Conformational analysis of replica exchange MD: Temperature-dependent Markov networks for FF amyloid peptides

Brajesh Narayan,^{1,2,a)} Colm Herbert,^{1,2,a)} Ye Yuan,^{1,2} Brian J. Rodriguez,^{1,3} Bernard R. Brooks,⁴ and Nicolae-Viorel Buchete^{1,2,b)}

¹*School of Physics, University College Dublin, Belfield, Dublin 4, Ireland*

²*Institute for Discovery, University College Dublin, Belfield, Dublin 4, Ireland*

³*Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland*

⁴*Laboratory of Computational Biology, NHLBI, National Institutes of Health, Bethesda, Maryland 20892, USA*

(Received 3 March 2018; accepted 25 May 2018; published online 22 June 2018)

Recent molecular modeling methods using Markovian descriptions of conformational states of biomolecular systems have led to powerful analysis frameworks that can accurately describe their complex dynamical behavior. In conjunction with enhanced sampling methods, such as replica exchange molecular dynamics (REMD), these frameworks allow the systematic and accurate extraction of transition probabilities between the corresponding states, in the case of Markov state models, and of statistically-optimized transition rates, in the case of the corresponding coarse master equations. However, applying automatically such methods to large molecular dynamics (MD) simulations, with explicit water molecules, remains limited both by the initial ability to identify good candidates for the underlying Markovian states and by the necessity to do so using good collective variables as reaction coordinates that allow the correct counting of inter-state transitions at various lag times. Here, we show that, in cases when representative molecular conformations can be identified for the corresponding Markovian states, and thus their corresponding collective evolution of atomic positions can be calculated along MD trajectories, one can use them to build a new type of simple collective variable, which can be particularly useful in both the correct state assignment and in the subsequent accurate counting of inter-state transition probabilities. In the case of the ubiquitously used root-mean-square deviation (RMSD) of atomic positions, we introduce the relative RMSD (RelRMSD) measure as a good reaction coordinate candidate. We apply this method to the analysis of REMD trajectories of amyloid-forming diphenylalanine (FF) peptides—a system with important nanotechnology and biomedical applications due to its self-assembling and piezoelectric properties—illustrating the use of RelRMSD in extracting its temperature-dependent intrinsic kinetics, without *a priori* assumptions on the functional form (e.g., Arrhenius or not) of the underlying conformational transition rates. The RelRMSD analysis enables as well a more objective assessment of the convergence of the REMD simulations. This type of collective variable may be generalized to other observables that could accurately capture conformational differences between the underlying Markov states (e.g., distance RMSD, the fraction of native contacts, etc.). <https://doi.org/10.1063/1.5027580>

I. INTRODUCTION

The sustained trend of growth in both system sizes and time scales accessible to atomistic molecular dynamics (MD) simulations continues. This is largely due to high performance computing (HPC) hardware performance improvements. These have continued their strong development with numerical processing speeds increasing in agreement with Moore's Law.¹ However, the recent progress in algorithmic development has been equally impressive. Current MD packages are now capable of microsecond simulations. Millisecond simulations become more feasible by using either standard HPC facilities or hardware capabilities developed

specifically for MD simulations² or by using distributed computing approaches such as Folding@Home.³

A particularly promising pathway in terms of new algorithms for HPC studies of biomolecular systems has been opened by the development of methods based on Markovian analysis of MD trajectories (for a recent review, see Ref. 4). This is particularly welcome, as the capacity of current HPC facilities remains severely limited in its ability to support the increasingly large number of standard atomistic MD simulation studies using explicit solvent molecules of even modestly sized molecular systems, due to difficulties in quantifying the intrinsic complexity of the conformational dynamics of even rather simple biomolecular systems. Additionally, the system sizes (e.g., hundreds of thousands to millions of atoms and beyond, when including explicit solvent molecules and ions) of simulations relevant to typical, more realistic biological time scales (milliseconds to seconds and larger) remain

^{a)}B. Narayan and C. Herbert contributed equally to this work.

