Supporting Information

Genetic Algorithm Based Design and Experimental Characterization of a Highly Thermostable Metalloprotein

*E. Bozkurt[‡], M. A. S. Perez[‡], R. Hovius[†], N. Browning[‡], and U. Rothlisberger[‡]**

‡ Laboratory of Computational Chemistry and Biochemistry, École Polytechnique Fédérale de Lausanne, EPFL, CH-1015 Lausanne, Switzerland.

† Laboratory of Protein Engineering, École Polytechnique Fédérale de Lausanne, EPFL, CH-1015 Lausanne, Switzerland.

Table of Contents

1. Experimental procedure

1.1. Protein expression and purification

The expression plasmid pET-17b carrying a codon region of wild-type GB1 (with T2Q mutation) and the designed mutant (CATATGCAGTTTAAACTGATTCTGAACGGCAAAACCCTGAAAGGTGTGATTACG ATCGAAGCAGTTGATCATGCGGAAGCCGAAAAATTTTTCAAACAATACGCTAAC GATAATGGCGTCGACGGTGAATGGACCTACGACGAAGCCACCCACACGTTCACC GTTACGGAATAACTCGAG, gene length: 180 bp), synthesized by Genescript (USA) were transformed into *E. coli* strain BL21 (DE3). For expression, 1 L LB medium, supplemented with ampicillin (100 µg/L), was inoculated with a small overnight culture and were grown to an OD_{600} of 0.6. Protein expression was then induced with 0.5 mM IPTG (isopropylthio-D-galactoside) and cell cultures were incubated at 18 ^{o}C overnight with shaking at 220 RPM. Cells were harvested by centrifugation at 3,900 g for 30 min at 4 °C.

The procedure to purify the soluble proteins were essentially the same as previously reported[1]. The only exception was the pH (7.5) of the buffers used in the anion exchange purification part for the metallo-GB1 variant. All purification steps were conducted at room temperature. Concentrated fractions of GB1 mutant were buffer exchanged with the Amicon centrifugal filters (3K) against a solution containing 20 mM Tris, 15 mM NaCI, pH 7.5 before proceeding to the crystallization experiments. Protein concentrations were determined by using the theoretical extinction coefficients (WT: 9970 M^{-1} cm^{-1}, mutant: 8480 M^{-1} cm^{-1}). The total isolated yield for the designed mutant was 40 mg per L of culture. Protein identities were confirmed by mass spectrometry.

Table S1. MS for purified proteins.

	$[M+H]^+$ m/z average	
	Calculated	**Observed**
WT-GB1	6222.0	6220.0
6B1$^{23H/47E/50H}$	6366.0	6364.0

