

Interpolation schemes for peptide rearrangements

Marianne S. Bauer, Birgit Strodel, Szilard N. Fejer, Elena F. Koslover, and David J. Wales

Citation: [The Journal of Chemical Physics](#) **132**, 054101 (2010); doi: 10.1063/1.3273617

View online: <https://doi.org/10.1063/1.3273617>

View Table of Contents: <http://aip.scitation.org/toc/jcp/132/5>

Published by the [American Institute of Physics](#)

Articles you may be interested in

[Geometry optimization for peptides and proteins: Comparison of Cartesian and internal coordinates](#)

[The Journal of Chemical Physics](#) **127**, 234105 (2007); 10.1063/1.2807227

[Comparison of double-ended transition state search methods](#)

[The Journal of Chemical Physics](#) **127**, 134102 (2007); 10.1063/1.2767621

[Many-body computer simulation models for alkali-metal ion-water interactions](#)

[Scilight](#) **2017**, 050002 (2017); 10.1063/1.4995273

[Unitary Symmetry of Oscillators and the Talmi Transformation](#)

[Journal of Mathematical Physics](#) **6**, 142 (1965); 10.1063/1.1704252

[Special Functions of Mathematical Physics from the Viewpoint of Lie Algebra](#)

[Journal of Mathematical Physics](#) **7**, 447 (1966); 10.1063/1.1704953

[On the accuracy of the MB-pol many-body potential for water: Interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice](#)

[The Journal of Chemical Physics](#) **145**, 194504 (2016); 10.1063/1.4967719

PHYSICS TODAY

WHITEPAPERS

ADVANCED LIGHT CURE ADHESIVES

Take a closer look at what these environmentally friendly adhesive systems can do

READ NOW

PRESENTED BY
 **MASTERBOND**
ADHESIVES | SEALANTS | COATINGS

Interpolation schemes for peptide rearrangements

Marianne S. Bauer,¹ Birgit Strodel,² Szilard N. Fejer,¹ Elena F. Koslover,³ and David J. Wales^{1,a)}

¹University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, United Kingdom

²Institut für Strukturbiologie und Biophysik, Strukturbiochemie (ISB-3), Forschungszentrum Jülich, Jülich 52425, Germany

³Biophysics Program, Stanford University, Stanford, California 94305, USA

(Received 10 September 2009; accepted 23 November 2009; published online 1 February 2010)

A variety of methods (in total seven) comprising different combinations of internal and Cartesian coordinates are tested for interpolation and alignment in connection attempts for polypeptide rearrangements. We consider Cartesian coordinates, the internal coordinates used in CHARMM, and natural internal coordinates, each of which has been interfaced to the OPTIM code and compared with the corresponding results for united-atom force fields. We show that aligning the methylene hydrogens to preserve the sign of a local dihedral angle, rather than minimizing a distance metric, provides significant improvements with respect to connection times and failures. We also demonstrate the superiority of natural coordinate methods in conjunction with internal alignment. Checking the potential energy of the interpolated structures can act as a criterion for the choice of the interpolation coordinate system, which reduces failures and connection times significantly.

© 2010 American Institute of Physics. [doi:10.1063/1.3273617]

I. INTRODUCTION

Building up a database of local minima and the transition states that connect them enables us to describe dynamics using a kinetic transition network.^{1–3} In particular, the discrete path sampling (DPS) approach^{4–8} subsumes several different schemes for growing a stationary point database systematically to represent the kinetic properties of interest. Transition networks can also be constructed using molecular dynamics (MD) for states that interconvert on MD accessible time scales.^{9–11} Locating transition states on the potential energy surface (PES) plays a key role in extracting the rate constants associated with connections in the transition network, and in characterizing rearrangement mechanisms using geometry optimization techniques. Here we employ the geometrical definition of a transition state, namely, a stationary point that has a Hessian matrix with precisely one negative eigenvalue.¹²

Most of the CPU time used in approaches based on geometry optimization is spent in characterizing the transition states. The efficiency of these calculations can be improved in double-ended searches by the initial alignment of the two end points and the choice of coordinates employed for the initial interpolation between them. The OPTIM (Ref. 13) program includes several double-ended transition state search methods, such as the evolving string,¹⁴ growing string,^{15,16} and doubly nudged¹⁷ elastic band^{18–23} (DNEB) approaches. The choice of coordinate systems for interpolation ranges from Cartesian, through general Z-matrix²⁴ and Z-matrix derived coordinate systems, to natural internal coordinates.^{25–27}

The principal advantage of Cartesian coordinates is their

simplicity. However, interpolation in Cartesian space may disrupt the covalent bonding network of the protein, especially if the two end points are sufficiently different from each other. For example, internal rotation of a methyl group or a ring is easily described in internal coordinates, while the interpolation using Cartesian coordinates involves atom clashes or unphysical structures. In many cases we can still optimize such unphysical structures, to yield a sensible transition state candidate. However, this optimization procedure is likely to require significant CPU time and often leads to pathways that are far from the kinetically relevant ones, with high energy barriers.

Problems have already been reported for transition state optimizations in natural internal coordinates, especially for large proteins.^{24,28} Hence, in the present work, we only used internal coordinates for generating the initial transition state candidates, employing Cartesian coordinates for the actual geometry optimization phase. The local minima corresponding to each transition state are also characterized using Cartesian coordinates. We present a detailed comparison of different interpolation and alignment schemes for the initial alignment and interpolation between end points, and test them for two peptides.

To represent the peptides we have chosen the all-atom force field AMBER ff03 (Ref. 29) and the united-atom force field CHARMM19.³⁰ CHARMM (Ref. 31) itself offers an option of using internal coordinates that are similar to Z-matrix redundant coordinates.³² We implemented the same coordinates for our interface of OPTIM with the AMBER9 package. Koslover and Wales²⁸ previously employed natural internal coordinates in OPTIM for use with CHARMM19, and we extended this scheme to the all-atom force fields in AMBER in the present work.

Interpolation with internal coordinates depends strongly

^{a)}Author to whom correspondence should be addressed. Electronic mail: dw34@cam.ac.uk.

on the initial alignment of the two end points. Alignment here refers to minimizing a specified distance metric (usually Euclidean) with respect to both the coordinates of the peptides and the exchange of permutable atoms. In addition to the Euclidean distance metric we also considered alignment according to dihedral angles, and for AMBER we implemented a method that prohibits the exchange of the methylene hydrogen atoms in order to avoid unphysical geometries.