^{b)}Author to whom correspondence should be addressed: buchete@ucd.ie

hard to reach in spite of the impressive recent HPC hardware advances.^{4–7}

Thus, the development of more efficient ways to extract the thermodynamic and kinetic properties from MD simulations of a biomolecular system remains an outstanding challenge, and a variety of modern enhanced sampling methods are increasingly replacing standard MD simulations in computational studies of proteins and other biomolecules.^{4,8,9} Replica-exchange MD (REMD)^{10–12} and simulated tempering^{13–15} are some of the most popular modern methods used to cross high energy barriers and to map the free energy landscape of biomolecular systems available in most MD simulation packages. In practice, REMD simulations provide accurate estimates of the populations of conformational states of a molecular system. However, extracting quantitative kinetic information from REMD trajectories regarding the transitions between the various conformational states is generally more challenging.^{16–23}

While REMD simulations were initially developed with the main goal of enhancing sampling (and thus the calculation of thermodynamic quantities) at ensembles corresponding to target physical parameters of interest (e.g., the lowest temperature), Markov-based analysis using master equations lead to accurate estimates of the corresponding kinetic properties (e.g., inter-state transition probabilities and rates) as well.¹⁶ For example, the direct transition counting (DTC) method for REMD data was recently developed and used^{20,24,25} in its simplest form as a method for calculating transition rates from REMD trajectories that is easier to implement and leads to similarly accurate temperature-dependent rates as compared to the alternative, more complex and indirect methods.^{16,23} A main advantage of Markov-based methods such as DTC is that they do not require prior knowledge of the functional form of the transition rates dependence on temperature (e.g., Arrhenius-like^{26,27}), or the need to use algorithms based on likelihood maximization (e.g., the maximum likelihood propagator-based, MLPB) and multi-dimensional optimization methods,^{16,23} or the complex statistical analysis of transition paths.²⁸ Nevertheless, such methods still require the possibility to identify and verify (e.g., using Chapman-Kolmogorov tests) good reaction coordinates that are typically low-dimensional collective variables, along which the projected dynamics captures accurately the intrinsic kinetic properties of the systems studied, over the entire range from fast to, especially, slow (and thus more difficult to sample) intrinsic relaxation times.^{8,9,29–34} Thus, applying automatically such methods to large molecular dynamics (MD) simulations, with explicit water molecules, remains limited both by the initial ability to identify good candidates for the underlying Markovian states and by the necessity to do so using good collective variables as reaction coordinates that allow the correct counting of inter-state transitions at various lag times.

Here, we show that in cases when representative molecular conformations (e.g., actual atomistic coordinates) can be identified for the corresponding Markovian states, and thus their corresponding collective evolution of atomic positions can be calculated along MD or REMD trajectories, one can use these representative conformations to build a new type of simple collective variable that can be particularly useful in both the

correct state assignment and in the subsequent accurate counting of inter-state transition probabilities. Focusing on the case of the ubiquitously used root-mean-square deviation (RMSD) of atomic positions,^{35,36} we introduce the relative RMSD (Rel-RMSD) measure as a good reaction coordinate candidate in Markov-based analysis. We apply this method to the analysis of REMD trajectories of amyloid-forming diphenylalanine (FF) peptides—a system with important nanotechnology and biomedical applications due to its self-assembling and piezoelectric properties³⁷—illustrating the use of Rel-RMSD in extracting its temperature-dependent intrinsic kinetics, without *a priori* assumptions on the functional form (e.g., Arrhenius or not) of the underlying intrinsic conformational transition rates, using a simple method for direct counting of interstate transitions from REMD data.^{23,24} Our RelRMSD analyses enable also a simple and objective assessment of the convergence of the REMD simulations.^{38–40} This type of collective variable may be generalized to other observables that are able to capture accurately conformational deviations between the underlying Markov states [e.g., distance RMSD (dRMSD),⁴¹ tree-based RMSD (T-RMSD),⁴² the fraction of native contacts,⁴³ etc.].

II. METHODS

A. REMD simulations

In this study, we use new atomistic REMD simulations of FF peptide monomers, using an implementation similar to the one described in Ref. 23 (though, in that case for different peptides, NNQQ, where N and Q are asparagine and glutamine residues, respectively), with the MD package Gromacs^{44,45} (GROningen MACHine for Chemical Simulations, version 5.1.4), using Langevin dynamics with a friction coefficient of 0.1 ps^{−1},⁴⁶ Parrinello-Rahman isotropic pressure coupling,⁴⁷ the particle-mesh Ewald implementation with a switching distance for the van der Waals interactions and non-bonded electrostatics of 8.5 Å and a cutoff distance of 12 Å, and with an integration time step of 1 fs. The simulations were performed in the NPT ensemble with the recent Chemistry at Harvard Macromolecular Mechanics (CHARMM)⁴⁸ 36 all-atom protein force field parameters (C36)⁴⁹ and using explicit TIP3P⁵⁰ water molecules. The simulation box contained 3379 atoms in total, including 1112 water molecules. To enhance the sampling, REMD is used with 12 replicas running at temperatures spaced according to an optimized protocol⁵¹ (the actual replica temperature are reported in Table S1 of the [supplementary material](#)) in the range of 310.00 K–385.10 K.²³