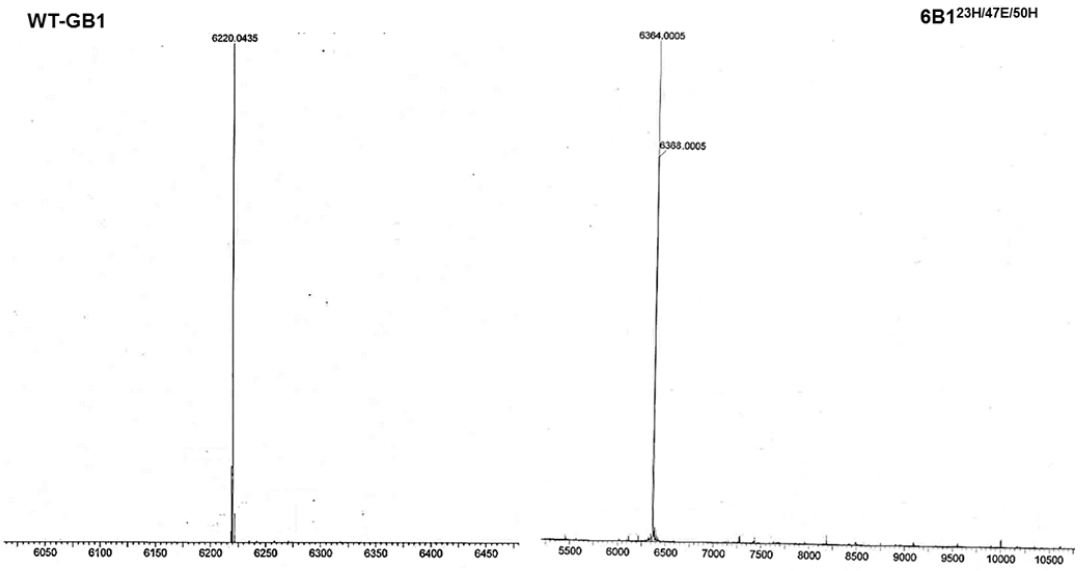

Figure S1. Mass spectra of wild-type and mutant proteins.

1.2. Crystallization, data collection and refinement

We performed both robotic and manual crystallization trials with 30 mg/mL protein. Good diffraction crystals were obtained in a home-made solution containing 70% MPD, 200 mM $ZnSO_4$, 50 mM NaCI at pH 4.5 within a few minutes by hanging-drop vapor diffusion with 500 µL reservoirs and droplets consisting of 2 µL of protein and 1 µL of precipitant solution. Crystals required freezing upon formation, maximum 1-2 hours, since they start to degrade if they are incubated for longer periods either at 18ºC or 4 ºC. Notably, we were able to create the best diffracting quality crystals by slightly differentiating this condition (70% MPD, 100

mM $ZnSO_4$, 100 mM NaCl) and drop ratio (5:1) without degradation. All crystals were cryoprotected in their corresponding mother liquor containing 25% (v/v) ethylene glycol and frozen in liquid nitrogen for data collection. Molecular replacement was carried out using Phaser[2] with a GB1 monomer (PDB ID: 1IGD) as the search model. Data refinement and model building were conducted using REFMAC[3] and COOT[4]. Additionally, PDB_REDO was used for automated re-refinement[5].

Table S2. X-ray crystallography data collection and refinement statistics.

	Zn-GB1 (1)	**Zn-(GB1) (2)**
PDB code	**5O94**	**5OFS**
Data collection		
Space group	P 32 2 1	P 21 21 2
a, b, c (Å)	54.18 54.18 81.08	51.1 74.4 76.6
α, β, γ (°)	90 90 120	90 90 90
Resolution (Å)	1.9 - 46.92	1.1 – 42.37
<I/σ(I)>	1.9	1.66
Completeness	99.9%	96.0 %
Refinement		
Resolution (Å)	1.9	1.1
No. unique reflections	10822	106704
R_{work}/R_{free}	0.21/0.24	0.13/0.15
No. atoms		
Total number of atoms	532	2176
Anisotropy	0.1	1.4
R.m.s. deviations		
Bond lengths (Å)	0.020	0.020
Bond angles (°)	2.040	1.935
Average B-factor (Å²)	51.0	18.0
Wilson B-factor (Å²)	35.5	7.9

1.3. Circular dichroism (CD) scans

CD measurements were performed on a Jasco-810 spectropolarimeter using a 1-mm cuvette at 20 °C and 50 °C. CD data were obtained in different buffers (10 mM potassium phosphate and 10 mM Tris-HCl) and were recorded from 190 to 260 nm with a bandwidth of 0.5 nm and an average of 3 scans. All CD buffers were prepared with Mili-Q water and treated with Chelex@100 to remove trace metal ions. Protein and metal concentrations were 10 µM and 20 µM, respectively for single scans. All collected spectra were baseline corrected. The ellipticity at 222 nm was monitored to follow denaturisation of the protein in 2 °C increments. T_m values were determined by using 2 and 10 µM of GB1 variant. The apo form of the designed GB1 mutant was assessed by adding excess EDTA (> 5 equivalents). Data were analysed by CDPal software[6].