As test systems we have chosen the amyloidogenic GN-NQQNY peptide^{33,34} (NH_3^+ -Gly-Asn-Asn-Gln-Gln-Asn-Tyr-COO⁻) and the tryptophan zipper peptide, trpzip1^{35,36} (NH_3 -Ser-Thr-Trp-Glu-Asn-Gly-Lys-Trp-Thr-Trp-Lys-CH₃). For these two systems we first generated a database of local minima and the transition states that connect them. We then employed the different alignment and interpolation methods to test whether minima with known connections could be reconnected and how much CPU time was required.

II. METHODS

A. Geometry optimization

If we wish to analyze pathways in a coarse-grained framework of local minima connected by transition states, then it is essential to have efficient tools for locating such stationary points. In the absence of branch points,^{8,37} each transition state links two minima via the two steepest-descent paths defined by the Hessian eigenvector corresponding to the unique negative Hessian eigenvalue. Hence we characterize such connections by a minimum-transition state-minimum triplet, and define a discrete path between two minima, min_A and min_B , in terms of overlapping triplets: $\text{min}_A\text{-ts}_1\text{-min}_1\text{-ts}_2\text{-min}_2\cdots\text{min}_B$. The number of steps in a discrete path is defined as the number of transition states.^{4,5}

All the minimization procedures in the present work employed a slightly modified version of the limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm,^{38,39} as implemented in the OPTIM code.¹³ An optimization was deemed to have converged when the root-mean-square gradient fell below 10^{-6} kcal mol⁻¹ Å⁻¹.

Double-ended searches for connections between specified local minima were performed using the DNEB algorithm. The “band” consists of a series of image structures, which are optimized simultaneously using the LBFGS approach until local maxima become well defined.^{17,40} These local maxima are then taken as candidates for further refinement using hybrid eigenvector-following (EF),^{41–43} where a selected direction is searched uphill in steps that alternate with LBFGS minimization projected onto the tangent space to prevent interference with the uphill step.

Complete discrete paths between distant minima are unlikely to be found in a single connection attempt. For a given pair of minima we therefore consider successive DNEB/hybrid EF searches up to a maximum number of cycles, terminating if a connected path is found. Unless the connection attempt succeeds in one cycle, there is therefore another local decision to be made about which pair of minima to consider next. This problem is addressed using the Dijkstra missing connection algorithm.^{44,45}

Employing an internal coordinate representation can reduce the number of steps required for convergence of geometry optimizations with anisotropic potentials.²⁸ However, there is also a significant overhead associated with the coordinate transformation, and in the applications discussed below we employed Cartesian coordinates for all the geometry optimizations, but considered different coordinate systems for the interpolation between minima. Interpolating in this way can help avoid stationary points with high energies or unphysical geometries that are sometimes supported by empirical biomolecular force fields.

B. Coordinates for interpolation

CHARMM internal coordinates. The standard CHARMM internal coordinates are a set of redundant coordinates, which are defined via four atoms, I, J, K, L, and the information of whether the dihedral angle described by these four atoms is an improper angle or not.³² The actual coordinates are then the two bond lengths I-J and K-L, two bond angles, and the dihedral angle I-J-K-L. For an improper dihedral angle, K is the central atom and the two bond angles are taken as I-K-J and J-K-L, instead of I-J-K and J-K-L for an ordinary dihedral angle. For the AMBER all-atom force fields we used the same definition of the internal coordinates employed in the CHARM22 force field. We implemented these coordinates for all amino acids, C- and N-termini, and the ACE and NME capping groups.³²

We only interpolate proper dihedral angles, as we found that in the rare cases where bond lengths and angles differ significantly between the two end points, the bond lengths and angles of the interpolated structure after optimization usually correspond closely to the values at one of the end points. Dihedral angles are only interpolated if they are Ramachandran angles or twistable sidechain dihedrals. In general, we consider dihedral angles defining ring structures as nontwistable in order to maintain the ring geometry during the interpolation procedure. Thus, none of the coordinates in proline, and none of the coordinates defining the ring structure in histidine, phenylalanine, tryptophan, and tyrosine are interpolated. Planar structures are also categorized as nontwistable, and the value for the interpolated geometry is taken from the closer end point. The relevant dihedral angles here occur in the guanidinium group in arginine, the carboxyl groups in aspartic and glutamic acid, and the C-terminus and the amide groups in asparagine and glutamine. None of the improper (including chiral) dihedral angles are initially interpolated, however, we emphasize that all the coordinates are optimized in the DNEB phase and in the subsequent geometry optimizations.

To optimize the images in Cartesian coordinates²⁸ the interpolated structures have to be transformed back to a Cartesian representation. We employ an existing CHARMM routine for this purpose, and we replicated this procedure for the OPTIM interface to AMBER9.

Natural internal coordinates. Problems with the above internal coordinates arise when a protein is associated with groups that do not consist of amino acids and therefore have no predetermined set of parametrized internal coordinates.

For such cases, these coordinates have to be chosen and programmed directly. Pulay and co-workers introduced a set of “natural” internal coordinates,^{25–27} which have previously been implemented in OPTIM (Ref. 28) for united-atom force fields such as CHARMM19. The advantage of these natural internal coordinates over other options for internal coordinates that can be chosen automatically, such as Z-matrix or primitive internals, is that there is no redundancy present.^{26,46} These natural internal coordinates are derived from the symmetrized internal coordinates used in vibrational spectroscopy. They consist of linear combinations of bond lengths, angles, and dihedral angles and are local in character, which means that the Cartesian coordinates of only a few atoms are combined to create an internal coordinate. In general, bond lengths are taken as individual coordinates, while the angular coordinates are composed of linear combinations depending on the symmetry of a particular atomic center.²⁵ These symmetries, or rather pseudosymmetries, depend on the coordination number and geometry of the atomic center. For rings, an idealized D_{nh} symmetry is assumed, even though the actual symmetry may be lower.²⁵ Terminal atoms are considered as equivalent, while nonterminal atoms are distinguished according to whether or not they belong to a ring.²⁵ When applying this approach to proteins it was found that there exist up to eight types of atomic centers with different symmetries. Internal coordinates for each type can be used as building blocks for a full set of natural internals (NIs), in addition to bond lengths and linear dihedrals.⁴⁷

The most computationally expensive step associated with using NIs is the transformation from Cartesian to internal coordinates and back. Our current implementation uses the linear scaling algorithm of Németh *et al.*⁴⁶ This algorithm employs a shifted pseudoinverse of the rectangular transformation matrix between Cartesian and internal coordinates⁴⁸ and exploits the sparsity of these matrices. Further details of this algorithm and its implementation can be found in Refs. 28 and 46.