The FF system was built using VMD⁵² Molefacture Plugin's (version 1.3) protein builder tool, and it was minimized, heated, and equilibrated. The system was simulated using the Gromacs replica exchange script,⁵¹ with an average acceptance probability of ~20%. The integration time step was 2 fs and the velocity and positions were saved every 100 fs. After simulation, data were transformed into per temperature data using the Gromacs *demux* command, and the Gromacs *trajconv* command was used to only select every 250th frame (i.e., every 0.5 ps) for kinetic analysis.

The production simulations were done for 120 ns for each of the 12 replicas, giving a total REMD simulation time of

1.44 μ s, which was sufficient for achieving convergence of the relevant kinetic quantities (as shown in Fig. 8). As an additional test for convergence, we also investigated the “equal occupancy” of replicas at each temperature,³⁹ (as shown in Fig. S1 of the [supplementary material](#)) which is a useful method for assessing quickly the performance of parallel tempering simulations.^{39,40} Subsequently, kinetic data on the corresponding transition probabilities were calculated from the REMD trajectories as described below. Additional details about the parameters used for the FF REMD runs are reported in Table S1 of the [supplementary material](#).

B. Extracting transition probabilities and rates from REMD data

We analyze the FF temperature-dependent REMD data by following the workflow indicated in Fig. S2 of the [supplementary material](#) or Ref. 23, and the corresponding transition probabilities and rates were extracted and compared. Importantly, REMD data can be represented (redundantly) both as continuous “per replica” R-trajectories and as interrupted, “per temperature” T-trajectories. As presented in Ref. 23, the replica R-trajectories are continuous, even though they visit various temperatures (see Fig. 1 in Ref. 23), while the T-trajectories are actually interrupted (i.e., discontinue) at times when exchange attempts are accepted. While previous analysis methods have focused on T-trajectories, due to their well-defined temperatures, we start by using R-trajectories in order to take advantage of their time-continuity in the initial assignment of states.^{23,25} Moreover, we showed in Ref. 23 that a simple, analytical relation connects R-trajectories and T-trajectories, the propagators (i.e., conditional probabilities) for transition along R-trajectories being in effect the weighted geometric means of propagators for corresponding transitions in T-trajectories. This enables a useful direct application of kinetic analysis along R-trajectories, on which state assignment is easier due to their continuous nature, to inform the analysis of the discontinuous T-trajectories.

Following the procedure detailed in Ref. 25, here we assume that the conformational space of a system can be discretized into N distinct states that obey a master equation, which can be expressed in matrix notation as $\frac{d\mathbf{p}(t)}{dt} = \mathbf{K} \cdot \mathbf{p}(t)$, where $\mathbf{K}$ is the $N \times N$ rate matrix, $\mathbf{p}(t)$ is the time dependent the time-dependent column vector of probabilities with elements such that $p_n(t) > 0$, $n \in \{1, \dots, N\}$, with $p_n(t)$ being the probability of state n at time t , and the matrix element k_{nm} is the rate of transition from state m to state n .^{8,53–63} At equilibrium, we have $\mathbf{K}\mathbf{p}^o \equiv 0$, and the first right eigenvector of $\mathbf{K}$ (corresponding to the first eigenvalue, $\lambda_1 = 0$) is therefore given by $\mathbf{p}^o$ (i.e., the vector of equilibrium populations that has positive elements, $p_n^o > 0$, $n \in \{1, \dots, N\}$, and it is normalized according to the relation $\sum_{n=1}^N p_n^o = 1$).

As mentioned above, central to the DTC method^{20,24} for estimating rates (and to other Markov-based analysis methods) is the assignment of conformational states of the system. Here, the states are assigned by following each replica at different temperatures, using the transition based assignment (TBA) method described and used in previous studies.^{8,23,64} We first

introduced the TBA method for biomolecular simulations in Ref. 8, and it was also reviewed in detail in Ref. 64. This method relies on a good choice of initial reaction coordinates that allows a good discrimination between the different conformational Markov states. While these reaction coordinates need to be reasonably good, the subsequent state assignment step does not depend entirely on their absolute quality, as the TBA method adds more specific information from analyzing the transition paths (i.e., time sequence of transition events) to the state assignment process.⁶⁴ As described next, here we proposed a new measure, the relative RMSD (RelRMSD), as a useful choice for initiating the TBA analysis step.