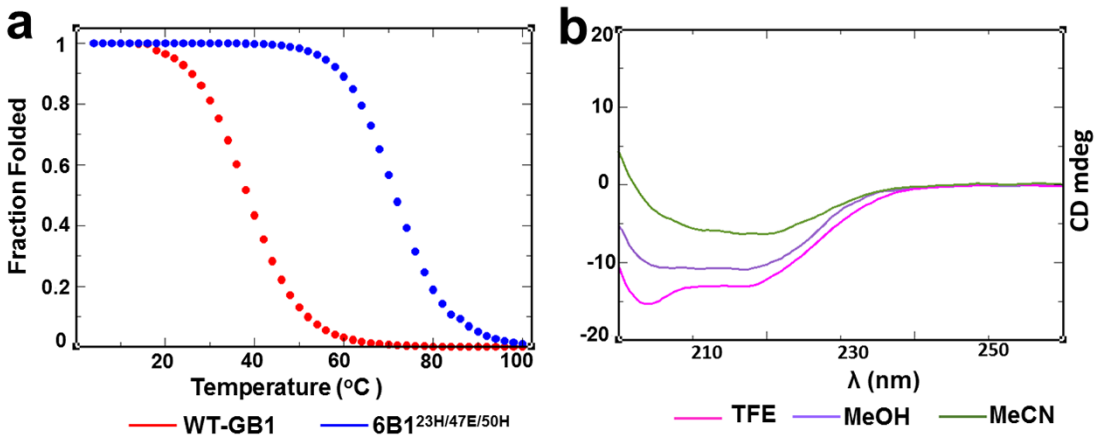

Figure S2. (a) thermal unfolding transitions of wild-type GB1 and computationally designed mutant. (b) Far-UV CD spectra of metallo-GB1 at 50°C in organic solvents.

1.4. Analytical ultracentrifugation (AUC)

Analytical ultracentrifugation of apo and holo samples were performed using a Beckman Coulter XL-A equipped with an An-60 Ti rotor at 20 °C. Samples were in 20 mM Tris, 100

mM NaCl (pH 7.5) buffer. The buffer was also treated with Chelex@100. 50 µM protein and excess $ZnCl_2$ (200 µM) were used. Absorbance was measured at a wavelength of 280 nm. The sedimentation velocity was assayed at a speed of 55000 rpm. Data were analysed by Sedfit[7] and AUC figures were created by Gussi[8].

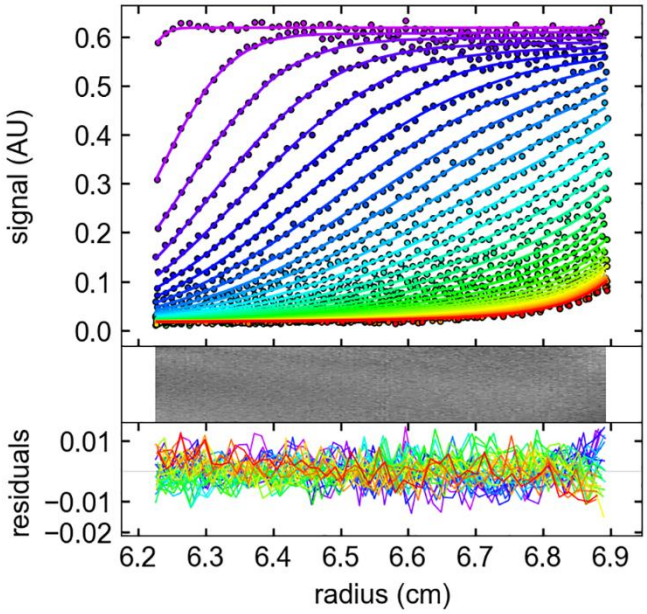

Figure S3. Representative sedimentation velocity data using 50 µM GB1 mutant without zinc. The data is evaluated at 280 nm.

2. Computational Procedure

2.1. Genetic algorithm (GA) optimization protocol

We have developed a python-based[9] genetic algorithm toolbox, which uses Open Babel[10] to interpret biomolecular graphs. With the goal of creating a favourable metal binding site in GB1 in an efficient way, we have employed GA optimization in the following way:

1. A representative frame (250 ns) is taken from the molecular dynamics simulation of wild-type GB1, and used as an initial seed.

2. The initial population is created by randomly sampling a rotamer library by mutating the desired residues of wild-type GB1. Here, we used the Richardson rotamer library,[11] which consists of 177 side chain conformations.