C. Structural alignment

For the interpolation employed in the DNEB algorithm the initial alignment of the end points is crucial. Optimal alignment means that the distance between the two structures is minimized according to a specific metric. Several groups suggested methods and algorithms to perform structural alignment.^{49–51} Here we utilized an algorithm that could be conveniently implemented in the OPTIM code.

Structural alignment has to be performed with respect to both overall translation and rotation, and with respect to permutational isomers. Such permutational isomers have the same energy due to the symmetry of the Hamiltonian,⁵² but differ with respect to the labels assigned to atoms of the same element. For example, a methyl group in structure A may have the indices of the hydrogens of the order of H1, H2, H3, whereas in structure B they are H3, H1, H2. Such permutational isomers have to be considered in conjunction with the orientational alignment.⁵³

Alignment according to Cartesian coordinates means that a Cartesian distance metric, namely the distance between

the coordinates of the starting ($\mathbf{r}_{\text{start}}$) and finishing structures ($\mathbf{r}_{\text{finish}}$), is minimized with respect to permutational isomerization; we denote this choice by $\min\{\sqrt{(\mathbf{r}_{\text{finish}} - \mathbf{r}_{\text{start}})^2}\}_{\text{perm}}$. The distance metric for alignment with natural internal coordinates is constructed from the n dihedrals in the whole structure, $\sqrt{\sum_{i=1}^n (\Phi_{\text{finish},i} - \Phi_{\text{start},i})^2}$, where $\Phi_{\text{start},i}$ and $\Phi_{\text{finish},i}$ refer to the values of dihedral angle i in the two end points.⁴⁷

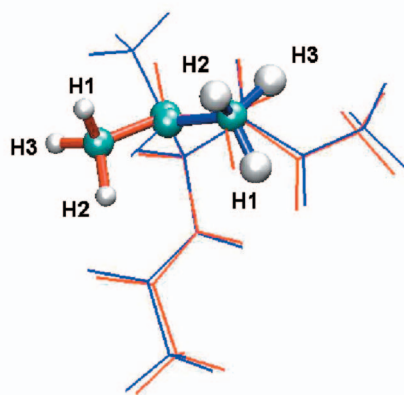
For methylene (CH_2) hydrogens, alignment with respect to the above metrics was often found to result in structures with hydrogen labels assigned in a way that requires an unphysical hydrogen exchange in the interpolation. This inappropriate alignment can be avoided by fixing the hydrogen labels so that a physically reasonable change in conformation is possible. We address this labeling procedure using the sign of the improper dihedral around a carbon atom C, namely, H1–H2–C–X, where X is the carbon substituent with the lower index in a fixed numbering scheme, and H1 and H2 are the two methylene hydrogens. To avoid unphysical swaps we fix the hydrogen labels in the two end points so that this improper dihedral has the same sign. Details of the tests used to compare alignment according to distance metrics and the fixed hydrogen labels can be found in the supplementary material.⁵⁴

Identification of the best permutational isomer in the alignment can greatly facilitate the path finding procedure, as the interpolation between the end points does not have to take into account additional degenerate rearrangements^{8,55} corresponding to permutational isomerization. Such rearrangements necessitate a superfluous transition state on the path, which, apart from decreasing the corresponding rate constant, also wastes CPU time. Figure 1 shows an example where two structures are practically the same except for the position of a methyl group. The optimal Cartesian alignment is shown in Fig. 1(a). Here two transition states are required to link the corresponding minima, one of which corresponds to internal rotation of the methyl group. A more appropriate alignment was achieved using internal coordinates, where this second transition state is not required [Fig. 1(b)].

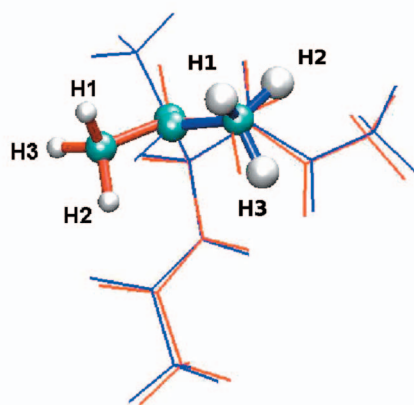
D. Test cases and parameters

Choice of methods. We tested seven methods, for which we now define a convenient shorthand notation. Four of these methods [backbone Cartesian, sidechains Cartesian (BCSC), backbone Cartesian, sidechains internal (BCSI), backbone internal, sidechains internal (BISI), and NONI] use Cartesian alignment. BCSC is the general approach using entirely Cartesian interpolation for generating the transition state candidates. BCSI and BISI employ CHARMM internal coordinates for the interpolation. This distinction between the different parts of the protein was introduced because the use of internal coordinates for backbone interpolation is likely to produce undesirable atom clashes.⁵⁶

Natural internal coordinates were used for the interpolation in the following four methods, in conjunction with internal coordinate alignment, except for method NONI (NO internal alignment), where natural internal coordinates are used with Cartesian alignment. In each case a user-determined number of images is generated in natural internal



(a) Cartesian alignment



(b) Internal alignment

FIG. 1. Alignment according to different coordinate systems. The rearrangement between the initial (red) and final (blue) end points involves mainly the motion of a single methyl group, highlighted as a ball and stick representation. When aligning according to a Cartesian distance metric, $\min\{\sqrt{(\mathbf{r}_{\text{finish}} - \mathbf{r}_{\text{start}})^2}\}_{\text{perm}}$, the blue H1 after the rotation of the methyl group would overlap with the red H3, the blue H3 with the red H2, and the blue H2 with the red H1. Hence an additional internal rotation of the methyl group would be required. When alignment according to the internal coordinate metric $\min\{\sqrt{\sum_{i=1}^n (\Phi_{\text{finish},i} - \Phi_{\text{start},i})^2}\}_{\text{perm}}$ is employed, the structures can interconvert via a single transition state.

coordinates and converted to a specified number of images in Cartesian coordinates for subsequent DNEB optimization. This conversion was achieved using a coordinate transformation, after which the Cartesian images generated were evenly distributed along the path using additional piecewise interpolation. NI is the general natural internal coordinate method.

