. This covers a major limitation in the field opening the door to the study of major complexes like the nucleosome core particle, promoter-DNA interactions, drug-nucleic acid binding. Finally, we show an example of the ability of GB-Neck2nu to reproduce the folding of a DNA, RNA hairpin and to reproduce the ligand binding to DNA minor groove.

## 4.2 Methods

### 4.2.1 Generalized-Born theory

In implicit solvent model, solvation energy is normally decomposed to two terms, polar and nonpolar: $\Delta G_{solvation} = \Delta G_{polar} + \Delta G_{np}$. Nonpolar term can be roughly approximated by $\Delta G_{np} = \sigma * A$ (where $\sigma$ is surface tension coefficient and the A term is solute surface area) although there are more sophisticated approaches.[9, 12, 28] Since the solvation energy is dominated by the polar part (particularly for highly charged nucleic acid),[11] most of the effort was spent in developing more accurate polar model.[20, 23-24, 26, 29]

Polar solvation energy can be calculated from the very accurate, but computationally expensive, Poisson Boltzmann (PB) method[77] or from much faster Generalized Born (GB) model. GB model approximates the polar solvation energy by summing energies of pairwise atomic interactions as well as self-interaction. The GB equation was first introduced by Still et al.[15] and it has been the basis for other GB model developments (eq. 4.1)

$$\Delta G_{GB} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\sum_{i,j}\frac{q_i q_j}{f_{ij}^{GB}(r_{ji})} \qquad (4.1)$$

Here $q_i$, $q_j$ are the partial charges of atom i and j with distance of $r_{ij}$. Function $f_{ij}^{GB}$ is defined by

$$f_{ij}^{GB}(r_{ij}) = \sqrt{r_{ij}^2 + R_i R_j \exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)} \qquad (4.2)$$

$R_i$ and $R_j$ are so-called effective radii of $i^{th}$ and $j^{th}$ atoms. They represent the degree of burial for each atom inside the solute. Effective radius of a given atom can be exactly calculated by solving the PB equation for the charge of the interested atom only (eq. 4.3). The effective radius calculated from PB is defined as 'perfect' radius. "Perfect" radii were shown to yield best agreement[16] between GB and PB energies if they were applied in the GB equation (eq. 4.1)

$$R_i = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\frac{q_i^2}{\Delta G_{self(i)}} \qquad (4.3)$$

In GB models, effective radii can be calculated using either the Coulomb Field Approximation (CFA) or non-CFA approach.[17] Although the former is notorious for

overestimating effective radii,[17, 19] it still is widely implemented in MD simulation due to its simple approximation that makes it easy to derive the analytic form of calculation effective radii. The more accurate non-CFA-based GB model, such as GBMV or GBMV2[20, 26] or recently developed R6 model, shows excellent agreement to PB calculation[17] but the slow calculation limits this from extensive use in MD simulation . Additionally, GBMV and GBMV2 uses a sharp molecular surface boundary between solute and solvent that leads to unstable numerical calculation in long time scale simulation.[170] The development of the analytical form of the R6 model still focuses on small molecule calculation and has not been extensively tested on protein simulation.[27]

Our work previously focused on improving the accuracy of CFA-based GB-Neck model[24] by introducing rigorous parameter training and testing for protein simulation.[29] Based on CFA, effective radii can be approximated by equation 4.[15]

$$R_i^{-1} = \rho_i^{-1} - I \qquad (4.4)$$

where I is 3D integral defined by $I = \frac{1}{4\pi}\int_{r>\rho_i} |r|^{->} dV$ (4.5). r is a vector center at atom i and the integral region stays inside the molecule but outside the atom i. $\rho_i$ is intrinsic radius of $i^{th}$ atom. Depending on the type of boundary between solute and solvent, integral region could be molecular volume ($I_{MS}$) or Van der Waals volume ($I_{vdw}$). Van der Waals (VDW) volume approach is more favorable because it is expensive to calculate effective radii using molecular volume.[14] Hawkins et al. followed VDW approach and introduced pairwise approximation to analytically calculate the effective radii (GB-HCT model).[21]

$$R_i^{-1} = \rho_i^{-1} - I \approx \rho_i^{-1} - I_{vdw} \approx \rho_i^{-1} - \frac{1}{4\pi}\sum_j \int_{|r_{ij}-r|<\rho_i} |r|^{-4} d^3r \qquad (4.6)$$

$I_{vdw}$ is approximated by summing all individual integrals contributed by atom $j \neq i$. To avoid the overestimation of $I_{vdw}$, a set of scaling factor $S_x$ (x = H, C, N, O, P, S …) was introduced ($\rho_i \rightarrow S_i * \rho_i$). However this approach neglects the interstitial region between atoms, which leads to underestimates of the effective radii for deeply buried atoms, in contrast to PB calculation, which uses molecular surface to define the solute/solvent boundary.[23] Onufriev et al.

introduced additional set of parameters (α, ß, γ) to empirically scale up the effective radii of those atoms (GB-OBC model).[23]

$$R_i^{-1} = \widetilde{\rho}_i^{-1} - \rho_i^{-1}\tanh(\alpha\,\omega - \beta\,\omega^2 + \gamma\,\omega^3) \qquad (4.7)$$

where $\widetilde{\rho}_i = \rho_i - \text{offset}$, $\omega = \widetilde{\rho}_i * I_{vdw}$. Mongan et al. later added $I_{neck}$ correction to $I_{vdw}$ to mimic the molecular surface boundary $I_{"MS"} = I_{vdw} + I_{neck}$ (GB-Neck model).[24] $I_{neck}$ is easily approximated following Mongan et al. approach.[24] The ω term can be re-calculated by $\omega = \widetilde{\rho}_i * I_{"MS"}$. To minimize the overlap of "neck" region, a scaling neck factor $S_{neck}$ was introduced. The GB-Neck model is theoretically better than the GB-OBC and GB-HCT models, but it was shown to quickly unfold the native structure in either protein or nucleic acid MD simulation.[24, 36, 49] We previously redesigned the training set and test set and performed more rigorous refitting GB-Neck parameters for protein. The new parameter set, named GB-Neck2, is better than previous models (GB-OBC, GB-Neck) in reproducing PB solvation energy and reproducing explicit solvent MD data such as secondary structure content.

Following the success of GB-Neck2 model for protein simulation, we also made [α, ß, γ] parameters (introduced by Onufriev et al.[23]) element-dependent. There are 20 parameters to be fitted, which are 5 scaling factors Sx (introduced by Hawkins et al.)[21] (with x=H, C, N, O, P) and 5 sets of $[\alpha, ß, \gamma]_x$ (x=H, C, N, O, P). The *offset* (introduced by Still et al.[15]) and $S_{neck}$ parameter[24] are kept to be identical to the values in GB-Neck2 model so that both nucleic acid and protein parameters can be combined in protein/nucleic acid complex MD simulation.

### 4.2.2 Fitting procedure

#### 4.2.2.1 Objective function

Twenty parameters $[S, \alpha, ß, \gamma]_x$ (x=H, C, N, O, P) were treated as variables in the objective function (eq. 4.8). Objective function is the sum of weighted normalized root-mean-square-deviation (RMSD) between GB and PB absolute energy, relative energy and the inverse of effective radii for different structure sets. The weighting factors are given to avoid any specific structure set bias. They were chosen in the similar way we have done previously.[29]

$$\text{obj\_funct} = w_{abs}\sum_i \text{abs\_rmsd}_i/\text{natom}_i + w_{rel}\sum_i \text{rel\_rmsd}_i/\text{natom}_i + w_r * \text{rad\_rmsd} \qquad (4.8)$$

Here abs_rmsd and rel_rmsd are absolute and relative energy RMSD, respectively, between GB and PB calculations. rad_rmsd is the RMSD between the inverse of GB and PB effective radii. $w_{abs}$, $w_{rel}$ and $w_r$ are weighting factors for abs_rmsd, rel_rmsd and rad_rmsd respectively. abs_rmsd, rel_rmsd are normalized by being divided by the number of atom ($natom_i$) for each training set.

Due to very large number of variables and very expensive objective function, we were seeking the best local minimized function values rather than the global value. The local optimization method NEWUOA[173] was chosen for objective function minimization because of its quick convergence compared to other local optimization methods.[174] Additionally, NEWUOA is an improved version of UOBYQA[80] which was successfully used for refitting protein parameters in our previous work. To make sure we get the best result as we can, a total of ~2400 optimization runs were carried out for each round of fitting; each optimization run started from a random guess in the following boundary $S_x \in [0.0, 2.0]$, $[\alpha_x, \beta_x, \gamma_x] \in [0.0, 5.0]$ where x = H, C, N, O, P. Weighting factors for radii and relative energies are also varied relative to absolute energies to see how they affect the fitting ($w_r$ = 1.5, 2.5, 5.0; $w_{rel}$ = 5.0, 10.0; $w_{abs}$ = 1.0). There are 5 rounds of fitting in which later round has more structures in training set, designed by the following protocol.

**4.2.2.2 Training set for parameter fitting**: To avoid over fitting due to a large number of parameters, we include as many structure variations as possible in training. We followed the iteratively designing procedure: (i) We first trained parameters by using only DNA duplex structures from the older GB models (GB-HCT,[21] GB-Neck[24]) and TIP3P MD simulations. The initial training has 200 structures, which were equally extracted from 10 ns MD simulations at 300 K in TIP3P and GB-HCT, starting from both canonical A and B forms and 1 ns MD simulations of GB-Neck2 starting from A-form. This initial training set was designed to have both 'good' (associated DNA dimer strands from TIP3P and GB-HCT MDs) and 'bad' (dissociated DNA dimer strands from GB-Neck MD) so that the new GB model does not bias any specific structure. However, the newly optimized parameter set from this original training set favor different compacted structure with wrong H-bond pattern from its own MD starting from canonical DNA A-form. We observed similar trends while optimizing our protein solvation model[29] and this was overcome by iteratively increasing the training set's size: After each round

of fitting, we performed 0.5 to 1.0 μs MD simulations for DNA and RNA structures in training set, we then equally extracted 50 to 100 structures from MD runs and introduced them into training set. The objective function was then minimized again with the new updated structure set. This procedure is repeated until the rel_rmsd difference between two consecutive fitting rounds is small (<2% of rel_rmsd of the former fitting round). Final training set has structures which were from TIP3P, GB-HCT, GB-Neck and GB inter-parameter set MD simulations.

Since the optimizations are expensive and needed several rounds of fitting each round includes: (i) calculating expensive PB energies, (ii) refitting the parameters, (iii) running long MD simulation (0.5 to 1 μs), (iv) adding resulting structures from the new simulations to training set. We chose a small size 10-base pair DNA $(CCAACGTTGG)_2$ and its complementary RNA $(CCAACGUUGG)_2$ duplex for training energy. These training sets are named dnadup and rnadup, respectively. Both of these DNA and RNA duplexes have base types (A, T, C, G or U) and are each long enough to form a complete duplex. Traditionally they were also extensively studied in GB simulation.[4, 169a, 169b, 175] The A and B-forms of DNA duplex $(CCAACGTTGG)_2$ were used for training effective radii (this set is named dnadupRad). The diversity of training sets are shown in supplement.

#### 4.2.2.3 Test set for comparing solvation energy between GB and PB

Following our previous work on proteins, we also designed two types of structure sets to test the transferability of the new GB parameters. Type I test set has all the structures in training and the structures from MD simulation using the final GB parameters. This test set was designed to check if the final GB parameters did not bias any structures coming from its own MD simulation. Type I have dnadup_plus150 (adding 150 structures that were equally extracted from 0.75 μs MD simulation of DNA duplex $(CCAACGTTGG)_2$ using the final GB parameter set to dnadup) and rnadup_plus200 (adding 200 structures that were equally extracted from 1.0 μs MD simulation of RNA duplex $(CCAACGUUGG)_2$ using the final GB parameter set to rnadup). Type II test set has structures with sequences different from the training set. Each test set for nucleic acid duplexes has structures coming from TIP3P MD simulations at 300K and from intermediate GB model developed during this project. Test set type II has structures of the Dickerson-Drew dodecamer DNA duplex $(CGCGAATTCGCG)_2$[176] (DNA DD) and its complementary RNA duplex $(CGCGAAUUCGCG)_2$ which are two popular DNA and RNA

models for experimental and computational studies .[177] We also used structures from the GCC-box binding domain in complex with DNA (PDB ID: 1GCC[178]) for testing the combination of GB-Neck2nu parameters for nucleic acid (this work) with GB-Neck2 parameters for protein.[29] 1GCC test set includes structures from 300 K and 500 K TIP3P simulations. High temperature was used to get the structure variety.

To avoid lengthy description, we characterize the training sets and test sets by their RMSD to their experimental structures and give further detail in the supplement.

#### 4.2.2.4 Test set for MD simulations

Comparing the agreement between GB and PB calculation is only initial step to justify the performance of a GB model. We first test if GB-Neck2nu is able to maintain stable DNA/RNA duplex and DNA/protein complex that was not seen in GB-Neck model.[24, 36] We also further test the stability of non-duplex system such as DNA quadruplex. Besides testing the structural stability, we also tested structural conversion, such as A to B form DNA, B to A form RNA. The folding of DNA/RNA hairpin and the process of ligand binding to minor groove of DNA duplex are also tested to show that GB-Neck2nu is suitable for various systems and applications beyond canonical duplexes. The summary of testing systems is given in table 4.S3.

Since GB-Neck model is infamous for destabilizing the duplex, we will test if GB-Neck2nu can keep DNA and RNA duplexes stable. Our structure stability testing will be compared to the available TIP3P simulation data (either from our own simulations or from previous work given in corresponding citation). The testing structures include DNA duplex $(CCAACGTTGG)_2$ and RNA duplex $(CCAACGUUGG)_2$ which were used in training parameters. It also includes popular Dickerson-Drew dodecamer (DD) DNA duplex $(CGCGAATTCGCG)_2$ and RNA duplex $(CGCGAAUUCGCG)_2$. Besides those structures, we further tested the longer DNA sequence $(CTAGGTGGATGACTCATT)_2$ (corresponding to "seq2" in Pérez et al.[179]). We also further tested for protein/nucleic acid complex simulation by choosing a protein/DNA complex system (PDB ID: 1GCC[178]). There are two runs for all DNA and RNA duplexes (except seq2 DNA) starting from both canonical A and B forms. The lengths of MD simulations in GB-Neck2nu are between 50 ns (Protein/DNA complex) to 1 μs, while TIP3P MD length is from 50 ns (Protein/DNA complex) to 100 ns. Since DNA and RNA duplexes were previously reported to be stable in μs timescale in TIP3P MD simulation with

bsc0 force field[177] (DNA) and bsc0$\chi_{OL3}$[180] (RNA), we only performed short MD simulations (100 ns) for TIP3P explicit solvent.

We also tested the folding of DNA and RNA hairpin. Our goal in developing a GB model is not just focusing on keeping DNA/RNA duplex stable, but also applying it to MD simulation of the system with large structural change, which requires lots of water molecules in explicit solvent. GB simulation's speed does not depend on the shape of the system. We choose a small DNA and RNA hairpin for testing the conformation change. Besides performing simulations for GB-Neck2nu, we did additional runs for GB-HC T model.[21] This model was developed 20 years ago and it is the foundation of later pairwise models (e.g. GB-OBC, GB-Neck, GB-Neck2). GB-HCT has been shown for strongly biasing compacted structure in protein MD simulations. [31] However, it is still being used for simulating DNA duplexes, RNA duplexes, and hairpin structures, since this model can keep duplexes stable.[36] We hypothesized that this GB model would still bias more compacted structures (compared to native one) in long time scale simulations of nucleic acid and the stability of duplex simulation in GB-HCT is just kinetically trapped. We chose a small DNA and RNA hairpin systems to test our hypothesis. We will test if two models can correctly predict the experimental structure of the hairpin. DNA hairpin with GCA loop (sequence GCGCAGC, PDB ID of its homologue: 1ZHU)[181] will be used as testing systems. This system is small enough to enable us to have very long μs simulations to observe the structural changes. This DNA hairpin was also successfully folded (to 1.5 Å heavy atom RMSD to NMR structure) from TIP3P simulation.[182] We also tested the folding of a RNA hairpin UUCG loop (PDB ID: 2KOC,[183] sequence GGCACUUCGGUGCC) with 5 base pairs in the stem. This hairpin system was shown to be stable in explicit solvent simulation in explicit solvent simulation.[180a]

Having more accurate implicit model for nucleic acids also opens the door to ligand binding calculations, often performed in protein-ligand studies.[184] Here we also show an example of how a ligand is able to diffuse into the correct binding site of the Dickerson-Drew dodecamer (PDB ID: 1D30).

### 4.2.3 PB calculation

We used similar input from our previous parameterization of proteins for PB solvation energy and 'perfect' radii calculation.[29] The mbondi2 radii[23] was used to define the boundary between solute/solvent for all PB calculations.