C. RelRMSD as a reaction coordinate for TBA analysis

Markov-based analysis of biomolecular trajectories can generally be performed in multiple dimensions (e.g., by using principal components of molecular trajectories).^{19,65,66} Nevertheless, for biomolecules, the final analysis step needs to lead to a discussion of the underlying mechanism that can be understood and discussed in terms of relevant molecular conformations and changes associated with them. Focusing on the case of the ubiquitously used root-mean-square deviation (RMSD) of atomic positions,^{35,36} we introduce the relative RMSD (RelRMSD) measure as a good reaction coordinate candidate in the Markov-based analysis.

Here, we denote by $RMSD_i$, the RMSD value calculated along a trajectory after aligning it with a certain set of atomic coordinates (i.e., molecular conformation “ i ”), with $RMSD_j$ being, accordingly, the same RMSD measure calculated along the same trajectory but after aligning it with respect to a second relevant molecular conformation “ j .” By alignment, here we understand the optimal three-dimensional rigid body superposition of two (or more) sets of atomic coordinates. Thus, in molecular alignment, in order to compare the structures of biomolecular conformations, their atomic coordinates are translated and rotated with respect to each other (or with a reference) in order to minimize their RMSD value. For two given coordinate sets, the optimal solid body transformation can be found automatically using an algorithm based on quaternions, which identifies the optimal rotation-translation move that minimizes the RMSD between two sets of vectors³⁶ and which is equivalent to the solution found via the popular Kabsch algorithm.³⁵ Using two different molecular conformations “ i ” and “ j ” (e.g., corresponding to representative structures of two Markovian conformational states), we can calculate the corresponding $RMSD_i$ and $RMSD_j$ values after subsequent alignments of the same molecular trajectory and define the relative RMSD values for the two conformations as a time-dependent measure using the expression

$$\text{RelRMSD}_{ij}(t) = \frac{RMSD_i(t)}{RMSD_i(t) + RMSD_j(t)}. \quad (1)$$

This definition has the advantage that such a time dependent collective variable will always have values bounded between 0 and 1, adopting very small values, close to 0 (yet non-negative) when the trajectory visits coordinates close to the molecular conformation “ i ” (denoted here as the primary reference state), and close to 1 (yet always sub-unitary) when

the trajectory visits coordinates close to the molecular conformation “j.” We construct the RelRMSD measure with the intention to use it as a convenient collective variable that can be used to discriminate between conformational states and be applied efficiently in the subsequent TBA analysis steps. We note that, as illustrated in this study for the 3 main conformations of FF molecules, the RelRMSD measure can be useful in kinetic analysis of more than two states, by using a pairwise comparison of different RMSD values that are calculated with respect to the same primary reference state “i.” (See the discussion below corresponding to Figs. 4–7.)

D. Markovian transition probabilities as a test of REMD convergence

The Markovian transition probabilities extracted from REMD data after applying the TBA method to project the R- and T-trajectories to states and performing the kinetic analysis (see Fig. 8) can also serve as an ultimate test of the convergence of the REMD simulations performed. As mentioned above, initially we tested the REMD data convergence by investigated the “equal occupancy” rule of replicas at each temperature³⁹ (as shown in Fig. S1 of the [supplementary material](#)) which is fast and useful to assess the performance of parallel tempering simulations.^{39,40} However, it is important to note that we can also use blocks of REMD data to estimate statistical errors for the final Markovian transition probabilities, as errors of the means for each data block. The analysis of the errors in the extracted intrinsic kinetic parameters offers a reliable assessment of the convergence of molecular simulations used (here REMD, see discussion below and Table S2 of the [supplementary material](#)).