The NIS (NI simple) method specifies that natural internal interpolation generates all the images required for the DNEB phase without respacing them in Cartesian coordinates to make the images equidistant. The NI choice (NIC) method introduces flexibility with respect to the coordinate system used depending on the energy of the images. This method uses natural internal coordinate interpolation followed by equispacing, as described for method NI, but also generates a set of images by linear Cartesian interpolation. The interpolation method that provides the lowest maximal energy for the images is then employed.⁴⁷

TABLE I. Summary of the different alignment and interpolation schemes.

Method	Alignment	Interpolation
BCSC	Cartesian	Cartesian for backbone and sidechains
BCSI	Cartesian	Cartesian for backbone, CHARMM internals for sidechains
BISI	Cartesian	CHARMM internal for backbone and sidechains
NONI	Cartesian	Natural internal
NI	Natural internal	Natural internal
NIS	Natural internal	Natural internal, using nonequispaced images
NIC	Natural internal	Natural internal or Cartesian, according to the lowest energy

For AMBER, all methods except ORIG employ alignment of CH₂ hydrogens to conserve the sign of a local dihedral angle and avoid unphysical exchange.

For the AMBER force field, we fixed the labels of the methylene CH₂ hydrogens in conjunction with each of the seven methods, unless otherwise specified. Results for method ORIG involving standard Cartesian interpolation and alignment according to a Cartesian distance metric are shown once to emphasize the improvement achieved by fixing these labels. A summary of all these methods can be found in Table I.

Systems. As test systems we have chosen the amyloidogenic GNNQQNY peptide^{33,34} and one of the tryptophan zipper (trpzip1) peptides.^{35,36} The free energy surface of the GNNQQNY monomer has previously been characterized using replica exchange MD (Ref. 57) and three stable conformations are shown in Fig. 2(b). From a replica exchange simulation for trpzip1, we extracted a stable folded structure and an unfolded structure, as shown in Fig. 2(a). Both peptides were represented by the united-atom force field CHARMM19 (Ref. 30) and the all-atom force field AMBER ff03.²⁹ To account for an aqueous environment, CHARMM19 was used together with the implicit solvation potential EEF1,⁵⁸ and for the AMBER simulations the generalized Born solvation model GB^{OBC} was employed.⁵⁹ Small modifications of the CHARMM force field and topologies as well as the AMBER topologies were made to ensure that sidechain rotamers have identical energies and geometries.^{17,53} All CHARMM calculations were performed with the c31a2 version, and for the AMBER calculations the AMBER9^{60–62} package was used.

Parameters. We first identified parameter choices required for OPTIM that provide a good overall performance for the two peptides. Parameters to be specified for DNEB include the image density, which describes how many images should be used per angstrom, the maximal number of images allowed for any given connection attempt, and the iteration density, which determines the maximal number of iterations allowed for optimizing the DNEB images before the transition state candidates for accurate refinement with hybrid EF^{41–43} are chosen.¹⁷ This maximal number of iterations is defined by iteration density multiplied by the number of images used for the connection attempt. Connections are

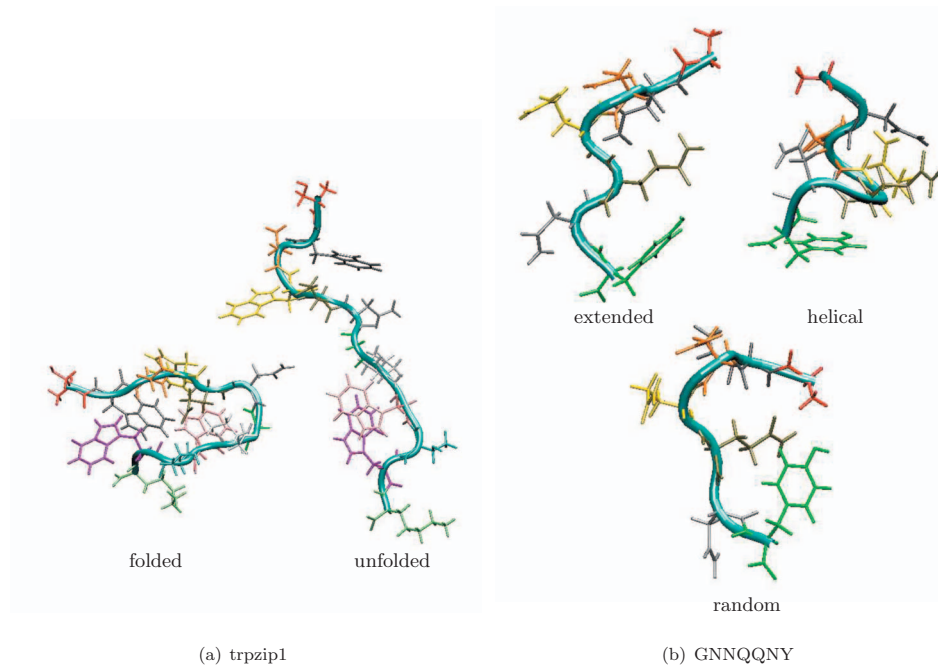


FIG. 2. Folded and unfolded conformations of trpzip1 and extended, helical, and random conformations of the GNNQQNY peptide. These structures were used as starting points for generating a database of minima and transition states.

usually found much faster for the CHARMM19 potential, due to the smaller number of atoms.

We tested different parameter sets for the DNEB calculations and the overall connection strategy. Two of these sets were investigated in detail and found to give the same general trends. For brevity we present results for only one set, which was optimized to minimize the number of failures, with the constraint that overall CPU time required should still be reasonable. We considered various image (1, 2, 3, 5, 10) and iteration density combinations (100, 80, 50, 30, 20).