### 4.2.4 Simulation protocol

#### 4.2.4.1 Basic setup

Popular force field 99SB[57] was used for protein while force field bsc0[177a] was used for DNA and bsc0$\chi_{OL3}$[180] was used for RNA simulations. Canonical A and B-forms of DNA and RNA duplexes were built by NAB program in Ambertools 12.[168] Topologies and coordinates for MD simulations were generated by LEaP.[168] All MD simulations were carried out by using either sander or pmemd program in Amber version 12 or 14.[111, 168] Long μs MD runs were performed with pmemd GPU version.[48] Amber code was modified to accept new GB parameters. The combination of GB protein parameters (GB-Neck2)[29] with GB nucleic acid pars can be accessed by specifying igb = 8 in Amber 14 or later version. Before production run, all simulations followed this protocol: (i) minimizing structure, (ii) equilibrating. MD simulations were performed at 300 K with time step of 2 fs. SHAKE was used to constrain bonds having hydrogen. GB simulations used Langevin thermostat with no cutoff while TIP3P simulations used Berendsen thermostat[88] with Particle Mesh Ewald (PME) method[89] for long range interaction with cutoff of 8 Å. Systems in explicit solvent simulation were solvated by a TIP3P truncated octahedron box with a buffer size of 10 Å. Since there was no explicit ion in GB simulation, we did not include ion in TIP3P simulations to have a more direct comparison. However, the TIP3P trajectories from other groups used ion. The details of TIP3P simulation was given in the corresponding citation. All GB-Neck2nu MD simulations used mbondi2 radii[23] for nucleic acid and mbondi3[29] for protein. Radii set mbondi3 is a small adjustment of mbondi2 for GB-Neck2 to correctly reproduce the TIP3P PMFs of salt bridge profile of Arg (and Lys) and Glu (and Asp).[29] For nucleic acid, mbondi3 should be identical to mbondi2. GB-HCT MD used a suggested mbondi radii set with offset = 0.13.[4]

We performed MD simulation for all tested structures except the RNA hairpin UUCG loop. This structure has 5 base pairs in the stem and is very stable in MD simulation (data not shown). We then used replica exchange molecular dynamics (REMD)[2] simulation to accelerate

the sampling for this system. Each run has 6 replicas (3-4 μs/replica) with temperature of [300.0, 317.0, 334.9, 353.9, 373.9, 395.1] to give acceptance ratio of 0.25. Exchange was attempted every 1 ps. Simulations started from both NMR structure and A-form conformations.

#### 4.2.4.2 Equilibration

In GB equilibration, the starting structures were first minimized in 500 steps and then were heated from 100 K to 300 K with 10.0 kcal/mol/Å$^2$ atomic positional restraints on heavy atoms. In the next 3 steps (250 ps each), the temperature was kept at 300 K and the restraint force constant was reduced from 10.0, 1.0 to 0.1 kcal/mol/Å$^2$. In TIP3P equilibration, the solvated structure was minimized in 10000 steps, then was heated from 100 to 300 K in NVT ensemble, then was equilibrated in NPT ensemble with 100.0, 10.0, 1.0 and 0.1 kcal/mol/Å$^2$ positional restraints for heavy atoms in next four 250ps-steps. Production runs were performed in NVT ensemble.

#### 4.2.4.3 Ligand binding simulation

Ten MD simulations were performed for DNA DD with its ligand 6-amidine-2-(4-amidino-phenyl) indole (DAPI) (PDB ID: 1D30[185]). General Amber force field (GAFF)[186] with AM1-BCC charge model[187] was used for ligand. The ligand topology was prepared by Antechamber[188] in Ambertools 12.[168] Initially ligand was taken out of its binding site in DNA minor groove with O2 (DT7) - HN1 (DAPI) distance 36 Å. Distance O2 (DT7) - HN1 (DAPI) is chosen to monitor the binding process since O2 (DT7) - HN1 (DAPI) forms H-bond in X-ray structure. To avoid ligand diffusing far away from DNA, a distant restraint (force constant of 0.1 kcal/mol/Å$^2$) was applied if the distance between P (DA18) and C2 (DAPI) is larger than 100 Å (which is 2 times longer than DNA length).

#### 4.2.4.4 Data analysis

Backbone RMSD (BB-RMSD) calculation and cluster analysis were carried out by the ptraj and cpptraj[189] program in Ambertools version 12 and 14.[111, 168] Cluster analysis used means algorithm with BB-RMSD as metric. The whole trajectory for each simulation was grouped into 50 clusters. Backbone atoms in DNA/RNA are defined as heavy atoms in the phosphate group and in the sugar pucker. Heavy atoms in the ligand are considered as back bone atom. DNA and

RNA helical parameter analysis were performed by CURVES+ program (version 1.31).[190] The major and minor groove width in the outputs from CURVES+ were added 5.8 Å to account for the P diameter as suggested.[190] The H-bond fraction is defined as the ratio between the number of H-bonds of each trajectory frame and the starting structure. Average of H-bond fraction is endcalculated for whole trajectory. The number of H-bond for each base pair was calculated by using "nastruct" command in cpptraj.[189]

### 4.3 Result and Discussion

#### 4.3.1 Parameter fitting

Our goal is to extend and re-optimize the parameters for GB-Neck model using PB solation energy (absolute energy and relative energy between structure pair) and 'perfect' radii as benchmark. We designed the objective function as the sum of weighted contribution from energy and effective radii RMSD between GB and PB calculation similar to our previous work for protein parameters.[29] A total of ~2400 optimization runs were performed for minimizing objective function with 6 combinations of weighting factor ($w_r$ = 1.5, 2.5, 5.0; $w_{rel}$ = 5.0, 10.0; $w_{abs}$ = 1.0) for each round of fitting. Each optimization started from a random guess within the boundary given in method section. We stopped minimization runs after 5 rounds when obtaining close error between the old training set (in 5$^{th}$ round) and the new training set (in 6$^{th}$ round) (figure 4.S2).

Among 6 weighting factor combinations, we chose the optimized parameters from ($w_r$ = 2.5, $w_{rel}$ = 5.0, $w_{abs}$ = 1.0) as our final candidate since this combination gave the best compromise between having low error for both energy and effective radii (table 4.S4). Those weighting factors are different from the ones in protein training,[29] reflecting the difference in training set size, charge … between protein and nucleic training set. The results for the top 10 optimization runs for ($w_r$ = 2.5, $w_{rel}$ = 5.0, $w_{abs}$ = 1.0) are given in table 4.S1. The final parameter set was named GB-Neck2nu parameters and they are given in table 4.1.

**Table 4.1**. GB parameters after training for GB-Neck2nu

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $S_H$ | 1.697 | $\alpha_N$ | 0.686 |
| $S_C$ | 1.269 | $\beta_N$ | 0.463 |
| $S_N$ | 1.426 | $\gamma_N$ | 0.139 |
| $S_O$ | 0.184 | $\alpha_O$ | 0.606 |
| Sp | 1.545 | $\beta_O$ | 0.463 |
| $\alpha_H$ | 0.537 | $\gamma_O$ | 0.142 |
| $\beta_H$ | 0.363 | $\alpha_P$ | 0.418 |
| $\gamma_H$ | 0.117 | $\beta_P$ | 0.290 |
| $\alpha_C$ | 0.332 | $\gamma_p$ | 0.106 |
| $\beta_C$ | 0.197 | | |
| $\gamma_C$ | 0.093 | | |

Table 4.2 shows the abs_rmsd, rel_rmsd and rad_rmsd for individual training set. The results are compared to GB-Neck model to show the improvement. GB-Neck2nu reduced about 80% error for absolute energy and reduced 65% and 15% for relative energy of dnadup and rnadup, respectively. Figure 4.S1 shows the energy comparison for individual structures in dnadup training between GB (GB-Neck, GB-Neck2nu) and PB. GB-Neck2nu has better agreement to PB for each structure, while GB-Neck underestimates energies for most of the structures. GB-Neck only has close energy to PB calculation for more extended structures (having large RMSD to both A and B-form DNA). The same trend is also observed for rnadup training set (figure 4.S4).

The effective radii errors are also reduced to 34% and 49% (compared to the original GB-Neck model) for A and B forms of the dnadupRad training set. Figure 4.S8 shows better correlation between GB-Neck2 effective radii and PB 'perfect' radii than the ones in the GB-Neck model. GB-Neck tends to overestimate effective radii for most atoms. This trend is also observed in the protein set.[29]

**Table 4.2.** abs_rmds, rel_rmsd aand rad_rmsd for individual training set. %reduced_error shows degree of improvement of GB-Neck2nu compared to GB-Neck, defined by %reduced_error =

100*(rmsdGB-Neck-rmsdGB-Neck2nu)/rmsdGB-Neck where "rmsd" is either abs_rmsd, rel_rmsd, rad_rmsd or obj_funct. "natom" is the number of atoms for each structure in the training set. Weighting factor "w" is also shown for each set.

| | Solvation energy rmsd (kcal/mol) | | | | Inverse of effective radii rmsd (1/Å) | | obj_funct |
|---|---|---|---|---|---|---|---|
| | dnadup | | rnadup | | A-form | B-form | |
| | abs_rmsd *w = 1.0* natom = 632 | rel_rmsd *w = 5.0* natom = 632 | abs_rmsd *w = 1.0* natom = 640 | rel_rmsd *w = 5.0* natom = 640 | rad_rmsd *w = 2.5* natom = 632 | rad_rmsd *w = 2.5* natom = 632 | |
| GB-Neck | 68.3 | 29.5 | 144.6 | 13.8 | 0.068 | 0.075 | 0.850 |
| GB-Neck2nu | 14.0 | 10.3 | 25.4 | 11.7 | 0.045 | 0.038 | 0.338 |
| %reduced_error | 80% | 65% | 82% | 15% | 34% | 49% | 60% |

### 4.3.2 Improving solvation energy and effective radii calculation: Comparison between GB and PB calculation

#### 4.3.2.1 Effective radii comparison

We trained effective radii for only A and B forms of DNA duplex $(CCAACGUUGG)_2$ and it is of interest if the improvement for those structures will translate to other nucleic acid structures, such as protein/nucleic acid complexes, or other DNA and RNA duplexes or non-duplexes. We chose 8 systems to test this. Two of them are DNA duplex $(CGCGAATTCGCG)_2$ in A and B forms. Four of them are RNA duplexes in A and B forms. We also tested a DNA quadruplex (PDB ID: 1L1H[191]) and a protein/DNA complex (PDB ID: 1GCC). Table 4.3 shows the RMSD of the inverse of effective radii (GB) and the inverse of 'perfect' radii (PB). Overall GB-Neck2nu modestly improved the radii. For example, rad_rmsd of A-form RNA $(CCAACGUUGG)_2$ is 0.051 for GB-Neck2nu and 0.069 for GB-Neck model.

The only special case GB-Neck2 out-performs GB-Neck is for B-form RNA. GB-Neck strongly overestimates the effective radii of B-form RNA while GB-Neck2 has better agreement to 'perfect' radii (figure 4.1). In this case, GB-Neck2nu reduced 65% of error from GB-Neck model. Manually inspecting the effective radii calculated from GB-Neck reveals that this model overestimates the effective radii for group of atoms that are close to HO2' atoms. We do not see this strong overestimation for those atoms in A, B form DNA and A form RNA since those structures are less compacted than B-form RNA. There are several reasons showing GB-Neck2nu performs better than GB-Neck for B-form RNA. Firstly, GB-Neck parameters were not

trained for nucleic acid while we explicitly trained GB-Neck2nu parameters. Secondly we followed the iterative process of designing training set (as described in Method section) where 'bad' structures (either with broken H-bond or very compacted structure) were also included.

**Table 4.3**. RMSD between the inverse of GB effective radii and the inverse of PB 'perfect' radii (1/Å)

| | GB-Neck | GB-Neck2nu | %reduced_error |
|---|---|---|---|
| A-form DNA $(CGCGAATTCGCG)_2$ | 0.05 | 0.05 | 0% |
| B-form DNA $(CGCGAATTCGCG)_2$ | 0.05 | 0.04 | 20% |
| A-form RNA $(CCAACGUUGG)_2$ | 0.07 | 0.05 | 30% |
| B-form RNA $(CCAACGUUGG)_2$ | 0.11 | 0.04 | 64% |
| A-form RNA $(CGCGAAUUCGCG)_2$ | 0.07 | 0.05 | 29% |
| B-form RNA $(CGCGAAUUCGCG)_2$ | 0.11 | 0.04 | 64% |
| DNA G-quadruplex (PDB ID: 1L1H) | 0.07 | 0.06 | 14% |
| DNA-protein complex (PDB ID: 1GCC) | 0.07 | 0.06 | 14% |

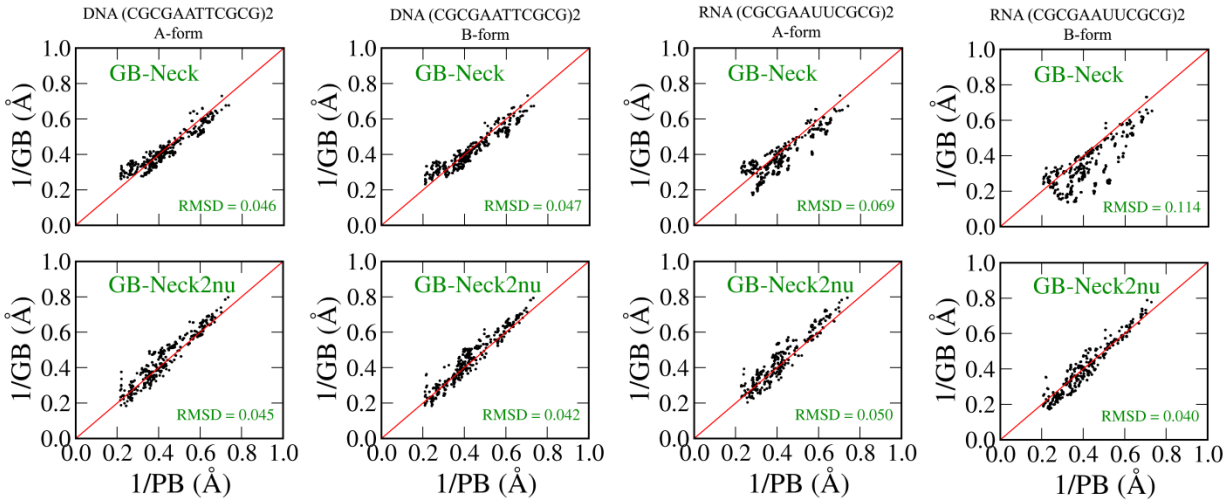


**Figure 4.1**. Comparison of inverse of effective radii between GB-Neck (top), GB-Neck2nu (bottom) and inverse of PB "perfect" radii for A and B forms of DNA duplex (CGCGAATTCGCG)2, A and B forms of RNA duplex $(CGCGAAUUCGCG)_2$. The red line in each subplot indicates the ideal agreement between GB and PB effective radii.

#### 4.3.2.2 Solvation energy comparison

To test the transferability of GB-Neck2nu parameters for solvation energy calculation from training to test set, we compared abs_rmsd, rel_rmsd for test set type I, named dnadup_plus150 and rnadup_plus200 (adding more structures from 0.75-1 μs MD simulation using the final parameters to the current training set) and type II (having sequences that are not in used in training). The test set type II include structures of DNA duplex $(CGCGAATTCGCT)_2$, RNA duplex $(CGCGAAUUCGCG)_2$ and protein/DNA complex 1GCC. Table 4.4 shows the abs_rmsd and rel_rmsd for different test sets. For test set type I, the abs_rmsd and rel_rmsd for dnadup_plus150 and rnadup_plus200 are similar to the training set dnadup and rnadup. For example abs_rmsd of dnadup and dnadup_plus150 are 68.3 and 70.8 kcal/mol respectively. This indicates that there is no new structure in MD simulation using the final parameters; adding more structures from this simulation and redo the fitting does not improve the agreement between GB and PB.

For test set type II, both absolute and relative energy RMSD are significantly reduced in GB-Neck2nu as compared to GB-Neck model. Specifically, the abs_rmsd is about 69 to 85 % reduced and the rel_rmsd is about 18 to 26% reduced. The abs_rmsd and rel_rmsd are also 69% and 18% reduced for 1GCC protein/DNA complex, although we have not included it in training. The comparison between GB and PB energies for individual structures of 3 test sets are also given in supplements.

**Table 4.4**. abs_rmsd, rel_rmsd for test set type I and II. We applied the original GB-Neck parameters for both protein and DNA (RNA) for this model.

| | | GB-Neck | | GB-Neck2nu | | %reduced_error | |
|---|---|---|---|---|---|---|---|
| | Test set name | abs_rmsd | rel_rmsd | abs_rmsd | rel_rmsd | abs_rmsd | rel_rmsd |
| Type I | dnadup_plus150 | 70.8 | 26.2 | 16.4 | 10.7 | 77% | 59% |
| | rnadup_plus200 | 144.3 | 11.1 | 21.5 | 10.4 | 85% | 6% |
| Type II | DNA duplex (CGCGAATTCGCG)2 | 104.2 | 17.8 | 15.3 | 13.6 | 85% | 24% |
| | RNA duplex (CGCGAAUUCGCG)2 | 177.4 | 13.3 | 29.9 | 9.9 | 83% | 26% |
| | Protein/DNA complex 1GCC | 126.0 | 23.3 | 39.2 | 19.1 | 69% | 18% |

### 4.3.3 Improving structural stability

We have shown that GB-Neck2 reduces the abs_rmsd and rel_rmsd compared to the original GB-Neck model if PB energies are used as benchmark. Since the original GB-Neck breaks all H-bond in DNA duplex simulation,[36] we will further test if better performance of GB-Neck2nu in energy calculation could result better structural stability in MD simulation. We performed very long MD simulation (1 μs) GB-Neck2nu with different DNA and RNA duplexes as well as a protein/DNA complex. The structural stability from GB-Neck2nu is compared with TIP3P MD (0.05-0.1 μs). There are 6 tested systems. Two of them are used for training GB-Neck2nu (DNA duplex $(CCAACGTTGG)_2$, RNA duplex $(CCAACGUUGG)_2$). Three of them are used for comparing GB and PB energies (DNA duplex $(CGCGAATTCGCG)_2$, RNA duplex $(CGCGAAUUCGCG)_2$ and protein/DNA complex 1GCC). We also tested a longer DNA sequence (DNA duplex 18 basepair) that has a long TIP3P MD trajectory (100 ns) from Pérez et al.[179] This DNA duplex corresponds to "seq2" in their paper.

Table 4.5 shows the average of H-bond fraction (defined in method section) in GB-Neck2nu and TIP3P MD simulation. Within 1 μs simulation time, GB-Neck2nu can maintain 83 to 97% of H-bond for our tested DNA (RNA) duplex and DNA/protein complex system if all base pairs are included in calculation. TIP3P simulation can maintain 95 to 98 % H-bond (100 ns length).