III. RESULTS AND DISCUSSION

In this study, following the analysis in Refs. 25 and 37, we use the newly introduced RelRMSD measure in

conjunction with the TBA method, for a more detailed characterisation of REMD trajectories of amyloid-forming diphenylalanine (FF) peptides—a system with important nanotechnology and biomedical applications due to its self-assembling and piezoelectric properties³⁷—illustrating the usefulness of RelRMSD in extracting its temperature-dependent intrinsic kinetics, without the need for making prior assumptions on the functional form (e.g., Arrhenius or not) of the underlying intrinsic conformational transition rates.^{23,24}

Previous studies suggested that the d_{TT} distance between the charged termini of FF peptides may be a useful collective variable as it is correlated to backbone stretching and thus with the magnitude of the total dipole moment of FF peptides under typical conditions with charged termini [for charged termini, NTER (NH₃⁺) and CTER (COO[−]) caps were applied].³⁷ The d_{TT} distance is defined and discussed in Fig. S2 of the [supplementary material](#). However, as shown in Fig. 1, while the distributions show good convergence in both cases,⁴⁰ a more detailed analysis shows that the second pick corresponds actually to two different “trans” conformational states, with representative (i.e., most populated) structures for both peaks being illustrated in the middle of Fig. 1. Note that the data in Fig. 1 correspond to probability distributions of the d_{TT} distance between the charged termini of FF peptides for each of the REMD replica R-trajectories (left) and the temperature T-trajectories (right).

To resolve this issue, guided by careful inspection of the conformational dynamics along R-trajectories, we also monitored additional collective variables such as the d_{SS} distance (i.e., side chain–side chain). The d_{SS} distance is defined and discussed in detail in Fig. S3 of the [supplementary material](#). Once again, the d_{SS} probability distribution data (see Fig. 2) show a very good convergence of the REMD simulation. Figure 2 shows the corresponding probability distributions of the d_{SS} distance for the lowest (black) and highest REMD temperatures T-trajectories (blue) and the data including all the R-trajectories (red). However, we note that the d_{SS} distance

FIG. 1. Probability distributions of the d_{TT} distance between the charged termini of FF peptides for each of the REMD replica R-trajectories (left) and the temperature T-trajectories (right). While the distributions show good convergence in both cases, a more detailed analysis shows that the second pick corresponds actually to two different conformational states, with representative (i.e., most populated) structures for both peaks being illustrated in the middle.

FIG. 2. Probability distributions of the d_{SS} (i.e., side chain-side chain, see Fig. S3 of the [supplementary material](#)) distance for the lowest (black) and highest REMD temperatures T-trajectories (left) and the data including all the R-trajectories (red). Once again, these distributions show a very good convergence of the REMD simulation; however, the d_{SS} distance is able to group similar conformations under corresponding peaks and discriminate the main *cis* state (S_1) and the two *trans* states (S_2 and S_3) from each other. Representative structures for all peaks are also illustrated accordingly.

is now able to also group similar conformations under corresponding peaks and discriminate the main *cis* (state S_1) and the two *trans* (states S_2 and S_3) from each other. We also note a relatively weak temperature dependence of the d_{SS} distance: as temperature was increased, the width of the corresponding distribution increases only slightly, with the R-trajectory data from all trajectories (red) corresponding to intermediate values. Representative structures for all peaks are also illustrated accordingly in Fig. 2.

While the results presented above are encouraging, in order to avoid relying entirely on the d_{SS} distance for subsequent analysis, we strengthen the relevance of the conformational states revealed by the d_{SS} distance distributions by noting that another independent collective variable, the radius of gyration, R_g , leads to identical representative structures for the three main candidates for conformational Markovian states in our FF system. As shown in Fig. 3, the probability distributions of the radius of gyration, R_g , calculated for the lowest (black) and highest REMD temperatures T-trajectories (blue), and the data including all the R-trajectories (red), unveil the same representative conformations of the three main states. Once again, these distributions show a very good convergence of the REMD simulation, and support the results obtained for the d_{SS} distance by grouping similar conformations under corresponding peaks and discriminate the main *cis* (state S_1) and the two *trans* (states S_2 and S_3) from each other. The corresponding representative molecular structures for all peaks are also illustrated in Fig. 3.