The optimized parameter set employs an image density of 1.0 in a general connection run. Three attempts for the connection for every pair of minima are allowed, with an increasing number of images and iterations in each attempt. While we used the same parameters within the same subset of system and force field, e.g., GNNQQNY with CHARMM19, we allowed the parameters for other combinations to be different in order to minimize failures in all subsets. The parameters for the DNEB calculations differ in terms of the spring parameter, k , which keeps the images equispaced between the initial and final end points, and the number of images and iterations used for the first connection attempt.⁵⁴

Databases. To test the interpolation and alignment methods, we prepared a set of minima pairs for which connections had already been found in previous DNEB runs for both GNNQQNY and trpzip1. We could therefore check if each method was able to find a connection again, how fast this connection could be achieved, and whether there are differences in the pathways located in terms of the barrier heights and path lengths. For GNNQQNY, we started by finding an initial path between helical and random conformations, as well as between random and extended conformations. For trpzip1 we first generated a pathway leading from the unfolded to the folded structure. These initial pathways were optimized using a scheme where minima from the database are selected for connection attempts if they belong to groups

with similar free energies separated by a large free energy barrier.⁶³ We consider the ratio of the barrier height to the free energy difference as a measure of this frustration, and select local minima from such groups for connection attempts based upon their minimized Euclidean distance. The test databases contained 2983 and 5206 minima for GNNQQNY, and 10655 and 10942 minima for trpzip1, with AMBER and CHARMM, respectively.

For each system we generated a database of all possible combinations of minima that are connected via one, two, and three transition states (TS1, TS2, and TS3). The AMBER databases for GNNQQNY and trpzip1 contained 3764 and 8856 combinations for TS1, 9737 and 12 345 for TS2, and 15 004 and 22 365 for TS3, while the CHARMM databases contained 6682, 21 206 and 70 141 TS1, TS2, and TS3 possibilities for GNNQQNY and 10 430, 15 741, and 27 478 possibilities for trpzip1. We randomly chose 150 pairs of minima for each case and attempted to reconnect them using the various combinations of alignment and interpolation schemes described above.

III. RESULTS and DISCUSSION

A. Overall CPU time and barrier heights

Our results are shown in Table II, where we provide details of the number of failures, the CPU time required to find a connection, the number of transition states on the connected path (n_{ts}), and the maximum energy barrier (ΔE) on the path. The average CPU time excludes the F slowest runs for each combination of methods, F being the maximal number of failures for all methods within a set (TS1–TS3). For trpzip1, we additionally excluded the slowest five runs for every method, thus discarding eight runs in total, as a few connections took a disproportionately long time and distort the average. Hence, for the method with the maximum number of failures, the average CPU time for successful runs will be unchanged, but it will be smaller for all the other meth-

TABLE II. Results for GNNQQNY (left) and trpzip1 (right). For GNNQQNY, the average CPU time (s) required to find a connection is averaged over $150-F$ runs, where F is the maximum value of f , the number of failures in 150 connection attempts for a given scheme. n_{ts} denotes the number of transition states involved in the connection and ΔE is the average maximum barrier height in kcal mol⁻¹. For trpzip1, averages for the CPU time are taken over 142 runs, to avoid the average being dependent mostly on the slowest eight runs.

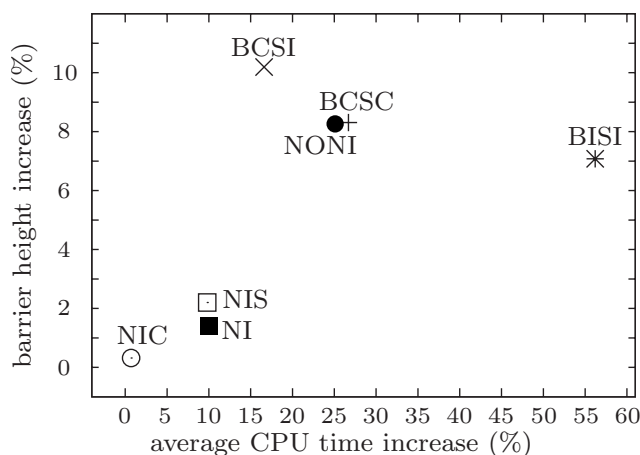
Method	GNNQQNY												trpzip1											
	TS1				TS2				TS3				TS1				TS2				TS3			
	f	Time	ΔE	n_{ts}	f	Time	ΔE	n_{ts}	f	Time	ΔE	n_{ts}	f	Time	ΔE	n_{ts}	f	Time	ΔE	n_{ts}	f	Time	ΔE	n_{ts}
AMBER																								
BCSC	0	90.2	2.14	1.3	0	211.3	4.25	2.8	0	328.0	4.70	3.8	0	528.5	2.71	1.2	2	1655.4	2.85	2.5	0	2512.3	3.39	3.6
BCSI	0	97.4	2.14	1.4	0	275.2	4.03	3.0	0	308.5	4.79	3.6	0	526.6	2.68	1.4	2	1775.9	2.83	2.6	0	3070.0	3.35	3.7
BISI	0	88.2	2.18	1.4	0	229.6	4.15	2.9	0	386.4	4.66	3.9	0	532.1	2.78	1.2	2	1366.1	2.78	2.5	0	2927.3	3.24	3.9
NI	0	88.3	2.04	1.4	0	277.4	3.74	2.9	0	324.0	3.83	3.7	1	452.3	2.71	1.3	1	1034.2	2.84	2.4	1	2308.9	3.15	3.5
NIS	0	78.0	2.02	1.4	1	288.0	3.75	2.9	1	339.9	3.78	3.6	0	409.6	2.70	1.2	0	979.8	2.87	2.3	1	2408.0	3.28	3.4
NIC	0	75.8	2.04	1.3	0	275.2	3.72	2.9	0	273.6	3.78	3.5	1	413.8	2.72	1.3	0	998.8	2.83	2.4	1	2315.5	3.15	3.5
NONI	0	138.0	2.24	1.5	0	272.2	4.30	2.9	0	361.4	4.47	3.7	2	576.0	2.72	1.3	2	1601.6	2.84	2.7	2	2869.8	3.25	3.7
ORIG	0	109.1	2.16	1.4	0	250.7	4.27	2.9	0	385.2	4.71	3.8	0	853.2	2.71	1.6	2	2577.2	2.89	3.0	1	3112.6	3.59	3.8
CHARMM																								
BCSC	0	14.3	2.81	1.1	0	25.8	4.03	2.5	0	53.0	5.30	3.9	0	50.3	1.81	1.3	0	159.0	2.55	2.9	1	301.4	3.25	3.9
BCSI	0	14.0	2.78	1.2	0	21.8	4.21	2.5	0	57.5	5.71	4.0	0	49.1	1.84	1.3	0	146.6	2.47	2.6	0	232.9	3.22	3.7
BISI	1	16.8	2.76	1.3	0	31.9	3.99	2.6	0	74.6	5.06	4.0	0	57.6	1.84	1.3	1	194.7	2.42	2.6	0	347.4	3.36	4.4
NI	0	11.5	2.67	1.1	0	24.5	3.80	2.4	0	50.8	4.45	3.8	0	45.1	1.84	1.3	2	139.1	2.32	2.6	1	201.6	3.23	3.4
NIS	0	11.1	2.67	1.1	0	24.6	3.85	2.5	0	51.8	4.44	3.8	0	47.2	1.85	1.3	1	133.7	2.36	2.7	1	210.5	3.19	3.4
NIC	0	11.3	2.69	1.1	0	24.0	3.75	2.4	0	42.4	4.28	3.6	0	45.2	1.79	1.3	0	141.1	2.34	2.7	1	204.8	3.14	3.4
NONI	0	13.5	2.75	1.2	0	28.7	4.19	2.6	0	55.8	5.54	4.0	0	49.4	1.86	1.3	2	172.7	2.31	2.6	1	243.7	3.19	3.6