Since we see the defraying of terminal base pair in both TIP3P and GB simulations, we also compared H-bond fraction by excluding one base pair in each terminal. Without the defraying base pairs, TIP3P can maintain almost 100% of H-bond while GB-Neck2nu can achieve 91 to 98% of H-bond. For first 4 DNA and RNA sequences with C-G terminal base pair, GB-Neck2 can achieve 97 to 98 % H-bond. For DNA sequence with terminal A-T base pair (DNA seq2 and DNA/protein complex 1GCC), the H-bond fraction is somewhat lower with 91%, which is 8% smaller than TIP3P MD (99%). If all A-T base pairs that are close to the terminal base pairs are excluded (last 3 base pairs in DNA seq2 and first 2 base pairs in protein/DNA complex), the H-bond fraction is almost 100 % for GB-Neck2nu (table 4.5).

The lower H-bond fraction for DNA (RNA) system with A-T terminal base pair indicates that H-bonds in this base pair are still weak. However we observe the same trend in TIP3P MD with lesser defraying degree. More quantitative benchmark, such as comparing H-bond PMF between GB and TIP3P MD, should be focused on future research. The weak H-bond problem

could be fixed by adjusting the Hydrogen radii by following previous work on both protein and DNA simulations.[4, 23]

We also report the average BB-RMSD over time to experimental structure between GB-Neck2nu and TIP3P MD simulation (table 4.7). For 4 DNA and RNA duplexes with one DNA/protein complex, the average BB-RMSD difference between GB-Neck2nu and TIP3P is only within 1.5 Å except the case of DNA seq 2 (6.7 Å in GB-Neck2nu vs. 4.4 Å TIP3P). For DNA seq 2, we want to stress that GB-Neck2nu MD is 1 μs long while TIP3P MD is only 0.1 μs. The difference in simulation length and the sampling in implicit (much better)[1a] and explicit solvent might be the reason leading to the significant difference in average BB-RMSD. The comparison between representative structure of the most populated cluster in GB-Neck2nu and TIP3P MD simulations are also given in figure 4.S13.

Besides testing the stability of the duplex, we also extend to the quadruplex systems. We chose a small DNA G-quadruplex $(GGGG)_4$ (antiparallel strand (aps))[177a] and a larger system four-stranded Oxytricha telomeric DNA (PDB ID: 1L1H). For GB-Neck2nu MDs, the structures are stable within the simulation time (1 μs) with very low average BB-RMSD (1.6-1.7 Å). The average H-bond fraction is almost 100%. In contrast, two quadruplex systems are not stable in TIP3P MD simulation (200-300 ns). The average BB-RMSD is 4.2 to 4.4 Å for both systems and most of native H-bonds were lost. Instead TIP3P MD simulations tend to favor different H-bond patterns (figure 4.S12). We want to stress that we performed all TIP3P MD simulation without ion. The quadruplex systems are shown to be stable in explicit MD simulation if proper ion is included.[164] For example, with the same DNA G-quadruplex aps system, the average BB-RMSD in TIP3P MD with ion is also very low (~1.2 Å) with almost 100% H-bond.[177a] The source of stability of quadruplex in GB-Neck2nu (without explicit ion) is not known yet. One possible reason could be that GB-Neck2nu introduces the overestimation of base stacking. The overestimation of H-bond is not likely the reason since we just showed that H-bonds in GB-Neck2nu are still weaker in TIP3P simulation for DNA duplex systems.

**Table 4.5.** Average H-bond fraction in GB-Neck2nu and TIP3P simulation for DNA (RNA) duplex and DNA/protein complex. The error is given in parenthesis. The errors are calculated from two runs (starting from A and B forms) for first 4 systems. For DNA seq2 and DNA-protein complex, the errors were calculated from first and second half of the trajectory of single MD run. a) For DNA systems having A-T base pairs in the terminal, we also report H-bond fraction excluding those base pairs.

| System | TIP3P | | GB-Neck2nu | |
|---|---|---|---|---|
| | All base pairs | Skip 2 terminal base pairs | All base pairs | Skip 2 terminal base pairs |
| DNA $(CCAACGTTGG)_2$ | 94±5 | 100±1 | 93±1 | 98±1 |
| DNA $(CGCGAATTCGCG)_2$ | 95±1 | 100±1 | 88±1 | 98±1 |
| RNA $(CCAACGUUGG)_2$ | 96±2 | 98±1 | 97±1 | 98±1 |
| RNA $(CGCGAAUUCGCG)_2$ | 98±1 | 98±1 | 92±1 | 97±1 |
| DNA seq2 (CTAGGTGGATGACTC**ATT**)$_2$ | 97±1 | 99±1 | 83±2 | 91±1; 98±1 [a] |
| DNA-protein complex (PDB ID: 1GCC) DNA sequence: (**TA**GCCGCCAGC)$_2$ | 95±1 | 99±1 | 86±1 | 91±1 ; 99±1 [a] |

### 4.3.4 Testing structural conversion

To further characterize the success of GB-Neck2nu model, we will test if GB-Neck2nu is able to reproduce the structural conversion from A to B form for DNA and B to A form for RNA from explicit solvent simulation. Those are traditional tests when developing new force field[177a, 180b] or testing solvent model.[4] The simulations that started from A form DNA and B form RNA converge to the B-form DNA and A form RNA, respectively.

Table 4.7 shows the convergence of A and B forms of DNA (or RNA): MD runs converge to the same structural ensemble with similar average RMSD. GB-Neck2 excellently reproduces the major and minor width as compared to TIP3P simulation for DNA (table 4.6). The minor groove widths from GB-Neck2nu MD for RNA simulations are also similar to TIP3P runs. The major grooves in GB-Neck2nu MD simulations for all tested RNA are underestimated about 4 Å as compared to TIP3P MD simulation (~15 Å for GB-Neck2nu and ~19 Å for TIP3P).

It is interesting that TIP3P MD simulation with bsc0χOL3 force field overestimates 2.5-3.2 Å of the major groove relative to X-ray and NMR data.[180b] The combination of GB-Neck2nu and bsc0χOL3 perform even better than TIP3P + bsc0χOL3 with similar major groove width (~ 15 Å). This suggests there is error cancellation between bsc0χOL3 force field and GB-Neck2nu solvent model.

**Table 4.6.** Groove width of DNA duplex $(CGCGAATTCGCG)_2$ and RNA duplex $(CGCGAAUUCGCG)_2$ from GB-Neck2nu and TIP3P MD simulations. There are two runs for each solvent model, starting from A and B-forms. Standard deviation for each run is shown in parenthesis.

| | DNA $(CGCGAATTCGCG)_2$ | | RNA $(CGCGAAUUCGCG)_2$ | |
|---|---|---|---|---|
| Groove width (Å) | GB-Neck2nu | TIP3P | GB-Neck2nu | TIP3P |
| Major | 18.7±0.1 | 18.8±0.1 | 15.1±0.5 | 19.2±0.1 |
| Minor | 12.8±0.1 | 12.1±0.1 | 15.9±0.1 | 15.1±0.1 |

**Table 4.7**. Summary of testing structural stability and structural conversion in MD simulations. "Stable" means getting a low average RMSD to either B form DNA or A form RNA in DNA and RNA simulations, respectively. "Average BB RMSD" column uses this format "RMSD to A-form (RMSD to B-form)" for DNA or RNA duplex simulations. "A → B" or "B → A" shows the conversion of A to B form in DNA simulation (starting from A-form) or B to A for in RNA simulation (starting from B-form), respectively. The RMSD plots for each system were given in supplement.

| | Length (ns) | | Average BB RMSD (Å) | | Observation |
|---|---|---|---|---|---|
| System | GB-Neck2nu | TIP3P | GB-Neck2nu | TIP3P | |
| A-form DNA $(CCAACGTTGG)_2$ | 1000 | 100 | 4.3 (4.3) | 4.2 (3.1) | A → B, stable |
| B-form DNA $(CCAACGTTGG)_2$ | 1000 | 100 | 4.2 (4.3) | 4.0 (3.1) | Stable |
| A-form DNA $(CGCGAATTCGCG)_2$ | 1000 | 100 | 5.2 (4.2) | 5.3 (3.0) | A → B, stable |
| B-form DNA $(CGCGAATTCGCG)_2$ | 1000 | 100 | 5.2 (4.2) | 5.4 (2.9) | Stable |
| A-form RNA $(CCAACGUUGG)_2$ | 1000 | 100 | 2.1 (6.1) | 2.8 (5.6) | Stable |
| B-form RNA $(CCAACGUUGG)_2$ | 1000 | 100 | 2.2 (6.4) | 2.8 (5.5) | B → A, stable |
| A-form RNA $(CGCGAAUUCGCG)_2$ | 1000 | 100 | 2.3 (6.7) | 3.6 (6.3) | Stable |
| B-form RNA $(CGCGAAUUCGCG)_2$ | 1000 | 100 | 2.7 (6.7) | 3.7 (5.8) | B → A, stable |
| B-form DNA seq2 $(CTAGGTGGATGACTCATT)_2$ | 1000 | 100 (Perez et al.[179]) | 6.7 | 4.4 | Stable |
| DNA quadruplex $(GGGG)_4$ | 1000 | 200 | 1.6 | 4.4 | Stable |
| DNA quadruplex $(GGGGTTTTGGGG)_2$ (PDB ID: 1L1H) | 1000 | 300 | 1.7 | 4.4 | Stable |
| DNA-protein complex (PDB ID: 1GCC) | 50 | 50 | 2.7 | 2.4 | Stable |

### 4.3.5 Folding DNA and RNA hairpin

We have shown that GB-Neck2nu is able to maintain stable DNA and RNA duplexes and is able to reproduce the structural conversion from A to B form for DNA and B to A form for RNA. The conversions in DNA and RNA duplexes are just small arrangement, so we further test larger conversion, specifically testing the folding DNA GCA hairpin loop and RNA UUCG hairpin loop. We also compare GB-Neck2nu MD/REMD simulations to GB-HCT simulations since this model has been used widely for DNA and RNA simulations.[36, 169a, 175]

For the DNA hairpin GCA loop (PDB of its homologue: 1ZHU), all GB MD simulations started from both NMR and B form conformations. In GB-HCT simulation starting from NMR structure, the native conformation was maintained for ~1 μs but it adopts different compacted structure (Figure 4.2) for another 6 μs (BB-RMSD of 3.0 Å). The B-form GB-HCT simulations also converged to the same misfolded structure. No refolding event is observed in GB-HCT simulations. Multiple folding/unfolding events appear in GB-Neck2nu MD simulations (Figure 4.2). The folded structure from this simulation has remarkably low BB-RMSD to NMR structure (1.2 Å). Those results are consistent with previous observations in protein folding simulation, in which GB-HCT favors more compacted structures.[31, 54]

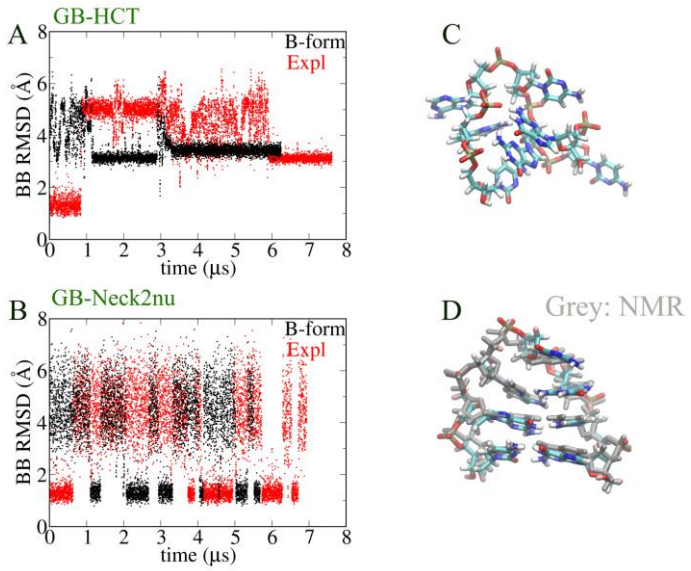


**Figure 4.2**. DNA GCA hairpin loop MD simulation, starting from B-form and NMR structures. (A) Backbone RMSD evolution for GB-HCT simulation, starting from native structure (red) and B-form conformation (black). (B) Backbone RMSD evolution for GB-Neck2nu simulation,

starting from B-form conformation. (C) Misfolded structure from GB-HCT simulation. (D) Overlap between experimental structure (grey color) and representative structure of the most populated cluster from GB-Neck2nu simulation. Only the simulation that started from B-form of GB-Neck2nu MD is shown since we observed multiple folding/unfolding events in this run.

We next test the folding of the RNA hairpin UUCG loop (PDB ID: 2KOC[183]). Since this hairpin has 5 base pairs in the stem, the structure is very stable in MD simulation (our preliminary result). To accelerate the folding/unfolding, we performed REMD run for each GB model (GB-HCT and GB-Neck2nu), starting from both NMR and A-form conformation. GB-HCT again favors different compacted structure (BB-RMSD to NMR structure: 8.6 Å) while GB-Neck2nu is able to fold to a structure similar to NMR structure with BB-RMSD of 1.9 Å and stem BB-RMSD of 1.1 Å. The loop geometry of the hairpin in GB-Neck2 simulation however is not correctly folded (figure 4.3).

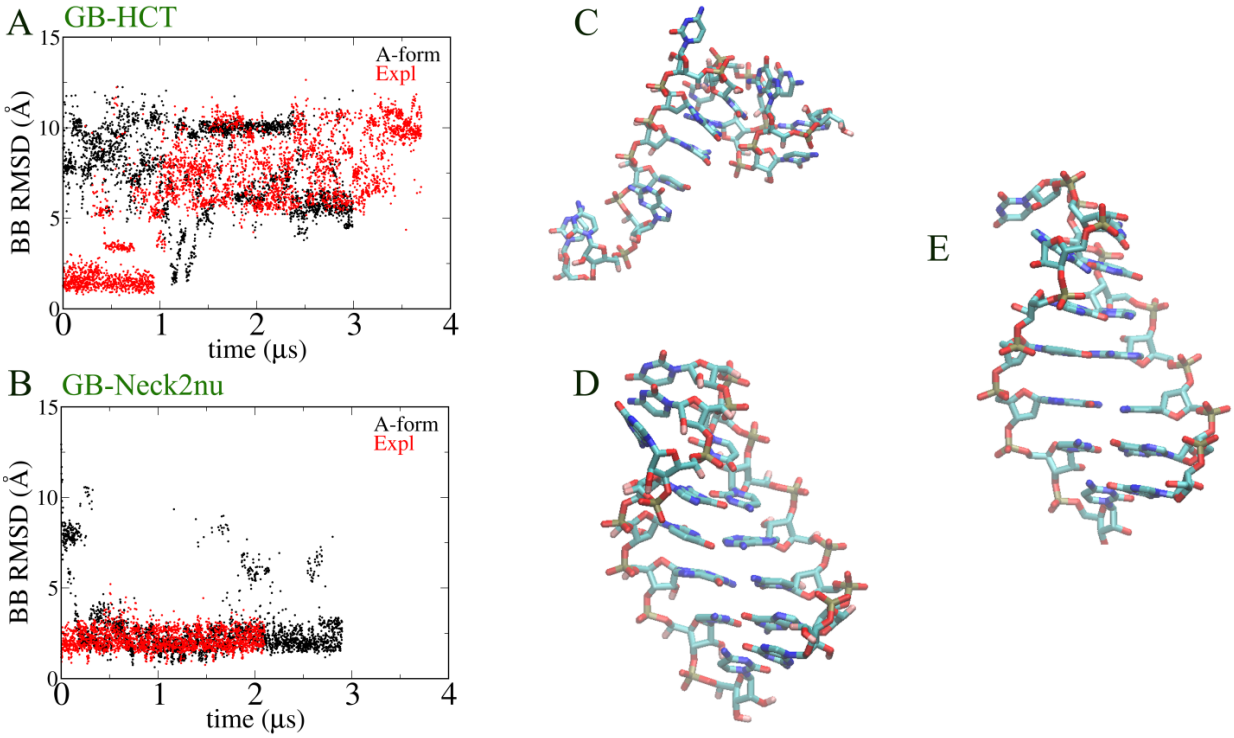


**Figure 4.3**. RNA UUCG hairpin loop REMD simulation starting from A-form and NMR structures. Only 300 K trajectories are shown. (A) Backbone RMSD versus time from GB-HCT simulation. (B) Backbone RMSD versus time from GB-Neck2nu simulation. (C) Misfolded structure from GB-HCT simulation in REMD run starting from NMR structure. (D) Folded

structure from GB-Neck2nu MD simulation. (E) Experimental structure (PDB ID: 2KOC[183]). For clarity, only heavy atoms are shown.

### 4.3.6 Reproducing ligand binding to DNA duplex

We have shown GB-Neck2nu can qualitatively reproduce the minor and major groove width of DNA duplex compared to TIP3P simulation. We then tested if this agreement could help GB-Neck2nu reproduce the binding of the ligand to the minor groove of DNA since the binding of ligand to its binding site is sensitive to minor groove's width.[192] We choose the complex of Dickerson-Drew dodecamer DNA duplex $(CGCGAATTCGCG)_2$ and its ligand DAPI for testing (PDB ID: 1D30). We performed 10 MD simulation runs (1.5-3 μs per run) in which the ligand was taken out of the binding site from X-ray structure. The initial ligand position is arbitrarily chosen (with O2 (DT7) - HN1 (DAPI) distance of 36.0 Å vs. 2.2 Å in X-ray structure). Ten different MD runs have different random starting velocities. In all the simulations, the ligand can bind to the minor groove within only 100 ns. Seven of 10 runs can sample correct binding site (justified by correctly reproduce the native hydrogen bond in crystal structure between HN1 (DAPI) and O2 (DT7) within our simulation time. Figure 4.4 shows one of examples of the binding process, illustrated by the BB-RMSD to the complex, BB-RMSD to only DNA and by the distance between HN1 (DAPI) and O2 (DT7). The representative structure of the closest cluster has HN1 (DAPI) and O2 (DT7) distance of 2.7 Å which the close to the distance in X-ray structure (~2.2 Å). The most populate cluster does not have native H-bond between HN1 (DAPI) and O2 (DT7), however the ligand still stays inside the minor groove (figure 4.4). The ligand binding process for all 10 MD runs are given in figure 4.S14.