3. The stability of each individual is evaluated through a fitness function. The fitness function f is a measure akin to the folding free energy difference, $\Delta\Delta$G, which is defined as the energy difference between a given mutant and the native structure ($E^{Mut}_{amber,folded}$ - $E^{WT}_{amber,folded}$) plus the difference between the sum of the single amino acid reference energies for the native and those of the mutated sequence $\left(\sum E^{WT}_{amber,reference} - \sum E^{Mut}_{amber,reference}\right)$. The energies of the wild-type and mutant proteins structures were evaluated with the force field Amber ff99SB[12] using the AMBER suite of programs. Each minimization procedure consists of 14000 steps of steepest descent followed by 1000 steps of conjugate gradient minimization. Solvation effects were represented implicitly using the Generalized Born/Solvent-Accessible surface area model (GB/SA)[13], developed by A. Onufriev et al.[14,15] with the parameters alpha, beta and gamma set to 1.0, 0.8 and 4.8, respectively. These settings have previously been tested on a prototypical alpha helix in solution. Electrostatic and van der Waals interactions were truncated at a cutoff of 16 Å. We used a dielectric constant of 80 to mimic the water environment.

The amino acid reference energy, $E_{amber,reference}$, was calculated for isolated amino acids in the same implicit solvent. For each amino acid, side chain conformations taken from the Richardson rotamer library are minimized using the force field based procedure as described previously and the rotamer with the lowest energy is taken as reference.

4. Mutants and their fitness values are stored.

5. A tournament selection with replacement[54] is made based on the fitness of this initial population to select parent individuals.

6. Simulated binary crossover with polynomial order 10 and genewise swap probability of 0.5 and polynomial mutation[16,17] operations are conducted to create a new pool of candidates for the next generation.

7. This procedure loops iteratively until an optimal combination of mutations of wild-type GB1 is achieved.

Optimal values for mutation probability, crossover probability and population size were 0.04, 0.7 and 200, respectively.

2.2. Molecular dynamics (MD) simulations

All-atom molecular dynamics simulations were conducted based on the ff99SB[12] force field using the AMBER 14 software. The NMR structure was obtained from the Protein Data Bank (PDB entry: 3GB1[18], 22nd conformation) and employed as the starting structure. Mutations were introduced utilizing the Leap module of AMBER14. Protonation states of charged residues were assigned at pH 7.4 using the H++ server (http://biophysics.cs.vt.edu/)[19]. Monomeric systems were solvated with a TIP3P water model in cubic boxes with a 0.15 nm distance from the protein to the edge of the box and dimeric forms were solved in rectangular boxes with 0.22 nm to the edge of the box. Sodium and/or chloride ions were added to maintain neutrality of the systems. All systems were first minimized to remove bad contacts between the solute and solvent. Subsequently, all systems were heated from 10 to 300 K in 300 ps with a time step of 1 fs applying weak restraints on the backbone and side chain atoms. Restraints are gradually released within 3.75 ns, followed by extensive MD simulations at a constant temperature of 300 K with a time step of 1.5 fs. The temperature was maintained by a Langevin[20] thermostat. A cutoff of 12 Å is used for non-bonded interactions and the long-range interactions are treated by the Particle Mesh Ewald (PME)

method[21,22]. The SHAKE[23] algorithm was applied to constrain the bonds involving hydrogen atoms.

We carried out two separate MD simulations for 1) the metal-unbound and 2) the metal-bound forms. All simulations performed in the present study are summarized in Supplementary Table S3. The protonation states of His23 and His50 were determined carrying out two MD simulations with two different protonation states of the histidines (50 ns) and found to be in good agreement with the crystal structure. Based on these simulations, the metal binding site was built by using the active site of thermolysin. We used a non-bonded model for the Zn^{+2} ion (default model in AMBER) to simulate $His_2GluWat$ type coordination since in a previous study, Ryde et al. concluded that nonbonded models produce more stable results for certain zinc coordination spheres[24]. Four different non-bonded parameter sets[25,26] were examined for the zinc ion in the His_2Glu_2 type coordination sphere. Nonetheless, all of them lead to pentacoordination with an extra water molecule therefore, we used the crystal structure as basis for the QM/MM simulations.