ods, as some slow runs that achieve connections are excluded. This presentation of the results was chosen because we expect that the methods that succeed for end points where other methods fail will require more CPU time for these difficult connections. Looking at the average over successful connections only would produce lower values for methods that fail for the more difficult connections.

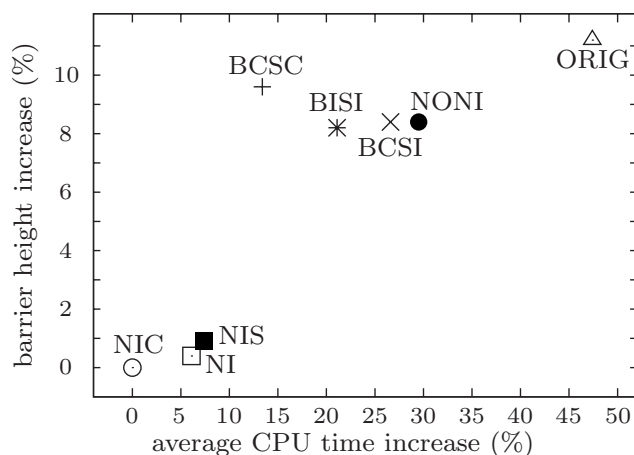
Plots of the average increase in barrier height versus average CPU time increase with respect to the best method are shown in Fig. 3. Here the fastest CPU time and lowest barrier are mapped in the lower left corner of the figure,

while the slowest CPU time and highest barrier appear in the top right corner. We averaged over the three runs for each peptide with every interpolation scheme and calculated the CPU time and maximum barrier height increase with respect to the best value (usually provided by the NIC scheme), before averaging over both peptides. The efficiency of all three natural internal coordinate methods in terms of barrier heights and CPU time is clearly visible from Fig. 3.

Figure 4 shows the CPU time improvement averaged over both peptides relative to the standard Cartesian interpolation method. The results again illustrate the efficiency of



(a) CHARMM



(b) AMBER

FIG. 3. Average increase in maximum barrier height (%) is plotted against average CPU time increase (%) with respect to the best method for (a) CHARMM and (b) AMBER. The actual CPU times are shown in Table II. These times were averaged over all TSx sets for each peptide and potential, and the percent increase with respect to the best value was averaged over the two peptide systems. One additional data point is included here for AMBER, namely, the ORIGINAL Cartesian interpolation with methylene hydrogen alignment according to a Cartesian distance metric (i.e., the sign of the local dihedral angle was not necessarily conserved).

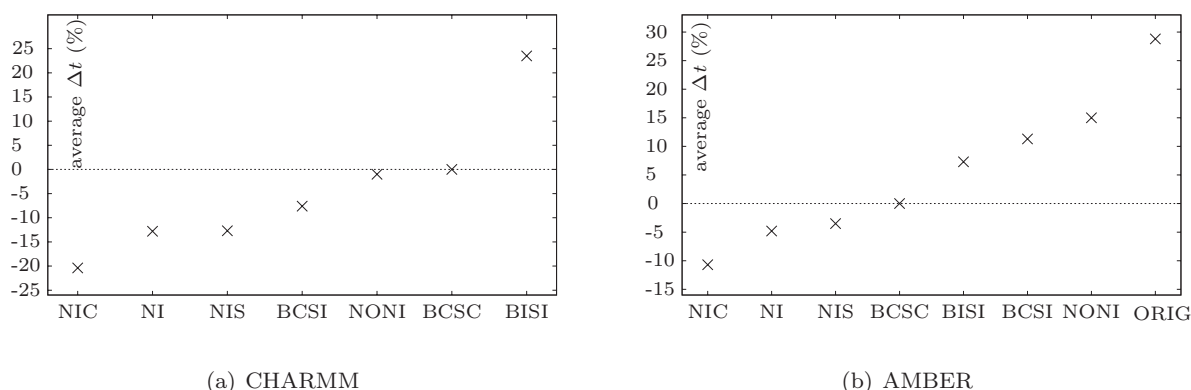


FIG. 4. Average improvement of CPU time in percent with respect to Cartesian coordinate scheme BCSC, plotted for each method with (a) CHARMM and (b) AMBER. (Here negative percentages indicate an improvement.) Natural internal interpolation is the fastest, especially the combination NIC, while CHARMM internal backbone interpolation BISI is the slowest. Note the poorer performance of the former default method ORIG (standard Cartesian interpolation BCSC where CH_2 hydrogens are permuted to minimize a Euclidean distance metric). BCSI and NONI exhibit intermediate performance, as does BCSC, which tends to outperform these methods on average, but does badly in a few cases.

natural internal interpolation. The NIC, NI, and NIS combinations are the methods with the fastest connections, while the failures shown in Table II are probably acceptable. Natural internal coordinate interpolation performs better when the rest of the protein is aligned according to an internal distance metric (NI, NIS, NIC) rather than a Cartesian metric (NONI). For AMBER the interpolation method BCSC also performed well. The results obtained when CH_2 hydrogens are permuted to minimize a distance metric (ORIG) are shown for comparison in Figs. 3(b) and 4(b). When the CH_2 hydrogens are fixed to conserve the sign of the local dihedral angle, all methods show a decrease in CPU time of about 40%.⁵⁴

B. Transition state convergence

Table III shows the average CPU time for convergence

TABLE III. Average CPU time (s) taken for transition state refinement with hybrid EF for trpzip1 in all 150 runs, for each test set (TS1–TS3). We ordered the methods according to the shortest convergence time averaged over all three sets, for each force field separately. Comparing the interpolation in natural internal coordinates, the NIC combination produces the best transition state candidates on average, while the NIS combination is best for TS3. We note the good performance of BISI for AMBER.