There are several reasons that made the most populated cluster not to have correct binding site for ligand. First we used very-approximated GAFF parameters with AM1-BCC charge model for quickly testing our prediction. More careful force field development for DAPI ligand should investigated if one wants to get more accurate result. Secondly, the force field bsc0 for DNA was shown to produce wider minor groove in MD simulation than in experiment.[36] The wider minor groove width might weaken the tight binding in experimental structure.

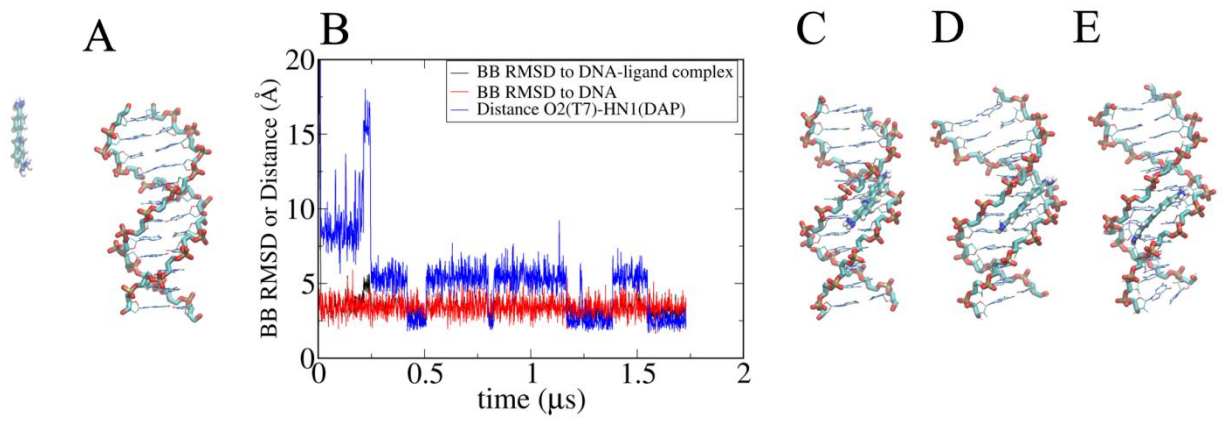


**Figure 4.4**. (A) Starting structure for GB-Neck2nu MD simulations of ligand binding. The ligand was initially taken out of its binding site in DNA minor groove. The initial distance of O2 (DT7) - HN1 (DAPI) is 36 Å. (**B**) Backbone RMSD and O2 (DT7) - HN1 (DAPI) distance versus time from GB-Neck2nu simulation. The overlapping of BB-RMSD to the complex and to only DNA in the plot indicates the ligand stays inside the minor groove after finding its binding site. (**C**) Crystal structure (PDB ID: 1D30[185]) of the complex between DNA duplex $(CGCGAATTCGCG)_2$ with ligand DAPI. The O2 (DT7) - HN1 (DAPI) distance is 2.2 Å. (**D**) The representative of the closest cluster to crystal structure from GB-Neck2nu simulation. The O2 (DT7) - HN1 (DAPI) distance is 2.7 Å. (**E**) Representative of the most populate cluster from GB-Neck2nu simulation. The O2 (DT7) - HN1 (DAPI) distance is 5.3 Å.

### 4.4 Conclusion

In this study, we have extended and refitted the GB-Neck model for the MD simulations of nucleic acid and its complex with protein. The fitting reduces 70%-80% error for absolute energy and 15% to 65% for relative energy calculation from GB-Neck if using PB calculation as benchmark. The effective radii calculation is also modestly improved. The improvement in energy and effective radii calculation translate for better structural stability for duplex, quadruplex and duplex/protein complex simulations. The model is also able to fold DNA and RNA hairpin loop and correctly reproduce the ligand binding to its binding site in minor groove of DNA.

We also show that the A-T base pair H-bonds are still weak in GB-Neck2nu simulation. Future research will focus on its stability, such as by adjusting atomic radii as done previously.[169b]

## Appendix 4. Supporting Document

**Table 4.S1**. Parameters for first 10 of 600 runs that have the lowest objective function values. "Sx < 0" means one of the scaling factors is negative while "Sx > 0" means all the scaling factors are positive. Three last rows show RMSD between GB and PB energies for dnadup and rnadup and the effective radii RMSD for dnadupRad training set. Relative energy RMSD is shown in parenthesis. Run #3 and #4 have similar objective functions but they have completely different parameters. We chose parameter set #4 as our final set run since it has slightly lower relative energies from both training sets and it has all positive scaling factor values (as compared to #1 and #2). Only absolute energy, relative energy and effective radii RMSD are shown for two best runs (#3 and #4).

| | #1 | #2 | #3 | **#4** | #5 | #6 | #7 | #8 | #9 | #10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Pars | Sx < 0 | Sx < 0 | Sx > 0 | **Sx > 0** | Sx < 0 | Sx > 0 | Sx > 0 | Sx < 0 | Sx > 0 | Sx > 0 |
| $S_H$ | -0.556 | -0.546 | 1.175 | **1.697** | 0.536 | 1.225 | 1.127 | 1.201 | 1.211 | 1.229 |
| $S_C$ | 0.920 | 0.887 | 0.669 | **1.269** | 0.817 | 0.644 | 0.611 | 0.178 | 0.175 | 0.095 |
| $S_N$ | 1.118 | 1.083 | 1.066 | **1.426** | 1.065 | 1.075 | 0.997 | 1.011 | 1.016 | 1.032 |
| $S_O$ | -0.309 | 0.221 | 0.184 | **0.184** | -0.333 | 0.405 | 0.183 | -0.019 | 0.184 | 0.184 |
| *Sp* | 1.500 | 1.451 | 1.487 | **1.545** | 1.432 | 1.491 | 1.418 | 1.434 | 1.448 | 1.476 |
| $\alpha_H$ | 1.373 | 1.359 | 1.184 | **0.537** | 1.241 | 1.193 | 1.421 | 1.368 | 1.180 | 1.315 |
| $\beta_H$ | 2.114 | 2.146 | 1.592 | **0.363** | 1.985 | 1.575 | 2.184 | 2.011 | 1.503 | 1.853 |
| $\gamma_H$ | 1.338 | 1.453 | 1.067 | **0.117** | 1.494 | 1.022 | 1.543 | 1.527 | 1.189 | 1.375 |
| $\alpha_C$ | 0.750 | 1.165 | -0.204 | **0.332** | -0.402 | 0.789 | 0.036 | 0.452 | 1.198 | 1.794 |
| $\beta_C$ | -0.384 | 0.578 | -1.198 | **0.197** | -2.769 | 1.024 | -0.826 | 0.039 | 1.901 | 3.236 |
| $\gamma_C$ | -0.337 | 0.223 | -0.233 | **0.093** | -1.473 | 0.934 | 0.039 | 0.572 | 1.673 | 2.357 |
| $\alpha_N$ | 2.361 | 2.773 | 1.503 | **0.686** | 0.364 | 2.104 | 1.944 | 2.096 | 2.565 | 2.905 |
| $\beta_N$ | 2.648 | 3.843 | 1.953 | **0.463** | -1.239 | 3.314 | 3.055 | 3.456 | 4.612 | 5.418 |
| $\gamma_N$ | 1.013 | 1.772 | 1.208 | **0.139** | -0.837 | 1.952 | 2.000 | 2.373 | 3.071 | 3.511 |
| $\alpha_O$ | 1.277 | 1.234 | 1.137 | **0.606** | 1.898 | 0.947 | 1.063 | 1.294 | 1.081 | 1.005 |
| $\beta_O$ | 2.470 | 2.459 | 1.937 | **0.463** | 4.436 | 1.381 | 1.794 | 2.419 | 1.814 | 1.587 |
| $\gamma_O$ | 1.918 | 2.075 | 1.396 | **0.142** | 3.514 | 0.990 | 1.464 | 1.928 | 1.477 | 1.277 |
| $\alpha_P$ | 1.222 | 0.812 | 1.077 | **0.418** | 0.612 | 1.104 | 1.109 | 1.234 | 1.143 | 1.102 |
| $\beta_P$ | 3.150 | 2.028 | 2.321 | **0.290** | 0.963 | 1.782 | 2.164 | 1.911 | 1.776 | 1.525 |
| $\gamma_P$ | 4.847 | 4.199 | 3.061 | **0.106** | 3.062 | 2.044 | 3.197 | 2.524 | 2.570 | 2.202 |
| obj_funct | 0.324 | 0.327 | 0.336 | **0.338** | 0.339 | 0.340 | 0.340 | 0.349 | 0.351 | 0.353 |
| dnadup | 11.6 (12.7) | 11.6 (13.0) | 14.8 (11.4) | **14.0 (10.3)** | 13.6 (11.8) | 15.6 (11.8) | 15.8 (11.7) | 15.5 (11.8) | 15.8 (12.2) | 16.1 (12.5) |
| rnadup | 9.3 (9.8) | 9.5 (10.4) | 24.0 (12.4) | **25.4 (11.7)** | 14.5 (10.4) | 25.1 (12.8) | 24.6 (12.7) | 25.4 (12.5) | 24.1 (12.5) | 23.8 (12.7) |
| dnadupRad | 0.046 | 0.044 | 0.035 | **0.041** | 0.048 | 0.033 | 0.037 | 0.037 | 0.038 | 0.037 |

**Table 4.S2.** Summary of training and test set for GB-Neck2nu

| | Name | #structures |
|---|---|---|
| Training set | dnadup $(CCAACGTTGG)_2$ | 370 |
| | rnadup $(CCAACGUUGG)_2$ | 187 |
| Type I test set | dnadup_plus150 $(CCAACGTTGG)_2$ | 520 |
| | rnadup_plus200 $(CCAACGUUGG)_2$ | 387 |
| Type II test set | DNA duplex $(CGCGAATTCGCG)_2$ | 650 |
| | RNA duplex $(CGCGAAUUCGCG)_2$ | 600 |
| | DNA/protein complex 1GCC | 850 |

**Table 4.S3.** Summary of structures used in this study. "*GB vs. PB*" means the structure was used for comparing GB and PB calculation. "*MD simulation*" means the structure was used for MD simulation. "x" mark indicates the structure was used. Blank indicates there is no test.

| System | System size (Number of residue) | Total Charge (C) | GB vs. PB | MD simulation | |
|---|---|---|---|---|---|
| | | | | GB | TIP3P |
| DNA duplex $(CCAACGTTGG)_2$ | 20 | -18 | x | x | x |
| DNA duplex $(CGCGAATTCGCG)_2$ | 24 | -22 | x | x | x |
| RNA duplex $(CCAACGUUGG)_2$ | 20 | -18 | x | x | x |
| RNA duplex $(CGCGAAUUCGCG)_2$ | 24 | -22 | x | x | x |
| DNA duplex $(CTAGGTGGATGACTCATT)_2$ | 36 | -24 | | x | x |
| DNA G-quadruplex $(GGGG)_4$ | 16 | -12 | | x | x |
| DNA G-quadruplex 2 (PDB ID: 1L1H) | 24 | -22 | x | x | x |
| DNA-protein complex (PDB ID: 1GCC) | 85 | -13 | x | x | x |
| DNA GCA hairpin loop (PDB: 1ZHU) | 7 | -6 | | x | |
| RNA UUCG hairpin loop (PBD: 2KOC) | 14 | -13 | | x | |
| DNA and ligand complex (PDB: 1D30) | 25 | -22 | | x | |

**Table 4.S4**. Comparison of energy and effective radii RMSD between GB and PB for training sets with different runs using different weighting factors (wr = 1.5, 2.5, 5.0; wrel = 5.0, 10.0). The default wabs is 1.0. We performed 300-600 function minimization runs for each choice. The fitting parameters from (wr = 2.5, wrel = 5.0, wabs = 1.0) are chosen as the final parameters since they have the best compromise between low energy RMSD and low effective radii RMSD to PB calculation. Third and fourth rows show absolute and relative energy RMSD between GB and PB. Fifth row shows effective radii RMSD to 'perfect' radii.

| Training set | | GB-Neck2nu | | |
|---|---|---|---|---|
| | $w_r$=1.5 $w_{rel}$=5.0 | $w_r$=2.5 $w_{rel}$=5.0 | $w_r$=5.0 $w_{rel}$=5.0 | $w_r$=2.5 $w_{rel}$=10.0 |
| dnadup | 18.7 (10.2) | 14.0 (10.3) | 17.1 (12.5) | 14.7 (12.0) |
| rnadup | 26.4 (11.3) | 25.4 (11.7) | 29.6 (12.9) | 31.1 (11.2) |
| dnadupRad | 0.051 | 0.041 | 0.030 | 0.050 |

**Table 4.S5**. Groove width of DNA duplex $(CCAACGTTGG)_2$ and RNA duplex $(CCAACGUUGG)_2$ from GB-Neck2nu and TIP3P MD simulations. There are two runs for each solvent model, starting from A and B-forms. Standard deviation for each run is shown in parenthesis. Those DNA and RNA duplexes were used for training GB-Neck2nu parameters.

| Groove width (Å) | DNA $(CCAACGTTGG)_2$ | | | | RNA $(CCAACGUUGG)_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | GB-Neck2nu | | TIP3P | | GB-Neck2nu | | TIP3P | |
| | A-form | B-form | A-form | B-form | A-form | B-form | A-form | B-form |
| Major | 18.6 (2.0) | 18.6 (2.0) | 18.0 (1.9) | 18.2 (1.9) | 15.2 (3.0) | 15.3 (2.9) | 19.0 (2.1) | 19.0 (2.1) |
| Minor | 13.0 (1.3) | 13.1 (1.3) | 12.4 (1.7) | 12.4 (1.6) | 15.9 (0.7) | 15.9 (0.7) | 15.4 (0.7) | 15.4 (0.7) |

**Table 4.S6**. Major and minor groove widths of DNA duplex $(CTAGGTGGATGACTCATT)_2$ from GB-Neck2nu and TIP3P MD simulations. Both simulations started from B-form. Standard deviation for each run is shown in parenthesis. 100 ns TIP3P MD trajectory was taken from Pérez et al.[179]

| Groove width (Å) | GB-Neck2nu | TIP3P |
|---|---|---|
| Major | 19.2 (2.0) | 18.0 (1.9) |
| Minor | 13.1 (1.3) | 12.8 (1.8) |

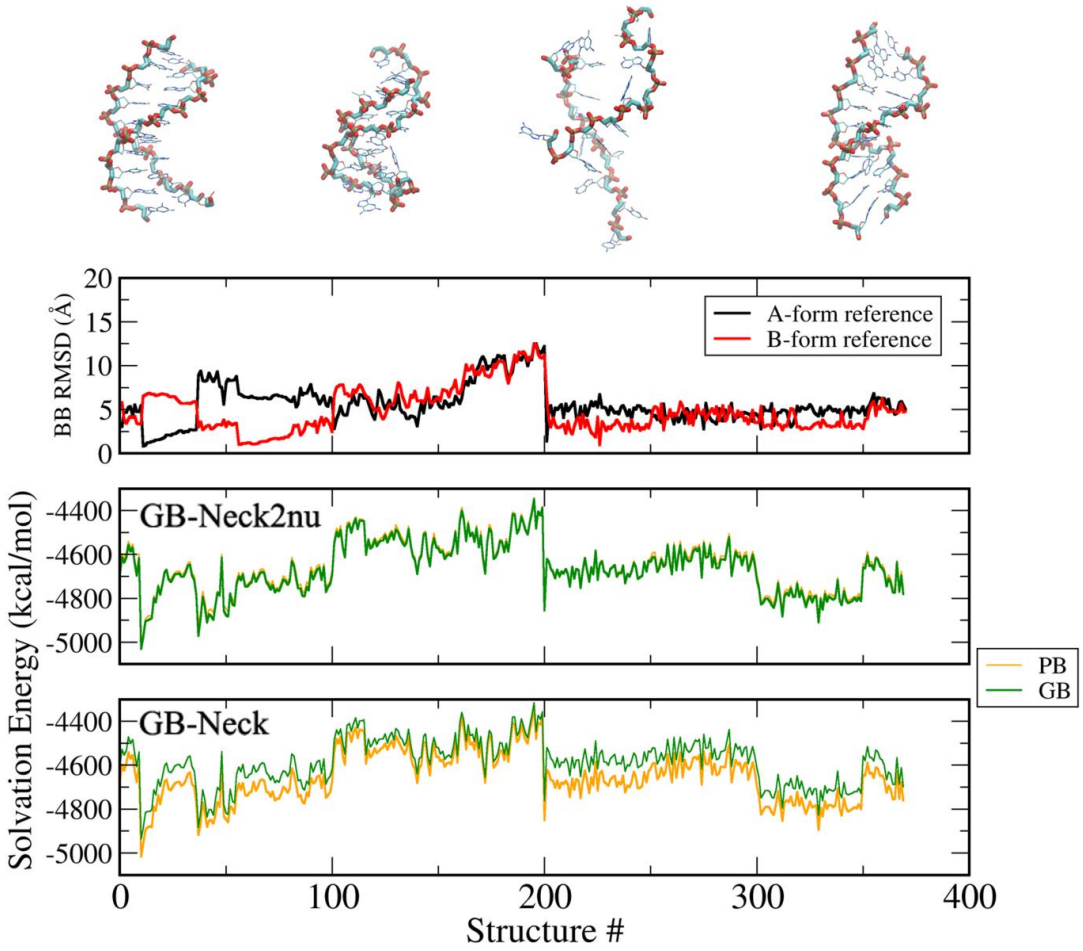


**Figure 4.S1.** Comparison between GB and PB energies for individual structures in the DNA training set for GB-Neck2nu and GB-Neck. Top panel shows the structures for 10th, 55th, 200th, 370th frames as an example of structural diversity in training. Second panel shows the backbone RMSD of each structure to canonical A and B-forms of DNA.