2.3. QM/MM molecular dynamics

QM/MM Born-Oppenheimer (BO) MD simulations were performed using the CPMD[27] code. The QM systems were composed of the coordinating histidine residues (23H and 50H), glutamate (47E), the zinc ion and one water molecule (monomer and side-by-side dimer) or the C-terminus of glutamate (56E) (head-to-tail dimer). We utilized the BLYP[28] exchange-correlation functional to treat the QM region and Troullier-Martins norm-conserving pseudopotentials[29]. For zinc, we used a pseudopotential with cutoff radii r_s = 1.8 au, r_p = 2.0 au, r_d = 1.8 au and r_f = 1.2 au[30]. The rest of the system was described at the classical level (MM) by the AMBER ff99SB force field. The time step was 0.24 fs (10 au). The QM/MM boundary was treated with monovalent carbon pseudopotentials (cut at Cα)[31]. The

wavefunction was expanded in a plane wave basis set up to an energy cutoff of 75 Ry. Constant temperature simulations (300 K) were performed using a Nose-Hoover thermostat[32] with a coupling frequency of 2000 cm^{-1}.

A representative frame with the lowest energy from the last 100 ns MD simulations of the apo form of the designed mutant was selected to build each metal bound structure. Thereafter, an equilibrated structure of the classical MD simulation of the zinc (II) bound protein was used as starting point to model monomer and dimer forms. Additionally, crystal structures were used to initiate the QM/MM simulations of the monomer and the head-to-tail dimer. Initially, the systems were equilibrated keeping the QM region fixed. Subsequently, a short unrestrained molecular dynamics was carried out. Then, production runs were conducted. Metal coordination distances were constrained in the beginning of the QM/MM simulations and gradually released within 4 ps.

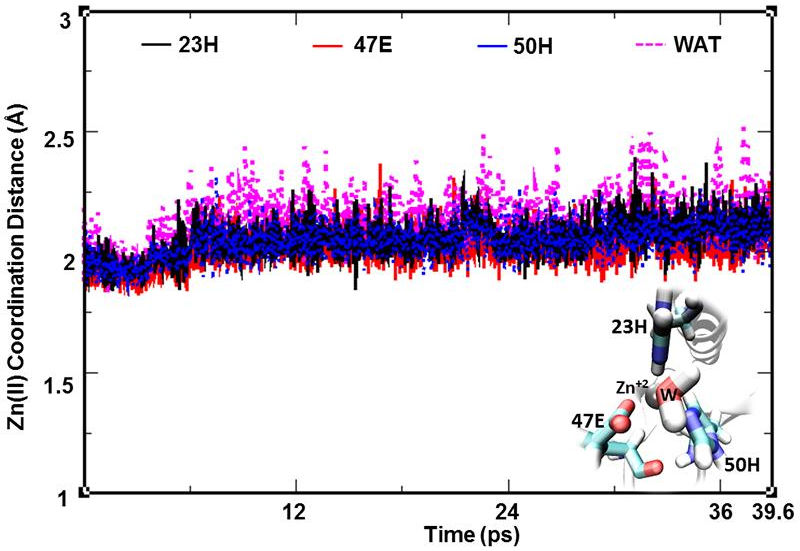

Figure S4. Illustration of zinc coordination distances during QM/MM molecular dynamics simulation of monomeric metallo-GB1 in solution.

Table S3. Summary of the simulations performed on the monomer and dimer form of GB1.

Entry	Duration of MD Simulation	Duration of QM/MM MD Simulation
WT-GB1	300 ns	-
6B1 (monomer)	500 ns	-
$6B1^{23H/47E/50H}$ (monomer)	500 ns	-
Zn-$6B1^{23H/47E/50H}$ (monomer)	10 ns	34.8 (x2)
Head-to-tail (with Zn^{+2})	250 ns (x4)	25 ps (x2)

6B1 refers to a mutant, which contains six wild-type thermostabilizing mutations (3F/15V/16I/18I/25E/29F) and additionally T2Q mutation.

The Ptraj module[33] of AMBER 14 and the VMD software[34] were used for analysis and rendering of the three-dimensional structures, respectively.