	TS1	TS2	TS3
AMBER			
BISI	103.26	103.03	116.43
NIC	108.28	97.36	121.35
NIS	110.56	102.37	115.53
BCSC	105.11	105.83	118.34
BCSI	96.70	115.13	124.68
NONI	110.73	102.76	123.36
NI	114.87	121.66	121.31
ORIG	119.29	136.97	129.38
CHARMM			
NIC	7.87	10.80	10.06
NIS	8.99	11.27	9.54
BCSC	8.79	9.92	11.56
BCSI	9.25	10.81	10.64
BISI	10.55	12.08	11.92
NI	8.18	14.89	10.09
NONI	9.63	14.03	10.38

of all the transition states for trpzip1. Transition state convergence times here refer to the CPU time spent in the hybrid EF procedure after a transition state candidate is obtained from the DNEB phase. The times for GNNQQNY showed a similar trend. However, because the differences in CPU time were so small for GNNQQNY, trpzip1 probably provides a better idea of likely performance.

A detailed analysis of Table III and the number of failures (Table II) shows that an improved transition state convergence time is usually associated with fewer failures. This observation also implies that the interpolation in natural coordinates should do well for connections between more distant end points, i.e., rearrangements for larger polypeptides. Initial tests showed this was indeed the case.

The NIC combination is generally the best. High energy natural internal coordinate images are avoided by switching to Cartesian interpolation if this scheme provides lower energies. Case studies have shown that in some complicated cases involving two ring rotations the NI, NONI, and NIS schemes proposed a pathway involving one ring moving *through* the other. The NIC scheme is able to avoid such unphysical interpolations, as the image energy is calculated. This ability to probe the quality of the pathway is probably responsible for NIC having relatively few failures.

We note that the transition state convergence times achieved following BISI interpolation were very fast, especially for AMBER. This good performance is not reflected in the total connection time, as the number of transition states involved per path was considerably higher than for the other methods. BISI in general tends to find connections involving more steps than the other methods (see Table II), but often has lower overall barriers. For the BCSC, BCSI and BISI methods transition state convergence times increase approximately linearly from TS1 to TS3. This increase is due to the fact that the DNEB images are less well converged for longer paths, which suggests that for longer connections these three methods will be relatively slow. The NIS interpolation method did not exhibit such an increase, implying that this procedure, or the NIC procedure with an appropriate number of internal images, should manage to retain good behavior for longer connections.

IV. CONCLUSIONS

We tested seven different combinations of internal and Cartesian coordinate schemes for interpolation and end point alignment to analyze rearrangement mechanisms of two polypeptides. We find that the alignment can influence the interpolation pathway significantly. Each interpolation scheme was used to generate starting images for subsequent refinement by the DNEB scheme.¹⁷ Local maxima in the resulting DNEB profile were then refined using hybrid EF^{41–43} and complete pathways were achieved by further connection attempts using the Dijkstra missing connection scheme.⁴⁵ All the developments described in this report have been implemented within the OPTIM program,¹³ available for use under the Gnu General Public License.

For all-atom force fields, such as AMBER, fixing the methylene hydrogens labels in order to avoid unphysical hydrogen exchanges provides significant improvements. Employing natural internal coordinates during the interpolation also increases efficiency compared to the previous Cartesian scheme. The CHARMM internal coordinate interpolation procedures did not perform as well, and it is better to use CHARMM internal coordinates for sidechain interpolation only, and not for the backbone.

The three best methods were normally the ones employing natural internal coordinates for both interpolation and alignment. In general we found that the most robust approach is the procedure where both natural internal coordinate and Cartesian coordinate interpolation are considered, and the set of images with the lower maximal energy is subsequently refined to locate transition state candidates. Further improvements can be expected from a method that links these two approaches in an optimal fashion.

The good performance for interpolation in natural internal coordinates was observed for both CHARMM19 and AMBER, which we chose to represent united-atom and all-atom force fields. The same trends were found for both of the test peptides for a variety of different parameter sets in both the DNEB and hybrid EF phases of the calculation. The success of different parameter sets suggests that the best schemes should also work efficiently for different systems. We conclude that the construction of kinetic transition networks for proteins and peptides using geometry optimization should be significantly more efficient when natural internal coordinates are used for interpolation.

ACKNOWLEDGMENTS

M.B. gratefully acknowledges the Gates Cambridge Trust for financial support.

¹F. Noé and S. Fischer, *Curr. Opin. Struct. Biol.* **18**, 154 (2008).

²D. Prada-Gracia, J. Gómez-Gardenes, P. Echenique, and F. Fernando, *PLOS Comput. Biol.* **5**, e1000415 (2009).

³R. E. Kunz and R. S. Berry, *J. Chem. Phys.* **103**, 1904 (1995).

⁴D. J. Wales, *Mol. Phys.* **100**, 3285 (2002).

⁵D. J. Wales, *Mol. Phys.* **102**, 891 (2004).

⁶D. J. Wales and J. P. K. Doye, *J. Chem. Phys.* **119**, 12409 (2003).

⁷D. Wales, *Int. Rev. Phys. Chem.* **25**, 237 (2006).

⁸D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).

⁹V. S. Pande and D. S. Rokhsar, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1273

(1999).

¹⁰W. Swope, J. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. Fitch, R. Germain, A. Rayshubski, T. Ward, Y. Zhestkov, and R. Zhou, *J. Phys. Chem. B* **108**, 6582 (2004).

¹¹S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14766 (2004).

¹²J. N. Murrell and K. J. Laidler, *Trans. Faraday Soc.* **64**, 371 (1968).

¹³D. J. Wales, OPTIM, a program for optimizing geometries and calculating pathways, <http://www-wales.ch.cam.ac.uk/software.html>, 2009.

¹⁴W. E. W. Ren, and E. Vanden-Eijnden, *Phys. Rev. B* **66**, 052301 (2002).

¹⁵B. Peters, A. T. Bell, and A. Chakraborty, *J. Chem. Phys.* **121**, 4453 (2004).