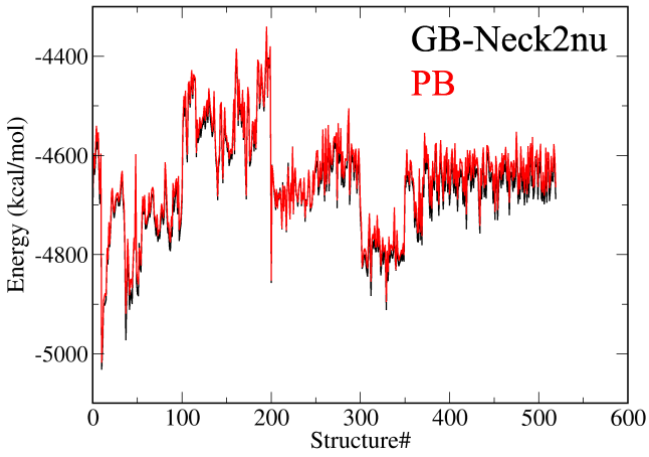


**Figure 4.S2**. Comparison between GB and PB energies for individual structure in DNA training set in the 5th round (first 370 structures) and structures taken from 0.75 microsecond MD simulation of DNA duplex using GB parameters from the 5th round (last 150 structures). We stopped the function minimization after the 5th round since there is no strong energy bias for the new structures. Including the last 150 structures in training set did not reduce the error between GB and PB energies (data not shown). This indicates that our training set converged after 5th rounds.

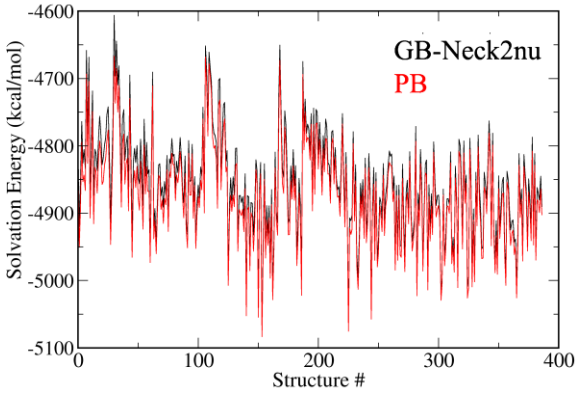


**Figure 4.S3**. Comparison between GB and PB energies for individual structure in RNA training set in the 5th round (first 187 structures) and structures taken from 1.0 microsecond MD simulation of RNA duplex using GB parameters from the 5th round (last 200 structures). We stopped the function minimization after the 5th round since there is no strong energy bias for the new structures. Including the last 200 structures in the training set did not reduce the error between GB and PB energies.

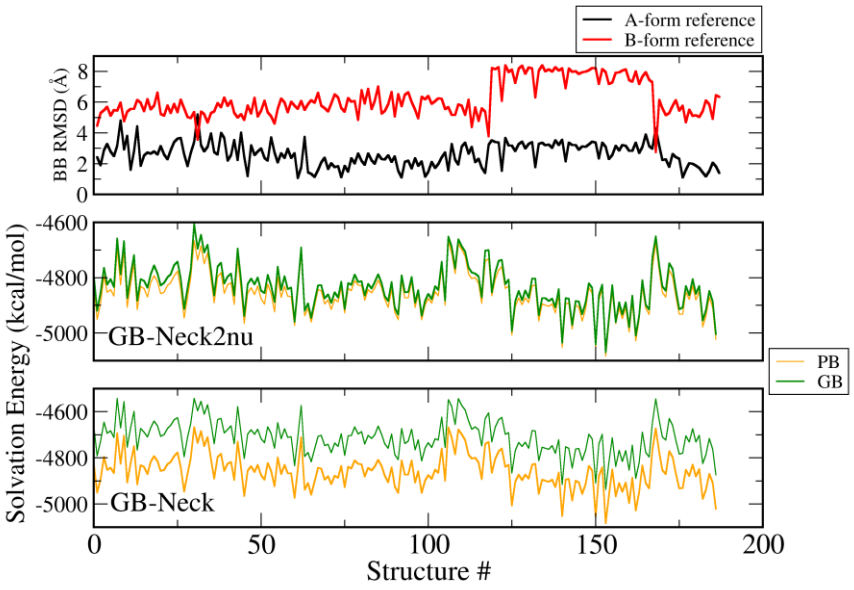


**Figure 4.S4**. Comparison between GB and PB energies for individual structures in RNA $(CCAACGUUGG)_2$ training set for GB-Neck2nu and GB-Neck. Top panel shows the backbone RMSD of each structure to canonical A and B-form RNA.

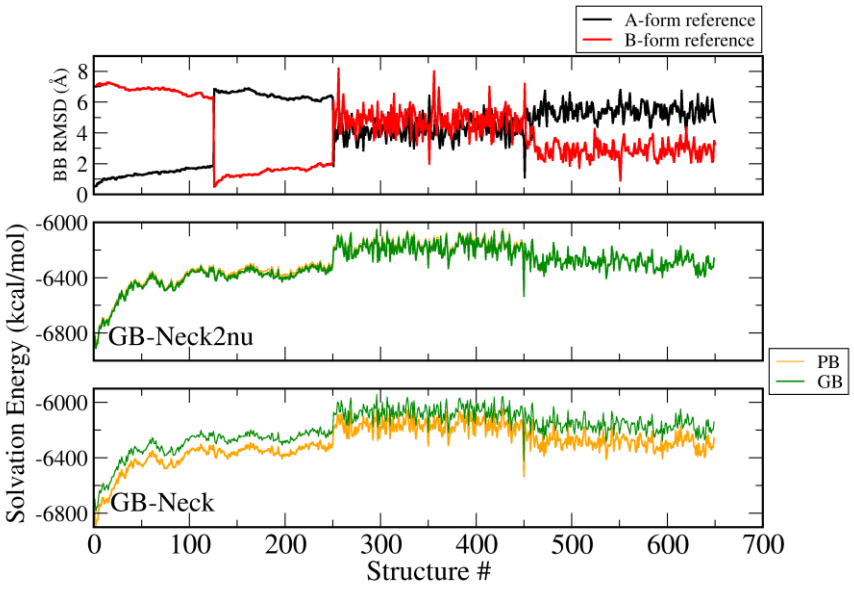


**Figure 4.S5**. Comparison between GB and PB energies for individual structure in DNA $(CGCGAATTCGCG)_2$ test set for GB-Neck2nu and GB-Neck. First 450 structures were from GB-intermediate MD simulations and last 200 structures were from TIP3P MD simulation. Top panel shows the backbone RMSD of each structure to canonical A and B-form DNA.

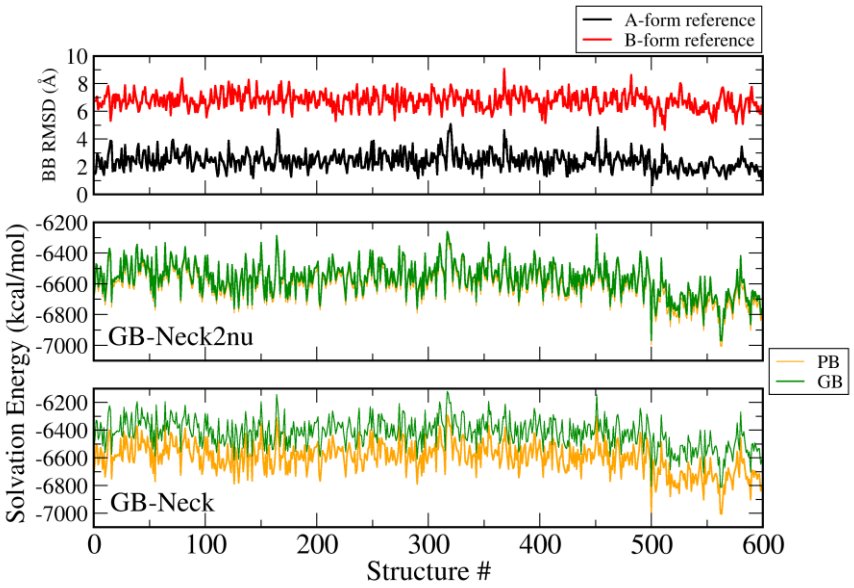


**Figure 4.S6**. Comparison between GB and PB energies for individual structure in RNA $(CGCGAAUUCGCG)_2$ test set for GB-Neck2nu and GB-Neck. First 500 structures were from GB-intermediate MD simulation and last 100 structures were from TIP3P MD simulations. Top panel shows the backbone RMSD of each structure to canonical A and B-form RNA.

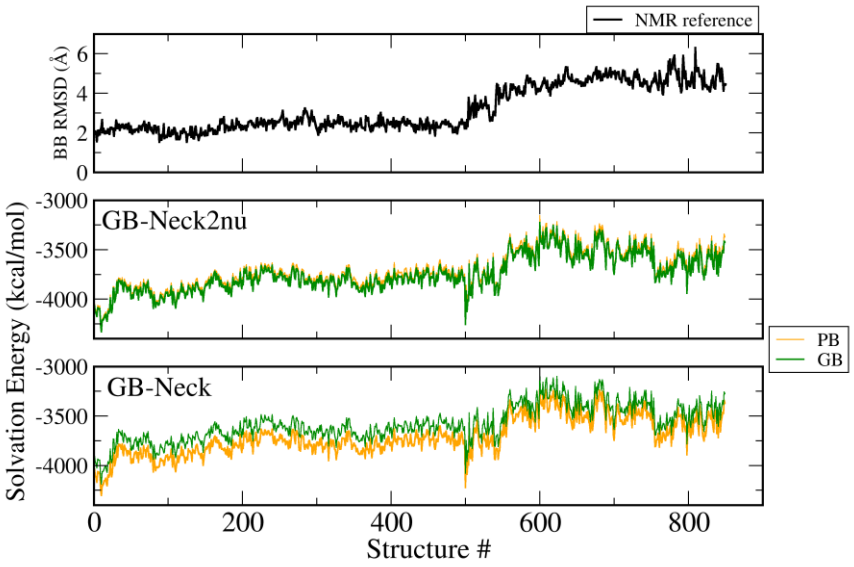


**Figure 4.S7**. Comparison between GB and PB energies for individual structure in DNA/protein complex 1GCC test set for GB-Neck2nu and GB-Neck. First 500 and last 350 structures were from TIP3P MD simulations at 300 and 500 K, respectively. Top panel shows the backbone

RMSD to NMR structure (PDB ID: 1GCC). Flexible termini were skipped for RMSD calculation (residue 23$^{th}$ to 26$^{th}$ and residue 75$^{th}$ to 85$^{th}$ in the complex).

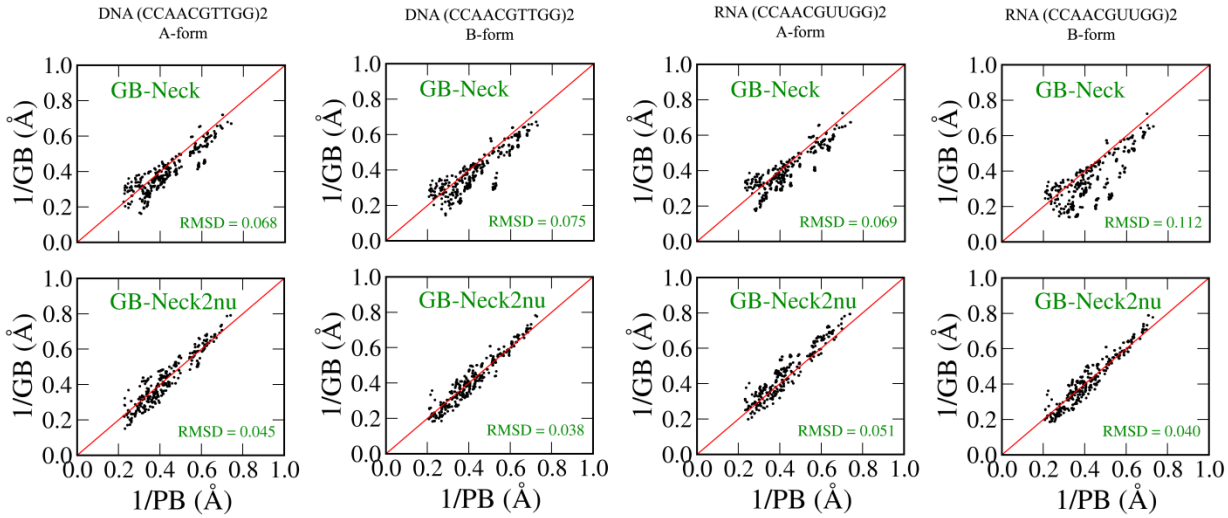


**Figure 4.S8.** Comparison of the inverse of effective radii between GB-Neck (top), GB-Neck2nu (bottom) and the inverse of PB "perfect" radii for A and B form DNA duplex $(CCAACGTTGG)_2$, A and B form RNA duplex $(CCAACGUUGG)_2$. A and B-form DNA duplex $(CCAACGTTGG)_2$ were used for radii training of GB-Neck2nu.

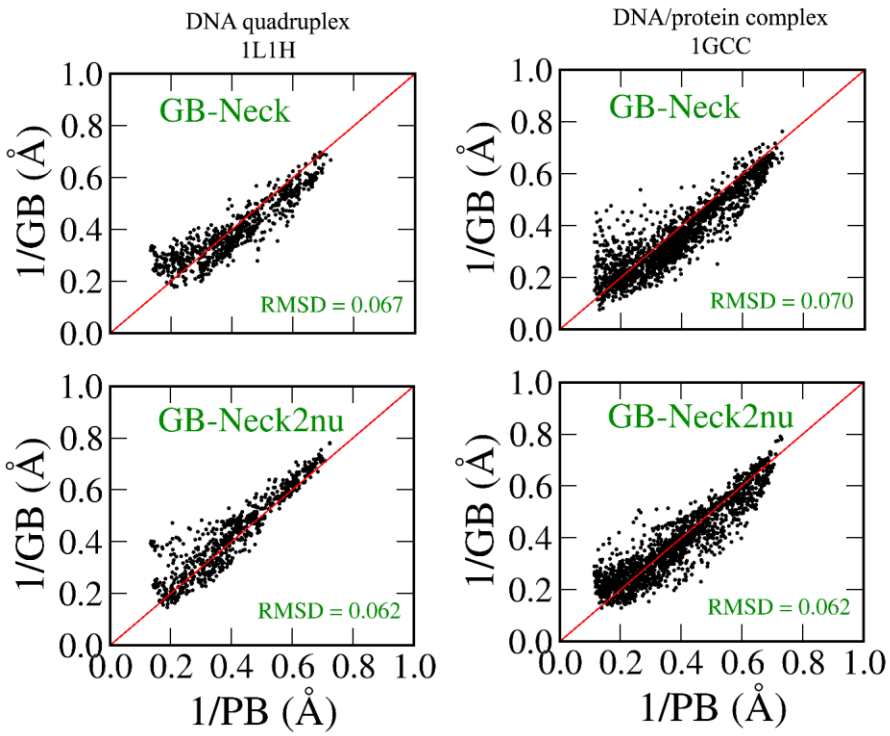


**Figure 4.S9**. Comparison of the inverse of effective radii between GB-Neck (top), GB-Neck2nu (bottom) and the inverse of PB "perfect" radii for DNA quadruplex (PDB ID 1L1H)[193] and DNA/protein complex (PDB ID 1GCC).[178]

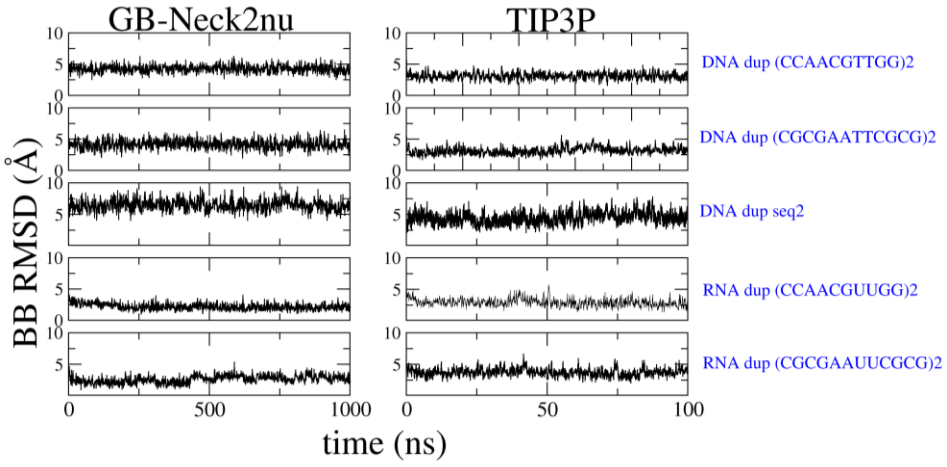


**Figure 4.S10.** Backbone RMSD evolution of DNA and RNA duplexes for GB-Neck2nu (left) and TIP3P (right) MD simulations. Stable structure in experiment was used as reference for RMSD calculation. For DNA duplexes, MD simulations started from A-form. For RNA duplexes, MD simulations started from B-form. Experimental structures were used as reference

structure for RMSD calculation. GB-Neck2nu MD simulations are 10-fold longer than TIP3P MD ones (1000 and 100 ns for GB-Neck2 and TIP3P, respectively). "DNA dup seq2" corresponds to DNA duplex (CTAGGTGGATGACTCATT)2 and the TIP3P trajectory was taken from Pérez et al.[179]