Our implementation of the TBA method starts from identifying the regions of conformational space that correspond to local free energy minima and thus equilibrium probability maxima along a reasonably good reaction coordinate set. From any equilibrium MD simulations, including REMD, the location of the main occupation probability maxima presents

FIG. 3. Probability distributions of the radius of gyration, R_g , calculated for the lowest (black) and highest REMD temperatures T-trajectories (blue) and the data including all the R-trajectories (red). Once again, these distributions show a very good convergence of the REMD simulation and support the results obtained for the d_{SS} distance by grouping similar conformations under corresponding peaks and discriminating the main *cis* (state S_1) and the two *trans* (states S_2 and S_3) from each other. Representative structures for all peaks are also illustrated.

the highest confidence (i.e., as these conformational regions are easy to sample) and can be used to find the representative molecular conformations for Markov state regions (i.e., regions separated by large enough free energy barrier such that subsequent transitions from one state to another are statistically independent). Here, for the 3 main states proposed above for FF molecules, we use their corresponding representative conformations as alignment references for generating the corresponding pairwise RelRMSD measure defined above, along the molecular trajectory. Figure 4 shows an illustration of the transition-based assignment method (TBA) using the relative RMSD values as reaction coordinates. In a first step, the relative pairwise RMSD values are calculated [as shown in Fig. 4(a), for all replica data] by using the representative conformations of the S_1 , S_2 , and S_3 states. For multiple states, one state (here, S_1) is chosen as the primary reference state and a $N - 1$ dimensional analysis (here 2D for 3 states) is performed where N is the number of candidate states. In this $N - 1$ space, the use of RelRMSD allows the definition of a spatial parameter, the neighborhood radius, r_{RC} , that can be used in the TBA analysis. An optimal r_{RC} would correspond to a situation for which the extracted transition probabilities (and the associated kinetic parameters) are essentially insensitive to variations of the precise choice of r_{RC} values. Using initially a small value for the neighborhood radius, r_{RC} , in this $N - 1$ dimensional plane (here, 2D for 3 states), all the points closer than r_{RC} from the S_1 representative conformation (red, see Fig. 4) are assigned to state 1, all points in the corresponding circle for S_2 are assigned to state 2 (green), and similarly for S_3 (blue points). The trajectory points outside the assigned regions (black) are first assigned to a temporary, intermediate state (I, black). Pr_{12} , Pr_{13} , Pr_{23} , and Pr_{123} denote state-space areas that are typically prohibited for the MD trajectory if the states used are well defined (i.e., exclusive).^{18,64} In a second

FIG. 4. Illustration of the transition-based assignment method (TBA) using the relative RMSD values as reaction coordinates. (a) In a first step, the relative pairwise RMSD values are calculated (here, shown for all replica data) by using the representative conformations of the S_1 , S_2 , and S_3 states. Using initially a small value for the neighborhood radius, r_{RC} , in this plane, all the points closer than r_{RC} from the S_1 representative conformation (red) are assigned to state 1, all points in corresponding circle for S_2 are assigned to state 2 (green), and similarly for S_3 (blue points). The trajectory points outside the assigned regions (black) are first assigned to a temporary, intermediate state (I, black). Pr_{12} , Pr_{13} , and Pr_{123} denote state-space areas that are typically prohibited for the MD trajectory if the states used are well defined (i.e., exclusive). (b) Trajectory segments from the intermediate state (I, black) of the previous step are subsequently re-assigned to their corresponding states in the TBA method by considering the history of their transition paths.

TBA step [see Fig. 4(b)], trajectory segments from the intermediate state (I, black) of the previous step are subsequently re-assigned to their corresponding states in the TBA method by considering the history of their transition paths (i.e., which state the transition paths come and go to). See Ref. 64 for a detailed review of the TBA method.

Figure 5(a) illustrates the 2D transition-based assignment method (TBA) using the RelRMSD values as reaction coordinates, for the entire T-REMD data (i.e., using all replicas), which is applied for different r_{RC} values. At the end of the TBA steps, all the conformations along the molecular (here, REMD) trajectory are assigned to one of the 3-states: S_1 (red), S_2 (green), or S_3 (blue points). Here, the circles correspond to the “high confidence zone” (HCZ) with $r_{RC} = 0.127$, which was selected based on the calculations for several r_{RC} values as shown in Fig. 5(b). The optimal r_{RC} value is found following an analysis of the sensitivity of the slowest relaxation time extracted from the trajectory to various r_{RC} values in the TBA analysis [Fig. 5(b)].

Similarly, Fig. 6(a) illustrates the application of the 2D transition-based assignment method using the RelRMSD₁₂ and RelRMSD₁₃ values as reaction coordinates, but this time for the T-trajectory data corresponding to the highest temperature $T_{high} = 385.1$ K. Circles correspond to high confidence state assignment zones with a radius of $r_{RC} = 0.103$, which was selected based on the calculations for several r_{RC} values as shown in (b). Figure. 6(b) presents the results of the sensitivity analysis of the slowest relaxation time extracted from the trajectory to various r_{RC} values in the TBA analysis for MD data at $T_{high} = 385.1$ K.