References

(1) Cunningham, T. F.; McGoff, M. S.; Sengupta, I.; Jaroniec, C. P.; Horne, W. S.; Saxena, S. *Biochemistry* **2012**, *51*, 6350–6359.

(2) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. *Journal of Applied Crystallography* **2007**, *40*, 658–674.

(3) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D: Biological Crystallography*, 1997, *53*, 240–255.

(4) Emsley, P.; Cowtan, K. *Acta Crystallographica Section D: Biological Crystallography* **2004**, *60*, 2126–2132.

(5) Joosten, R. P.; Joosten, K.; Cohen, S. X.; Vriend, G.; Perrakis, A. *Bioinformatics* **2011**, *27*, 3392–3398.

(6) Niklasson, M.; Andresen, C.; Helander, S.; Roth, M. G. L.; Zimdahl Kahlin, A.; Lindqvist Appell, M.; Lundström, P. Robust and convenient analysis of protein thermal and chemical stability. *Protein Science*, 2015, *24*, 2055–2062.

(7) Brown, P. H.; Schuck, P. *Computer Physics Communications* **2008**, *178*, 105–120.

(8) Brautigam, C. A. In *Methods in Enzymology*; 2015; Vol. 562, pp. 109–133.

(9) Behnel, S.; Bradshaw, R.; Citro, C.; Dalcin, L.; Seljebotn, D. S.; Smith, K. *Computing in Science and Engineering* **2011**, *13*, 31–39.

(10) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *Journal of Cheminformatics* **2011**, *3*.

(11) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. *Proteins: Structure, Function and Genetics* **2000**, *40*, 389–408.

(12) 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 Genetics*, 2006, *65*, 712–725.

(13) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *Journal of the American Chemical Society* **1990**, *112*, 6127–6129.

(14) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins: Structure, Function and Genetics* **2004**, *55*, 383–394.

(15) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *Journal of Computational Chemistry* **2004**, *25*, 265–284.

(16) Deb, K.; Agrawal, R. B. *Complex Systems* **1994**, *9*, 1–34.

(17) Deb, K.; Kumar, A. *Complex Systems* **1995**, *9*, 431–454.

(18) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration [16]. *Journal of the American Chemical Society*, 1999, *121*, 2337–2338.

(19) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic Acids Research* **2005**, *33*.

(20) Lzaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. *Journal of Chemical Physics* **2001**, *114*, 2090–2098.

(21) Darden, T.; York, D.; Pedersen, L. *The Journal of Chemical Physics* **1993**, *98*, 10089.

(22) Petersen, H. G. *The Journal of Chemical Physics* **1995**, *103*, 3668–3679.

(23) Miyamoto, S.; Kollman, P. A. *Journal of Computational Chemistry* **1992**, *13*, 952–962.

(24) Hu, L.; Ryde, U. *Journal of Chemical Theory and Computation* **2011**, *7*, 2452–2463.

(25) Li, P.; Roberts, B. P.; Chakravorty, D. K.; Merz, K. M. *Journal of Chemical Theory and Computation* **2013**, *9*, 2733–2748.

(26) Li, P.; Song, L. F.; Merz, K. M. *Journal of Physical Chemistry B* **2015**, *119*, 883–895.

(27) Hutter, J.; Iannuzzi, M. *Zeitschrift fur Kristallographie* **2005**, *220*, 549–551.

(28) Becke, A. D. *Physical Review A* **1988**, *38*, 3098–3100.

(29) Troullier, N.; Martins, J. L. *Physical Review B* **1991**, *43*, 8861–8869.

(30) Magistrato, A.; DeGrado, W. F.; Laio, A.; Rothlisberger, U.; VandeVondele, J.; Klein, M. L. *Journal of Physical Chemistry B* **2003**, *107*, 4182–4188.

(31) Von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Journal of Chemical Physics* **2005**, *122*, 1–7.

(32) Nosé, S. *The Journal of Chemical Physics* **1984**, *81*, 511.

(33) Roe, D. R.; Cheatham, T. E. *Journal of Chemical Theory and Computation* **2013**, *9*, 3084–3095.

(34) Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38.