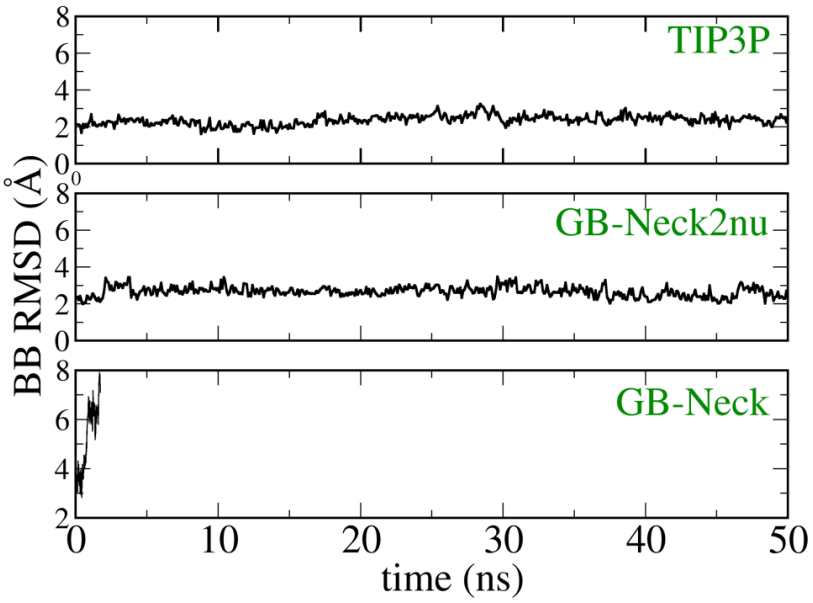


**Figure 4.S11**. Backbone RMSD evolution of Protein/DNA complex 1GCC.

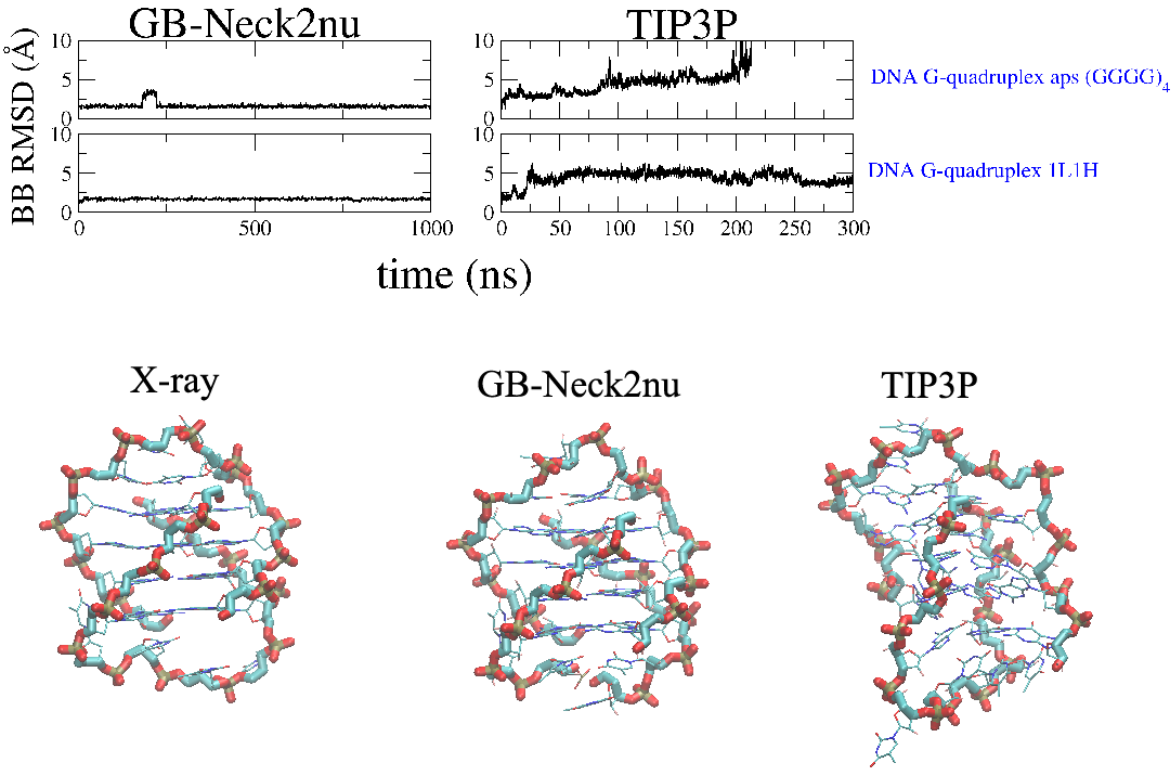


**Figure 4.S12**. (Top) Backbone RMSD of two DNA quadruplexes for GB-Neck2nu (left) and TIP3P (right) MD simulations. (Bottom) X-ray structure of DNA quadruplex 1L1H (PDB: 1L1H) with the representative structure of the most populated cluster from GB-Neck2nu (1000 ns) and TIP3P (300 ns without ion) MD simulations. Without salt, TIP3P structure different compacted structure; this is consistent with previous study.[164]

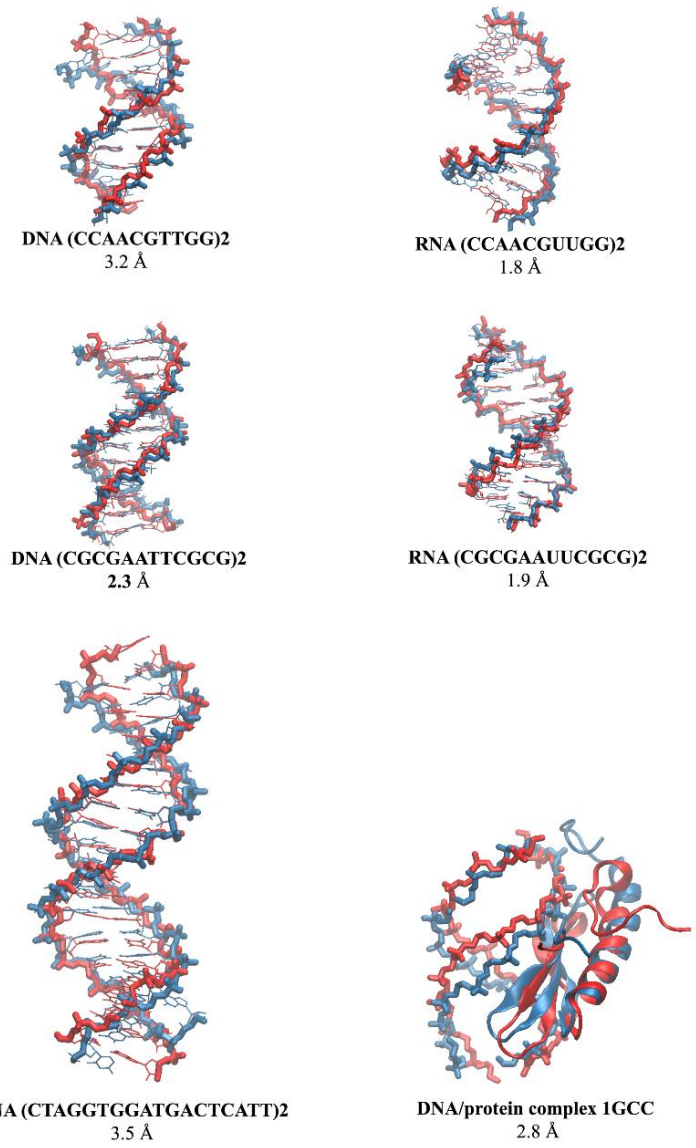


**Figure 4.S13**. Structural overlapping and BB-RMSD between representative structures of the most populated clusters from GB-Neck2nu (blue) and TIP3P (red) MD simulations. Only backbones of DNA are shown in DNA/protein 1GCC complex for clarity.

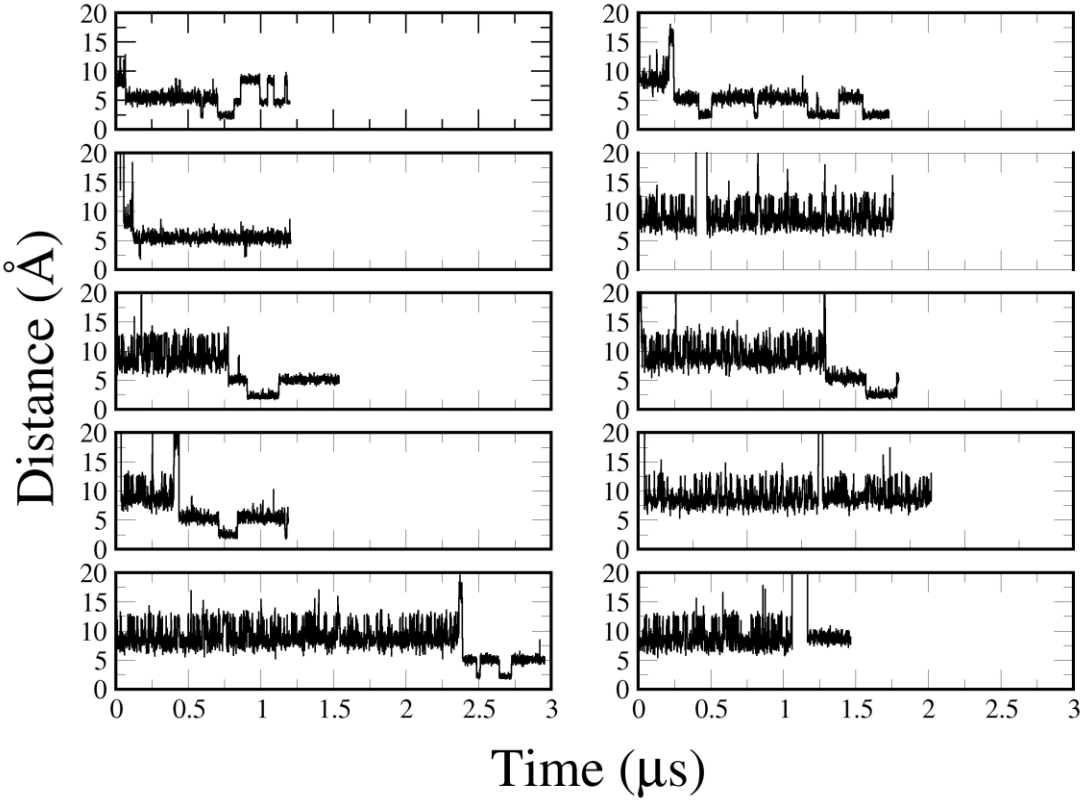


**Figure 4.S14**. Distance between O2 (DT7) and HN1 (DAPI) versus time in ten of GB-Neck2nu MD simulations of DAPI ligand binding to minor groove of DNA duplex $(CGCGAATTCGCG)_2$. The distance of O2 (DT7) and HN1 (DAPI) in X-ray structure is ~2.2 Å.

## Chapter 5. Conclusion and Future Direction

Molecular Dynamics simulation has been powerful to understand biological process at atomic detail. However the interesting process, such as protein folding, happens in large time scale, from microsecond to second.[37a] With the current computer power, most of research groups are only able to reach to microsecond timescale.[37a, 172] Alternatively, implicit solvents (especially fast pairwise Generalized-Born solvent model) dramatically accelerate the sampling thanks to their low viscosity[1a] and their friendliness to parallel computation in CPU and GPU.[4, 48]

The speed comes with trade-off accuracy. The most accurate implicit solvent model such as Poisson–Boltzmann method[77] or GBMV2 model[20] is just too slow to be routine in long time scale MD simulation.[14] Much faster and more popular model to be used in MD simulations are GB-HCT[21] and GB-OBC[23] models but they favor more alpha helix[31, 33-34, 49, 55] and have much stronger salt bridge strength[30b, 34b] than explicit solvent simulation.

GB-Neck model[24] introduces more correction to GB-OBC with the hope that it is as fast as GB-OBC and its accuracy is comparable to more accurate (but slow) GBMV2 model. It turns out that this model tends to disfavor native structures for both protein and nucleic acid simulation.[24, 36, 49] Mongan et al.[24] showed the H-bond in beta hairpin was too weak in GB-Neck2 MD simulation. This group and others also showed DNA duplex strands quickly dissociates in this model.[24, 36]

Since the performance of a GB model is heavily affected by its set of empirical parameters, we hypothesize that GB-Neck has more correct theory level as compared to its ancestors (GB-HCT, GB-OBC) but its parameters need to be re-optimized with more careful design of training and testing data.

Our main goal is to develop a set of parameter for GB-Neck to reduce the limitation of the current fast pairwise GB models. Specifically, we want to reduce the helical bias, reproduce the salt bridge strength and have more stable DNA/RNA duplex simulation as compared to explicit MD data.

We have presented the further development of CFA-based GB-Neck model in chapter 2 and 4,[24] resulting two parameter sets: a set for protein simulation (GB-Neck2) and a set for nucleic acid simulation (GB-Neck2nu). GB-Neck2 is shown to have better secondary structure balance and salt bridge strength as compared to GB-Neck or GB-OBC model for protein simulation.[29] GB-Neck2 even goes further by showing that with the combination of a good GB model, a good force field and GPU, simulation of μs to millisecond protein folding is now routine for common hardware.[167] GB-Neck2nu set is not only able to maintain stable nuclei acid duplex simulation but also able to reproduce the folding of DNA, RNA hairpin or reproduce the ligand binding process. Unlike other GB models that only deal with either protein or nucleic acid, two sets GB-Neck2 and GB-Neck2nu can be combined for protein/nucleic acid complex simulation.

Our GB model is based on CFA approach (R4 model), which shows overestimate effective radii.[17] However, the rigorous parameter fitting introduce error cancelation and this fortuitous cancelation is transferrable from one to other systems. Our strategy in designing training set and test set, designing objective function will provide framework for developing later models.

The success of our GB model for nucleic acid simulation, at least for the case of DNA/RNA duplex or hairpin, shows two important things. First, there is still room for developing solvent model for nucleic acid model although this system is highly charged if proper parameterization is done. Secondly, PB calculation is still good benchmark for GB parameter fitting.

Although we have shown the robustness of GB-Neck2 and GB-Neck2nu for protein and nucleic acid simulations, there are still some limitations for those models:

- The solvation energy errors (as compared to PB method) are still large (from 10 to 40 kcal/mol for large systems such as lysozyme or protein/DNA complex 1GCC).

- The alpha helical is reduced (compared to GB-HCT and GB-OBC) but GB-Neck2 still overestimates the 3-10 helix (as compared to TIP3P data).[29] Additionally, we showed that the PP2 content is underestimated.[29]
- The salt bridge strength is reproduced from TIP3P data but the geometry of the salt bridge is different.[29]
- For GB-Neck2nu model, it tends to break the H-bond of A-T base pair. (chapter 4)

The model studied here and other analytical forms of GB model might be further improved to have better accuracy and better speed by several approaches if we have:

- Better intrinsic radius set to define the solute/solvent boundary. We have been showing that performance of a given GB model is sensitive to the intrinsic radii set. Thus, having good radii set is critically important for the development of GB model.
- More converged data from explicit solvent simulation for benchmarking GB simulation. The end-line-aim of developing a GB model is to replace explicit solvent in case of needing better sampling performance. Thus, direct comparison of GB and explicit solvent MD simulations should always be carried out to validate the accuracy of new GB model. While getting converged data for GB is not difficult, achieve the same task with explicit solvent seems to be laborious work. For example with even small trp-cage protein with 20 residues,[59] it was required to run 208 microsecond to get 12 folding and 12 unfolding event in TIP3P solvent using supercomputer Anton.[46] While we are able to get those numbers of folding/unfolding events using up to 2 microseconds. This took only 2 days of simulation using one GPU core.[167]
- More accurate non-polar term.
- Better treatment of ion in nucleic acid simulation.
- Better theoretical GB model: The analytical form of more theoretical accurate model, R6 model, is on progress.[27]
- New algorithm to accelerate GB effective radii calculation. The current approach of calculating effective has complexity of $O(N^2)$.[194] The solvation computation is thus significantly slow for large molecule.[14, 194]

**Reference**


1. (a) Zagrovic, B.; Pande, V., Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study. *J. Comput. Chem.* **2003,** *24* (12), 1432-1436; (b) Feig, M., Kinetics from Implicit Solvent Simulations of Biomolecules as a Function of Viscosity. *J. Chem. Theory Comput.* **2007,** *3* (5), 1734-1748.
2. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999,** *314* (1-2), 141-151.
3. Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S., Accelerating molecular dynamic simulation on graphics processing units. *J. Comput. Chem.* **2009,** *30* (6), 864-872.
4. Tsui, V.; Case, D. A., Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000,** *122* (11), 2489-2498.
5. (a) Simmerling, C.; Strockbine, B.; Roitberg, A. E., All-Atom Structure Prediction and Folding Simulations of a Stable Protein. *J. Am. Chem. Soc.* **2002,** *124* (38), 11258-11259; (b) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S., Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1−39). *J. Am. Chem. Soc.* **2010,** *132* (5), 1526-1528.
6. Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C., HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **2006,** *103* (4), 915-920.
7. Lopes, A.; Alexandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T., Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins: Structure, Function, and Bioinformatics* **2007,** *67* (4), 853-867.
8. Okur, A.; Strockbine, B.; Hornak, V.; Simmerling, C., Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins. *J. Comput. Chem.* **2003,** *24* (1), 21-31.
9. Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K., On the Nonpolar Hydration Free Energy of Proteins: Surface Area and Continuum Solvent Models for the Solute−Solvent Interaction Energy. *J. Am. Chem. Soc.* **2003,** *125* (31), 9523-9530.
10. Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E., A "universal" surface area correlation for molecular hydrophobic phenomena. *J. Am. Chem. Soc.* **1999,** *121* (39), 9243-9244.
11. Wagoner, J.; Baker, N. A., Solvation forces on biomolecular structures: A comparison of explicit solvent and Poisson–Boltzmann models. *J. Comput. Chem.* **2004,** *25* (13), 1623-1629.
12. Chen, J.; Brooks, C. L., Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.* **2008,** *10* (4), 471-481.
13. (a) Wagoner, J. A.; Baker, N. A., Assessing implicit models for nonpolar mean solvation forces: The importance of dispersion and volume terms. *Proc. Natl. Acad. Sci. USA* **2006,** *103* (22), 8331-8336; (b) Lee, M. S.; Olson, M. A., Comparison of volume and surface area nonpolar solvation free energy terms for implicit solvent simulations. *J. Chem. Phys.* **2013,** *139* (4), -.
14. Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Charles L. Brooks, I., Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **2004,** *25* (2), 265-284.
15. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990,** *112* (16), 6127-6129.
16. Onufriev, A.; Case, D. A.; Bashford, D., Effective Born radii in the generalized Born approximation: The importance of being perfect. *J. Comput. Chem.* **2002,** *23* (14), 1297-1304.