¹⁶E. F. Koslover and D. J. Wales, *J. Chem. Phys.* **127**, 134102 (2007).

¹⁷S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.* **120**, 2082 (2004).

¹⁸R. Elber and M. Karplus, *Chem. Phys. Lett.* **139**, 375 (1987).

¹⁹R. Czerminski and R. Elber, *J. Chem. Phys.* **92**, 5580 (1990).

²⁰G. Henkelman and H. Jónsson, *J. Chem. Phys.* **111**, 7010 (1999).

²¹G. Henkelman, B. P. Uberuaga, and H. Jónsson, *J. Chem. Phys.* **113**, 9901 (2000).

²²G. Henkelman and H. Jónsson, *J. Chem. Phys.* **113**, 9978 (2000).

²³G. Henkelman and H. Jónsson, *J. Chem. Phys.* **115**, 9657 (2001).

²⁴J. Baker and F. Chan, *J. Comput. Chem.* **17**, 888 (1996).

²⁵P. Pulay, G. Foragasi, F. Pang, and J. E. Boggs, *J. Am. Chem. Soc.* **101**, 2550 (1979).

²⁶P. Pulay and G. Fogarasi, *J. Chem. Phys.* **96**, 2856 (1992).

²⁷G. Fogarasi, X. Zhou, P. W. Taylor, and P. Pulay, *J. Am. Chem. Soc.* **114**, 8191 (1992).

²⁸E. F. Koslover and D. J. Wales, *J. Chem. Phys.* **127**, 234105 (2007).

²⁹Y. Duan, C. Wu, S. Chowdhury, M. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, and T. Lee, *J. Comput. Chem.* **24**, 1999 (2003).

³⁰E. Neria, S. Fischer, and M. Karplus, *J. Chem. Phys.* **105**, 1902 (1996).

³¹B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. S. Swaminathan, and M. Karplus, *J. Comput. Chem.* **4**, 187 (1983).

³²Intcor.doc, <http://www.lobos.nih.gov/charmm>, 2002.

³³M. Balbirnie, R. Grothe, and D. S. Eisenberg, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 2375 (2001).

³⁴R. Nelson, M. R. Sawaya, M. Balbirnie, A. O. Madsen, C. Riekel, R. Grothe, and D. Eisenberg, *Nature (London)* **435**, 773 (2005).

³⁵A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 5578 (2001).

³⁶C. D. Snow, L. Qiu, D. Du, F. Gai, S. J. Hagen, and V. S. Pande, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4077 (2004).

³⁷P. Valtazanos and K. Ruedenburg, *Theor. Chim. Acta* **69**, 281 (1986).

³⁸J. Nocedal, *Math. Comput.* **35**, 773 (1980).

³⁹D. Liu and J. Nocedal, *Math. Program.* **45**, 503 (1989).

⁴⁰D. Sheppard, R. Terrell, and G. Henkelman, *J. Chem. Phys.* **128**, 134106 (2008).

⁴¹L. J. Munro and D. J. Wales, *Phys. Rev. B* **59**, 3969 (1999).

⁴²Y. Kumeda, L. J. Munro, and D. J. Wales, *Chem. Phys. Lett.* **341**, 185 (2001).

⁴³Y. Kumeda and D. J. Wales, *Chem. Phys. Lett.* **374**, 125 (2003).

⁴⁴E. Dijkstra, *Numer. Math.* **1**, 269 (1959).

⁴⁵J. M. Carr, S. A. Trygubenko, and D. J. Wales, *J. Chem. Phys.* **122**, 234903 (2005).

⁴⁶K. Németh, O. Coulaud, G. Monard, and J. G. Ángyán, *J. Chem. Phys.* **113**, 5598 (2000).

⁴⁷E. F. Koslover, M.Phil. thesis, University of Cambridge, Cambridge, 2007.

⁴⁸E. B. Wilson, J. C. Decius, and P. C. Cross, *Molecular Vibrations* (Dover, New York, 1980).

⁴⁹I. N. Shindyalov and P. E. Bourne, *Protein Eng.* **11**, 739 (1998).

⁵⁰F. Teichert, U. Bastello, and M. Porto, *BMC Bioinf.* **8**, 425 (2007).

⁵¹W. R. Taylor and C. A. Orengo, *J. Mol. Biol.* **208**, 1 (1989).

⁵²P. R. Bunker and P. Jensen, *Molecular Symmetry and Spectroscopy*, 2nd ed. (NRC Research, Ottawa, 1998).

⁵³E. Małolepsza, B. Strodel, M. Khalili, S. Trygubenko, S. Fejer, and D. J. Wales, "Symmetrisation of the AMBER and CHARMM force fields," *J. Comput. Chem.* (in press).

⁵⁴See supplementary material at <http://dx.doi.org/10.1063/1.3273617> for details of the DNEB parameters and the tests used to compare alignment according to distance metrics and the fixed hydrogen labels.

⁵⁵R. E. Leone and P. v. R. Schleyer, *Angew. Chem., Int. Ed. Engl.* **9**, 860 (1970).

⁵⁶F. Noé, F. Ille, J. C. Smith, and S. Fischer, *Proteins: Struct., Funct.,*

- [Bioinf.](#) **59**, 534 (2005).
- ⁵⁷ B. Strodel, C. S. Whittleston, and D. J. Wales, [J. Am. Chem. Soc.](#) **129**, 16005 (2007).
- ⁵⁸ T. Lazaridis and M. Karplus, [Proteins: Struct., Funct., Genet.](#) **35**, 133 (1999).
- ⁵⁹ A. Onufriev, D. Bashford, and D. A. Case, [Proteins](#) **55**, 383 (2004).
- ⁶⁰ D. Case, T. Darden, T. C. Cheatham III, C. Simmerling, J. Wang, R. Duke, R. Luo, K. Merz, D. Pearlman, M. Crowley, R. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D. Mathews, C. Schafmeister, W. Ross and P. Kollman, AMBER9, University of California, 2006.
- ⁶¹ D. Pearlman, D. Case, J. Caldwell, W. Ross, T. C. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman, [Comput. Phys. Commun.](#) **91**, 1 (1995).
- ⁶² D. Case, T. C. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang, and R. Woods, [J. Comput. Chem.](#) **26**, 1668 (2005).
- ⁶³ J. M. Carr and D. J. Wales, [J. Phys. Chem. B](#) **112**, 8760 (2008).