17. Mongan, J.; Svrcek-Seiler, W. A.; Onufriev, A., Analysis of integral expressions for effective Born radii. *J. Chem. Phys.* **2007,** *127* (18), 185101.
18. Lee, M. S.; Salsbury, F. R.; Brooks, C. L., Novel generalized Born methods. *J. Chem. Phys.* **2002,** *116* (24), 10606-10614.
19. Grycuk, T., Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation. *J. Chem. Phys.* **2003,** *119* (9), 4817-4826.
20. Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L., New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **2003,** *24* (11), 1348-1356.
21. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G., Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **1995,** *246* (1-2), 122-129.
22. (a) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A., AMBER 14. *University of California, San Francisco* **2014**; (b) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A., AMBER 10. **2008**; (c) Case, D. A.; Darden, T. A.; Cheatham, T. E. I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossvry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER12*, University of California, San Francisco: 2012.
23. Onufriev, A.; Bashford, D.; Case, D. A., Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **2004,** *55* (2), 383-394.
24. Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A., Generalized Born Model with a Simple, Robust Molecular Volume Correction. *J. Chem. Theory Comput.* **2007,** *3* (1), 156-169.
25. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem* **1983,** *4* (2), 187-217.
26. Im, W.; Lee, M. S.; Brooks, C. L., Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003,** *24* (14), 1691-1702.
27. Aguilar, B.; Shadrach, R.; Onufriev, A. V., Reducing the Secondary Structure Bias in the Generalized Born Model via R6 Effective Radii. *J. Chem. Theory Comput.* **2010,** *6* (12), 3613-3630.
28. Gallicchio, E.; Paris, K.; Levy, R. M., The AGBNP2 Implicit Solvation Model. *J. Chem. Theory Comput.* **2009,** *5* (9), 2544-2564.

29. Nguyen, H.; Roe, D. R.; Simmerling, C., Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J. Chem. Theory Comput.* **2013,** *9* (4), 2020-2034.
30. (a) Chocholoušová, J.; Feig, M., Implicit Solvent Simulations of DNA and DNA−Protein Complexes: Agreement with Explicit Solvent vs Experiment. *J. Phys. Chem. B* **2006,** *110* (34), 17240-17251; (b) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C., Investigation of Salt Bridge Stability in a Generalized Born Solvent Model. *J. Chem. Theory Comput.* **2006,** *2* (1), 115-127.
31. Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C., Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation. *J. Phys. Chem. B* **2007,** *111* (7), 1846-1857.
32. Réblová, K.; Špačková, N. a.; Štefl, R.; Csaszar, K.; Koča, J.; Leontis, N. B.; Šponer, J., Non-Watson-Crick Basepairing and Hydration in RNA Motifs: Molecular Dynamics of 5S rRNA Loop E. *Biophys. J. 84* (6), 3564-3582.
33. Nymeyer, H.; Garcia, A. E., Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. USA* **2003,** *100* (24), 13934-13939.
34. (a) Zhou, R., Free energy landscape of protein folding in water: Explicit vs. implicit solvent. *Proteins: Struct., Funct., Bioinf.* **2003,** *53* (2), 148-161; (b) Okur, A.; Wickstrom, L.; Simmerling, C., Evaluation of Salt Bridge Structure and Energetics in Peptides Using Explicit, Implicit, and Hybrid Solvation Models. *J. Chem. Theory Comput.* **2008,** *4* (3), 488-498; (c) Ruhong, Z.; Bruce, J. B., Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water? *Proc. Natl. Acad. Sci. USA* **2002,** *99* (20), 12777-12782.
35. Cheatham, T. E.; Case, D. A., Twenty-five years of nucleic acid simulations. *Biopolymers* **2013,** *99* (12), 969-977.
36. Gaillard, T.; Case, D. A., Evaluation of DNA Force Fields in Implicit Solvation. *J. Chem. Theory Comput.* **2011**.
37. (a) Dill, K. A.; MacCallum, J. L., The Protein-Folding Problem, 50 Years On. *Science* **2012,** *338* (6110), 1042-1046; (b) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S., To milliseconds and beyond: challenges in the simulation of protein folding. *Current Opinion in Structural Biology* **2013,** *23* (1), 58-65.
38. Neudecker, P.; Robustelli, P.; Cavalli, A.; Walsh, P.; Lundström, P.; Zarrine-Afsar, A.; Sharpe, S.; Vendruscolo, M.; Kay, L. E., Structure of an Intermediate State in Protein Folding and Aggregation. *Science* **2012,** *336* (6079), 362-366.
39. Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K., Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain. *Biophysical Journal* **2008,** *94* (10), L75-L77.
40. Bowman, G. R.; Voelz, V. A.; Pande, V. S., Atomistic Folding Simulations of the Five-Helix Bundle Protein λ6−85. *Journal of the American Chemical Society* **2010,** *133* (4), 664-667.
41. Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S., Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1−39). *Journal of the American Chemical Society* **2010,** *132* (5), 1526-1528.
42. Kim, S. J.; Born, B.; Havenith, M.; Gruebele, M., Real-Time Detection of Protein–Water Dynamics upon Protein Folding by Terahertz Absorption Spectroscopy. *Angewandte Chemie International Edition* **2008,** *47* (34), 6486-6489.

43. Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B., Millisecond-scale molecular dynamics simulations on Anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, ACM: Portland, Oregon, 2009; pp 1-11.
44. Shirts, M.; Pande, V. S., Screen Savers of the World Unite! *Science* **2000,** *290* (5498), 1903-1904.
45. (a) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S., Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *Journal of the American Chemical Society* **2011,** *133* (45), 18413-18419; (b) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R., Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences* **2009,** *106* (45), 19011-19016.
46. Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., How Fast-Folding Proteins Fold. *Science* **2011,** *334* (6055), 517-520.
47. Piana, S.; Lindorff-Larsen, K.; Shaw, D. E., Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences* **2013**.
48. Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012,** *8* (5), 1542-1555.
49. Shell, M. S.; Ritterson, R.; Dill, K. A., A Test on Peptide Stability of AMBER Force Fields with Implicit Solvation. *J. Phys. Chem. B* **2008,** *112* (22), 6878-6886.
50. Feig, M.; Brooks, C. L., Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struc. Biol.* **2004,** *14* (2), 217-224.
51. Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A., BIOMOLECULAR SIMULATIONS: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions. *Annu. Rev. Bioph. Biom.* **2001,** *30* (1), 211-243.
52. Chen, J.; Brooks, C. L., Critical Importance of Length-Scale Dependence in Implicit Modeling of Hydrophobic Interactions. *J. Am. Chem. Soc.* **2007,** *129* (9), 2444-2445.
53. Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A., Computation of electrostatic forces on solvated molecules using the Poisson-Boltzmann equation. *J. Phys. Chem-us.* **1993,** *97* (14), 3591-3600.
54. Onufriev, A.; Bashford, D.; Case, D. A., Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000,** *104* (15), 3712-3720.
55. Zhu, J.; Alexov, E.; Honig, B., Comparative Study of Generalized Born Models:Born Radii and Peptide Folding. *J. Phys. Chem. B* **2005,** *109* (7), 3008-3022.
56. Shang, Y.; Nguyen, H.; Wickstrom, L.; Okur, A.; Simmerling, C., Improving the description of salt bridge strength and geometry in a Generalized Born model. *J. Mol. Graphics Model.* **2011,** *29* (5), 676-684.
57. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.* **2006,** *65* (3), 712-725.

58. Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H., Enhanced Hairpin Stability through Loop Design:The Case of the Protein G B1 Domain Hairpin. *J. Am. Chem. Soc.* **2004,** *126* (23), 7238-7243.
59. Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H., Designing a 20-residue protein. *Nat. Struct. Mol. Biol.* **2002,** *9* (6), 425-430.
60. (a) Fadrná, E.; Špačková, N. a.; Sarzyñska, J.; Koča, J.; Orozco, M.; Cheatham, T. E.; Kulinski, T.; Šponer, J. i., Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields. *J. Chem. Theory Comput.* **2009,** *5* (9), 2514-2530; (b) Showalter, S. A.; Brüschweiler, R., Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks: Application to the AMBER99SB Force Field. *J. Chem. Theory Comput.* **2007,** *3* (3), 961-975; (c) Showalter, S. A.; Brüschweiler, R., Quantitative Molecular Ensemble Interpretation of NMR Dipolar Couplings without Restraints. *J. Am. Chem. Soc.* **2007,** *129* (14), 4158-4159; (d) Lange, O. F.; van der Spoel, D.; de Groot, B. L., Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data. *Biophys. J.* **2010,** *99* (2), 647-655.
61. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983,** *79* (2), 926-935.
62. Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A., Tryptophan zippers: Stable, monomeric β-hairpins. *Proc. Natl. Acad. Sci. USA* **2001,** *98* (10), 5578-5583.
63. Kun Song; Stewart, J. M.; Fesinmeyer, R. M.; Andersen, N. H.; Simmerling, C., Structural insights for designed alanine-rich helices: Comparing NMR helicity measures and conformational ensembles from molecular dynamics simulation. *Biopolymers* **2008,** *89* (9), 747-760.
64. McKnight, C. J.; Matsudaira, P. T.; Kim, P. S., NMR structure of the 35-residue villin headpiece subdomain. *Nat. Struct. Mol. Biol.* **1997,** *4* (3), 180-184.
65. Wickstrom, L.; Bi, Y.; Hornak, V.; Raleigh, D. P.; Simmerling, C., Reconciling the Solution and X-ray Structures of the Villin Headpiece Helical Subdomain: Molecular Dynamics Simulations and Double Mutant Cycles Reveal a Stabilizing Cation Interation. *Biochemistry* **2007,** *46* (12), 3624-3634.
66. Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R., High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proc. Natl. Acad. Sci. USA* **2005,** *102* (21), 7517-7522.
67. Prabu-Jeyabalan, M.; Nalivaika, E. A.; King, N. M.; Schiffer, C. A., Structural Basis for Coevolution of a Human Immunodeficiency Virus Type 1 Nucleocapsid-p1 Cleavage Site with a V82A Drug-Resistant Mutation in Viral Protease. *J. Virol.* **2004,** *78* (22), 12446-12454.
68. Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I., Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. *Biochemistry* **1992,** *31* (40), 9665-9672.
69. Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J., Structure of ubiquitin refined at 1.8Å resolution. *J. Mol. Biol.* **1987,** *194* (3), 531-544.
70. Hodsdon, M. E.; Cistola, D. P., Ligand Binding Alters the Backbone Mobility of Intestinal Fatty Acid-Binding Protein as Monitored by 15N NMR Relaxation and 1H Exchange†. *Biochemistry* **1997,** *36* (8), 2278-2290.
71. Holt, D.; Luengo, J.; Yamashita, D.; Oh, H.; Konialian, A.; Yen, H.; Rozamus, L.; Brandt, M.; Bossard, M., Design, synthesis, and kinetic evaluation of high-affinity FKBP ligands

and the X-ray crystal structures of their complexes with FKBP12. *J. Am. Chem. Soc.* **1993,** *115* (22), 9925-9938.
72. Kuszewski, J.; Gronenborn, A. M.; Clore, G. M., Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. *J. Am. Chem. Soc.* **1999,** *121* (10), 2337-2338.
73. Schenck, H. L.; Gellman, S. H., Use of a Designed Triple-Stranded Antiparallel β-Sheet To Probe β-Sheet Cooperativity in Aqueous Solution. *J. Am. Chem. Soc.* **1998,** *120* (19), 4869-4870.
74. Sauter, C.; Otalora, F.; Gavira, J.-A.; Vidal, O.; Giege, R.; Garcia-Ruiz, J. M., Structure of tetragonal hen egg-white lysozyme at 0.94 A from crystals grown by the counter-diffusion method. *Acta. Crystallogr. D* **2001,** *57* (8), 1119-1126.
75. Roe, D. R.; Hornak, V.; Simmerling, C., Folding Cooperativity in a Three-stranded [beta]-Sheet Model. *J. Mol. Biol.* **2005,** *352* (2), 370-381.
76. Ding, F., Exploring the Structure and Dynamics of HIV-1 PR by MD Simulations. *Ph.D. diss., State University of New York at Stony Brook. (Publication No. AAT 3422802)* **2010**.
77. Gilson, M. K.; Sharp, K. A.; Honig, B. H., Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comput. Chem.* **1988,** *9* (4), 327-335.
78. Sigalov, G.; Scheffel, P.; Onufriev, A., Incorporating variable dielectric environments into the generalized Born model. *J. Chem. Phys.* **2005,** *122* (9), 094511-15.
79. Bondi, A., van der Waals Volumes and Radii. *J. Phys. Chem.* **1964,** *68* (3), 441-451.
80. Powell, M. J. D., UOBYQA: unconstrained optimization by quadratic approximation. *Math. Program.* **2002,** *92* (3), 555-582.
81. Forrest, S., Genetic algorithms: principles of natural selection applied to computation. *Science* **1993,** *261* (5123), 872-878.
82. Metcalfe, T. S.; Charbonneau, P., Stellar structure modeling using a parallel genetic algorithm for objective global optimization. *J. Comput. Phys.* **2003,** *185* (1), 176-193.
83. Leardi, R., Genetic algorithms in chemometrics and chemistry: a review. *J. Chemometr.* **2001,** *15* (7), 559-569.
84. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-2637.
85. Wickstrom, L.; Okur, A.; Song, K.; Hornak, V.; Raleigh, D. P.; Simmerling, C. L., The Unfolded State of the Villin Headpiece Helical Subdomain: Computational Studies of the Role of Locally Stabilized Structure. *J. Mol. Biol.* **2006,** *360* (5), 1094-1107.
86. Blanco, F. J.; Rivas, G.; Serrano, L., A short linear peptide that folds into a native stable [beta]-hairpin in aqueous solution. *Nat. Struct. Mol. Biol.* **1994,** *1* (9), 584-590.
87. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977,** *23* (3), 327-341.
88. Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984,** *81* (8), 3684-3690.
89. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993,** *98* (12), 10089-10092.
90. Simmerling, C.; Elber, R.; Zhang, J., MOIL-View - A Program for Visualization of Structure and Dynamics of Biomolecules and STO - A Program for Computing Stochastic Paths. *Modelling of Biomolecular Structures and Mechanisms* **1995**, 241-465.

91. Wickstrom, L.; Okur, A.; Simmerling, C., Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophys. J.* **2009,** *97* (3), 853-856.
92. Hsieh, M.-J.; Luo, R., Balancing Simulation Accuracy and Efficiency with the Amber United Atom Force Field. *J. Phys. Chem. B* **2010,** *114* (8), 2886-2893.
93. Day, R.; Paschek, D.; Garcia, A. E., Microsecond simulations of the folding/unfolding thermodynamics of the Trp-cage miniprotein. *Proteins: Struct., Funct., Bioinf.* **2010,** *78* (8), 1889-1899.
94. (a) Pitera, J. W.; Swope, W., Understanding folding and design: Replica-exchange simulations of ``Trp-cage" miniproteins. *Proc. Natl. Acad. Sci. USA* **2003,** *100* (13), 7587-7592; (b) Zhou, R., Trp-cage: Folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. USA* **2003,** *100* (23), 13280-13285; (c) Paschek, D.; Hempel, S.; García, A. E., Computing the stability diagram of the Trp-cage miniprotein. *Proc. Natl. Acad. Sci. USA* **2008,** *105* (46), 17754-17759.
95. Michael, S. L.; Freddie R. Salsbury, Jr.; Charles, L. B., III, Novel generalized Born methods. *J. Chem. Phys.* **2002,** *116* (24), 10606-10614.
96. Onufriev, A. V., (personal communication). **2010**
97. (a) Lyne, P. D.; Lamb, M. L.; Saeh, J. C., Accurate Prediction of the Relative Potencies of Members of a Series of Kinase Inhibitors Using Molecular Docking and MM-GBSA Scoring. *J. Med. Chem* **2006,** *49* (16), 4805-4808; (b) Guimarães, C. R. W.; Cardozo, M., MM-GB/SA Rescoring of Docking Poses in Structure-Based Lead Optimization. *J. Chem. Inf. Model.* **2008,** *48* (5), 958-970.
98. Yearly Growth of Protein Structures. http://www.pdb.org/pdb/statistics/contentGrowthChart.do?content=molType-protein (accessed March 27).
99. Pruitt, K. D.; Brown, G. R.; Hiatt, S. M.; Thibaud-Nissen, F.; Astashyn, A.; Ermolaeva, O.; Farrell, C. M.; Hart, J.; Landrum, M. J.; McGarvey, K. M.; Murphy, M. R.; O'Leary, N. A.; Pujar, S.; Rajput, B.; Rangwala, S. H.; Riddick, L. D.; Shkeda, A.; Sun, H.; Tamez, P.; Tully, R. E.; Wallin, C.; Webb, D.; Weber, J.; Wu, W.; DiCuccio, M.; Kitts, P.; Maglott, D. R.; Murphy, T. D.; Ostell, J. M., RefSeq: an update on mammalian reference sequences. *Nucleic acids research* **2014,** *42* (Database issue), D756-63.
100. Piana, S.; Lindorff-Larsen, K.; Shaw, D. E., Atomic-level description of ubiquitin folding. *P Natl Acad Sci USA* **2013,** *110* (15), 5915-5920.
101. Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D., Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **1999**, 171-176.
102. (a) MacCallum, J. L.; Perez, A.; Dill, K. A., Integrative Modeling of Protein Conformational Ensembles using Limited Data. *Biophys J* **2013,** *104* (2), 546a-547a; (b) Marks, D. S.; Hopf, T. A.; Sander, C., Protein structure prediction from sequence variation. *Nat Biotech* **2012,** *30* (11), 1072-1080.
103. Bonneau, R.; Tsai, J.; Ruczinski, I.; Chivian, D.; Rohl, C.; Strauss, C. E. M.; Baker, D., Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins-Structure Function and Genetics* **2001**, 119-126.
104. Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; Mcleavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y. B.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C.,

Anton, a special-purpose machine for molecular dynamics simulation. *Commun Acm* **2008,** *51* (7), 91-97.
105. Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* **2012,** *8* (5), 1542-1555.
106. Nguyen, H.; Roe, D. R.; Simmerling, C., Improved Generalized Born Solvent Model Parameters for Protein Simulations. *Journal of Chemical Theory and Computation* **2013,** *9* (4), 2020-2034.
107. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics* **2006,** *65* (3), 712-725.
108. Maier, J.; Martinez, C.; Wickstrom, L.; Simmerling, C., A comprehensive revision of AMBER protein dihedral corrections. **unpublished data**.
109. Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K., Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain. *Biophys. J.* **2008,** *94* (10), L75-L77.
110. (a) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K., On the Nonpolar Hydration Free Energy of Proteins: Surface Area and Continuum Solvent Models for the Solute−Solvent Interaction Energy. *Journal of the American Chemical Society* **2003,** *125* (31), 9523-9530; (b) Wagoner, J. A.; Baker, N. A., Assessing implicit models for nonpolar mean solvation forces: The importance of dispersion and volume terms. *Proceedings of the National Academy of Sciences* **2006,** *103* (22), 8331-8336.
111. D.A. Case, T. A. D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman, AMBER 14. *University of California, San Francisco* **(2014)**.
112. (a) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C., Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem* **1999,** *20* (8), 786-798; (b) Hopkins, C.; Roitberg, A., **unpublished data**.
113. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **2006,** *65* (3), 712-725.
114. Møller, C.; Plesset, M. S., Note on an Approximation Treatment for Many-Electron Systems. *Physical Review* **1934,** *46* (7), 618-622.
115. Ditchfield, R.; Hehre, W. J.; Pople, J. A., Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **1971,** *54* (2), 724-728.
116. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995,** *117* (19), 5179-5197.
117. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., General atomic and molecular electronic structure system. *J Comput Chem* **1993,** *14* (11), 1347-1363.

118. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 98, Revision A.7*, Gaussian, Inc.: Pittsburgh, PA, USA, 1998.
119. Case, D. A., Cheatham, T. E., Darden, H., Gohlke, R., Luo, R., Merz, K. M., Jr., Onufriev, A., Simmerling, C., Wang, B. and Woods, R., The Amber biomolecular simulation programs. *J Comput Chem* **2005,** *26*, 1668-1688.
120. Holland, J. H., *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press: Ann Arbor, 1975; p viii, 183 p.
121. Wall, M. *GAlib genetic algorithm*, Massachussetts Institute of Technology.
122. Roe, D. R.; Cheatham, T. E., PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013,** *9* (7), 3084-3095.
123. Gohlke, H.; Kiel, C.; Case, D. A., Insights into Protein–Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras–Raf and Ras–RalGDS Complexes. *Journal of Molecular Biology* **2003,** *330* (4), 891-913.
124. Lipari, G.; Szabo, A., Nuclear magnetic resonance relaxation in nucleic acid fragments: models for internal motion. *Biochemistry* **1981,** *20* (21), 6250-6256.
125. Prompers, J. J.; Brüschweiler, R., General Framework for Studying the Dynamics of Folded and Nonfolded Proteins by NMR Relaxation Spectroscopy and MD Simulation. *Journal of the American Chemical Society* **2002,** *124* (16), 4522-4534.
126. Li, D.-W.; Brüschweiler, R., Iterative Optimization of Molecular Mechanics Force Fields from NMR Data of Full-Length Proteins. *J Chem Theory Comput* **2011,** *7* (6), 1773-1782.
127. Feng, W.; Tejero, R.; Zimmerman, D. E.; Inouye, M.; Montelione, G. T., Solution NMR Structure and Backbone Dynamics of the Major Cold-Shock Protein (CspA) from Escherichia coli: Evidence for Conformational Dynamics in the Single-Stranded RNA-Binding Site†,‡. *Biochemistry* **1998,** *37* (31), 10881-10896.
128. (a) Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. S., HOW WELL CAN SIMULATION PREDICT PROTEIN FOLDING KINETICS AND THERMODYNAMICS? *Annual Review of Biophysics and Biomolecular Structure* **2005,** *34* (1), 43-69; (b) Yun-yu, S.; Lu, W.; Van Gunsteren, W. F., On the Approximation of Solvent Effects on the Conformation and Dynamics of Cyclosporin A by Stochastic Dynamics Simulation Techniques. *Molecular Simulation* **1988,** *1* (6), 369-383.
129. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983,** *79* (2), 926-935.

130. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *Journal of Chemical Physics* **1993,** *98* (12), 10089-10092.
131. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.
132. Ciccotti, G.; Ryckaert, J. P., Molecular dynamics simulation of rigid molecules. *Computer Physics Reports* **1986,** *4* (6), 346-392.
133. Honda, S.; Akiba, T.; Kato, Y. S.; Sawada, Y.; Sekijima, M.; Ishimura, M.; Ooishi, A.; Watanabe, H.; Odahara, T.; Harata, K., Crystal Structure of a Ten-Amino Acid Protein. *J. Am. Chem. Soc.* **2008,** *130* (46), 15327-15331.
134. Sarisky, C. A.; Mayo, S. L., The ββα fold: explorations in sequence space. *Journal of Molecular Biology* **2001,** *307* (5), 1411-1418.
135. Liu, F.; Du, D.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M., An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proceedings of the National Academy of Sciences* **2008,** *105* (7), 2369-2374.
136. Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E., Computational Design and Experimental Testing of the Fastest-Folding β-Sheet Protein. *Journal of Molecular Biology* **2011,** *405* (1), 43-48.
137. Gronwald, W.; Hohm, T.; Hoffmann, D., Evolutionary Pareto-optimization of stably folding peptides. *BMC Bioinformatics* **2008,** *9* (1), 109.
138. Horng, J.-C.; Moroz, V.; Raleigh, D. P., Rapid Cooperative Two-state Folding of a Miniature α–β Protein and Design of a Thermostable Variant. *Journal of Molecular Biology* **2003,** *326* (4), 1261-1270.
139. Neuweiler, H.; Sharpe, T. D.; Rutherford, T. J.; Johnson, C. M.; Allen, M. D.; Ferguson, N.; Fersht, A. R., The Folding Mechanism of BBL: Plasticity of Transition-State Structure Observed within an Ultrafast Folding Protein Family. *Journal of Molecular Biology* **2009,** *390* (5), 1060-1073.
140. Shah, P. S.; Hom, G. K.; Ross, S. A.; Lassila, J. K.; Crowhurst, K. A.; Mayo, S. L., Full-sequence computational design and solution structure of a thermostable protein variant. *J.Mol.Biol.* **2007,** *372* (1), 1-6.
141. Nauli, S.; Kuhlman, B.; Le Trong, I.; Stenkamp, R. E.; Teller, D.; Baker, D., Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2. *Protein Science* **2002,** *11* (12), 2924-2931.
142. Schindelin, H.; Jiang, W.; Inouye, M.; Heinemann, U., Crystal Structure of CspA, the Major Cold Shock Protein of Escherichia coli. *Proc.Natl.Acad.Sci.USA* **1994,** *91* (11), 5119-5123.
143. Kanagawa, M.; Yokoyama, S.; Kuramitsu, S., unpublished data.
144. Walsh, S. T. R.; Cheng, H.; Bryson, J. W.; Roder, H.; DeGrado, W. F., Solution structure and dynamics of a de novo designed three-helix bundle protein. *P Natl Acad Sci USA* **1999,** *96* (10), 5486-5491.
145. Beamer, L. J.; Pabo, C. O., Refined 1.8 Å crystal structure of the λ repressor-operator complex. *J.Mol.Biol.* **1992,** *227* (1), 177-196.
146. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D., Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **2003,** *302* (5649), 1364-1368.

147. Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C., Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model. *J Chem Theory Comput* **2006,** *2* (2), 420-433.
148. Reid, K. L.; Rodriguez, H. M.; Hillier, B. J.; Gregoret, L. M., Stability and folding properties of a model β-sheet protein, Escherichia coli CspA. *Protein Science* **1998,** *7* (2), 470-479.
149. Scalley-Kim, M.; Baker, D., Characterization of the Folding Energy Landscapes of Computer Generated Proteins Suggests High Folding Free Energy Barriers and Cooperativity may be Consequences of Natural Selection. *Journal of Molecular Biology* **2004,** *338* (3), 573-583.
150. Buck, M.; Boyd, J.; Redfield, C.; MacKenzie, D. A.; Jeenes, D. J.; Archer, D. B.; Dobson, C. M., Structural Determinants of Protein Dynamics: Analysis of 15N NMR Relaxation Measurements for Main-Chain and Side-Chain Nuclei of Hen Egg White Lysozyme. *Biochemistry* **1995,** *34* (12), 4041-4055.
151. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic acids research* **2000,** *28* (1), 235-242.
152. Davis, C. M.; Xiao, S.; Raleigh, D. P.; Dyer, R. B., Raising the Speed Limit for β-Hairpin Formation. *Journal of the American Chemical Society* **2012,** *134* (35), 14476-14482.
153. Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J., Smaller and Faster: The 20-Residue Trp-Cage Protein Folds in 4 μs. *Journal of the American Chemical Society* **2002,** *124* (44), 12952-12953.
154. Jäger, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M. E.; Noel, J. P.; Gruebele, M.; Kelly, J. W., Structure–function–folding relationship in a WW domain. *Proc. Natl. Acad. Sci. USA* **2006,** *103* (28), 10648-10653.
155. Wang, M.; Tang, Y.; Sato, S.; Vugmeyster, L.; McKnight, C. J.; Raleigh, D. P., Dynamic NMR Line-Shape Analysis Demonstrates that the Villin Headpiece Subdomain Folds on the Microsecond Time Scale. *Journal of the American Chemical Society* **2003,** *125* (20), 6032-6033.
156. Horng, J.-C.; Moroz, V.; Raleigh, D. P., Rapid Cooperative Two-state Folding of a Miniature α–β Protein and Design of a Thermostable Variant. *J. Mol. Biol.* **2003,** *326* (4), 1261-1270.
157. Neuweiler, H.; Sharpe, T. D.; Rutherford, T. J.; Johnson, C. M.; Allen, M. D.; Ferguson, N.; Fersht, A. R., The Folding Mechanism of BBL: Plasticity of Transition-State Structure Observed within an Ultrafast Folding Protein Family. *J.Mol.Biol.* **2009,** *390* (5), 1060-1073.
158. Johansson, M. U.; de Château, M.; Wikström, M.; Forsén, S.; Drakenberg, T.; Björck, L., Solution structure of the albumin-binding GA module: a versatile bacterial protein domain. *J.Mol.Biol.* **1997,** *266* (5), 859-865.
159. Wang, T.; Zhu, Y.; Gai, F., Folding of A Three-Helix Bundle at the Folding Speed Limit. *The Journal of Physical Chemistry B* **2004,** *108* (12), 3694-3697.
160. Gillespie, B.; Vu, D. M.; Shah, P. S.; Marshall, S. A.; Dyer, R. B.; Mayo, S. L.; Plaxco, K. W., NMR and Temperature-jump Measurements of de Novo Designed Proteins Demonstrate Rapid Folding in the Absence of Explicit Selection for Kinetics. *Journal of Molecular Biology* **2003,** *330* (4), 813-819.
161. Nauli, S.; Kuhlman, B.; Baker, D., Computer-based redesign of a protein folding pathway. *Nature Structural & Molecular Biology* **2001,** *8* (7), 602-605.

162. Zhu, Y.; Alonso, D. O. V.; Maki, K.; Huang, C.-Y.; Lahr, S. J.; Daggett, V.; Roder, H.; DeGrado, W. F.; Gai, F., Ultrafast folding of α3D: A de novo designed three-helix bundle protein. *Proceedings of the National Academy of Sciences* **2003,** *100* (26), 15486-15491.
163. Yang, W. Y.; Gruebele, M., Folding at the speed limit. *Nature* **2003,** *423* (6936), 193-197.
164. Stadlbauer, P.; Krepl, M.; Cheatham, T. E.; Koča, J.; Šponer, J., Structural dynamics of possible late-stage intermediates in folding of quadruplex DNA studied by molecular simulations. *Nucl. Acids Res.* **2013,** *41* (14), 7128-7143.
165. Cheatham, T. E., III; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A., Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins. *J. Am. Chem. Soc.* **1995,** *117* (14), 4193-4194.
166. (a) Roitberg, A. E.; Okur, A.; Simmerling, C., Coupling of Replica Exchange Simulations to a Non-Boltzmann Structure Reservoir. *J. Phys. Chem. B* **2007,** *111* (10), 2415-2418; (b) Fukunishi, H.; Watanabe, O.; Takada, S., On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002,** *116* (20), 9058-9067.
167. Nguyen, H.; Maier, J.; Simmerling, C., Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field, implicit solvent and GPU. *(submitted).*
168. D.A. Case, T. A. D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman, AMBER 12. *University of California, San Francisco* **(2012)**.
169. (a) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A., Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. *J. Am. Chem. Soc.* **1998,** *120* (37), 9401-9409; (b) Tsui, V.; Case, D. A., Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers* **2000,** *56* (4), 275-291; (c) Ruscio, J. Z.; Onufriev, A., A Computational Study of Nucleosomal DNA Flexibility. *Biophys. J.* **2006,** *91* (11), 4121-4132.
170. Chocholoušová, J.; Feig, M., Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations. *J. Comput. Chem.* **2006,** *27* (6), 719-729.
171. Feenstra, K. A.; Hess, B.; Berendsen, H. J. C., Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **1999,** *20* (8), 786-798.
172. Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S., To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struc. Biol.* **2013,** *23* (1), 58-65.
173. Powell, M. J. D., The NEWUOA software for unconstrained optimization without derivatives. In *Large-Scale Nonlinear Optimization*, Pillo, G.; Roma, M., Eds. Springer US: 2006; Vol. 83, pp 255-297.
174. Rios, L.; Sahinidis, N., Derivative-free optimization: a review of algorithms and comparison of software implementations. *J Glob Optim* **2013,** *56* (3), 1247-1293.

175. Cheatham, T. E.; Kollman, P. A., Molecular Dynamics Simulations Highlight the Structural Differences among DNA:DNA, RNA:RNA, and DNA:RNA Hybrid Duplexes. *J. Am. Chem. Soc.* **1997,** *119* (21), 4805-4825.
176. Wing, R.; Drew, H.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R., Crystal structure analysis of a complete turn of B-DNA. *Nature* **1980,** *287* (5784), 755.
177. (a) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham III, T. E.; Laughton, C. A.; Orozco, M., Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **2007,** *92* (11), 3817-3829; (b) Pérez, A.; Luque, F. J.; Orozco, M., Dynamics of B-DNA on the Microsecond Time Scale. *Journal of the American Chemical Society* **2007,** *129* (47), 14739-14745.
178. Allen, M. D.; Yamasaki, K.; Ohme Takagi, M.; Tateno, M.; Suzuki, M., A novel mode of DNA recognition by a β-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *EMBO J.* **1998,** *17* (18), 5484-5496.
179. Pérez, A.; Lankas, F.; Luque, F. J.; Orozco, M., Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucl. Acids Res.* **2008,** *36* (7), 2379-2394.
180. (a) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E.; Šponer, J. i.; Otyepka, M., Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **2010,** *6* (12), 3836-3849; (b) Zgarbová, M.; Otyepka, M.; Šponer, J. i.; Mládek, A. t.; Banáš, P.; Cheatham III, T. E.; Jurecka, P., Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **2011,** *7* (9), 2886-2902.
181. Zhu, L.; Chou, S.-H.; Xu, J.; Reid, B. R., Structure of a single-cytidine hairpin loop formed by the DNA triplet GCA. *Nat Struct Biol.* **1995,** *2* (11), 1012-1017.
182. Kannan, S.; Zacharias, M., Role of the closing base pair for d(GCA) hairpin stability: free energy analysis and folding simulations. *Nucleic Acids Res.* **2011,** *39* (19), 8271-8280.
183. Nozinovic, S.; Fürtig, B.; Jonker, H. R. A.; Richter, C.; Schwalbe, H., High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucl. Acids Res.* **2010,** *38* (2), 683-694.
184. Zhang, L. Y.; Gallicchio, E.; Friesner, R. A.; Levy, R. M., Solvent models for protein–ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. *J. Comput. Chem.* **2001,** *22* (6), 591-607.
185. Larsen, T. A.; Goodsell, D. S.; Cascio, D.; Grzeskowiak, K.; Dickerson, R. E., The Structure of DAPI Bound to DNA. *J.Biomol.Struct.Dyn.* **1989,** *7* (3), 477-491.
186. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25* (9), 1157-1174.
187. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002,** *23* (16), 1623-1641.
188. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Model.* **2006,** *25* (2), 247-260.
189. Roe, D. R.; Cheatham, T. E., PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013,** *9* (7), 3084-3095.

190. Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K., Conformational analysis of nucleic acids revisited: Curves+. *Nucl. Acids Res.* **2009,** *37* (17), 5917-5929.
191. Haider, S. M.; Parkinson, G. N.; Neidle, S., Structure of a G-quadruplex–Ligand Complex. *J.Mol.Biol.* **2003,** *326* (1), 117-125.
192. Kopka, M. L.; Yoon, C.; Goodsell, D.; Pjura, P.; Dickerson, R. E., The molecular origin of DNA-drug specificity in netropsin and distamycin. *Proc. Natl. Acad. Sci. USA* **1985,** *82* (5), 1376-1380.
193. Haider, S. M.; Parkinson, G. N.; Neidle, S., Structure of a G-quadruplex–Ligand Complex. *J Mol Biol.* **2003,** *326* (1), 117-125.
194. Anandakrishnan, R.; Daga, M.; Onufriev, A. V., An n log n Generalized Born Approximation. *J. Chem. Theory Comput.* **2011,** *7* (3), 544-559.