Similarly, Fig. 7(a) shows the corresponding 2D TBA method implementation using the 2D RelRMSD₁₂ and RelRMSD₁₃ values as reaction coordinates, but for the T-trajectory corresponding to the lowest temperature $T_{low} = 310$ K. Circles correspond to high confidence state assignment zones with a radius of $r_{RC} = 0.077$, which was selected based on the calculations for several r_{RC} values as shown in Fig. 7(b). The results of the corresponding detailed

FIG. 5. (a) 2D transition-based assignment method (TBA) using the relative RMSD values as reaction coordinates, for the entire T-REMD data (i.e., using all replicas), is applied for different r_{RC} values. After TBA, all the conformations along the trajectory are assigned to one of the 3-states: S_1 (red), S_2 (green), or S_3 (blue points). Here the circles correspond to the “high confidence zone” with $r_{RC} = 0.127$, which was selected based on the calculations for several r_{RC} values as shown in (b). (b) Sensitivity of the slowest relaxation time extracted from the trajectory to various r_{RC} values in the TBA analysis.

FIG. 6. (a) 2D transition-based assignment method (TBA) using the relative RMSD values as reaction coordinates for the T-trajectory corresponding to the highest temperature $T_{high} = 385.1$ K. Circles correspond to high confidence state assignment zones with a radius of $r_{RC} = 0.103$, which was selected based on the calculations for several r_{RC} values as shown in (b). (b) Sensitivity of the slowest relaxation time extracted from the trajectory to various r_{RC} values in the TBA analysis for MD data at $T_{high} = 385.1$ K.

FIG. 7. (a) 2D transition-based assignment method (TBA) using the relative RMSD values as reaction coordinates for the T-trajectory corresponding to the lowest temperature $T_{low} = 310$ K. Circles correspond to high confidence state assignment zones with a radius of $r_{RC} = 0.077$, which was selected based on the calculations for several r_{RC} values as shown in (b). (b) Sensitivity of the slowest relaxation time extracted from the trajectory to various r_{RC} values in the TBA analysis for MD data at $T_{low} = 310$ K.

sensitivity analysis of the slowest relaxation time extracted from the trajectory to various r_{RC} values in the TBA analysis for MD data at $T_{low} = 310$ K are shown in Fig. 7(b).

Figure 8 shows the temperature-dependent Markov kinetic networks extracted from our simulations of REMD trajectories for FF peptides. The transition rates (blue arrows) are shown between the three major conformational Markovian states S_1 , S_2 , and S_3 and the corresponding population (percentages) of these states for [see Fig. 8(a)] the T-trajectory data at $T_{high} = 385.1$ K (highest temperature). The data corresponding to all the replicas (all R-trajectories) and, finally, the T-trajectory REMD data at $T_{low} = 310$ K are also shown in Fig. 8(b). An arrow's thickness is proportional to the magnitude of its corresponding transition probability.

Finally, the data in Table S2 of the [supplementary material](#) show values of the statistical errors for the final extracted Markovian transition rates and the corresponding state equilibrium populations illustrated in Fig. 8 (main text), estimated using blocks of REMD data, as errors of the means calculated for each data block. As highlighted in Sec. II, it is important to note that the magnitude of the statistical errors in the extracted intrinsic kinetic parameters offers a reliable assessment method for the convergence of molecular simulations analyzed (as applied here to REMD). Together with the simpler “equal occupation” for each replica at each temperature, test results illustrated in Figure S1 of the [supplementary material](#), these errors allow for a more clear and quantitative assessment of the quality of the convergence of the molecular simulation trajectories used in the kinetic analysis study.

FIG. 8. Temperature-dependent kinetics of FF peptides from REMD trajectories. Shown are the transition rates (blue arrows, see the text) between the three major conformational Markovian states S_1 , S_2 , and S_3 and the corresponding population (percentages) of these states; (a) for the T-trajectory data at $T_{high} = 385.1$ K (highest temperature), (b) for the data corresponding to all the replicas (all R-trajectories), and, finally, (c) for the T-trajectory REMD data at $T_{low} = 310$ K. Each arrow's thickness is proportional to the magnitude of its corresponding transition rate. The corresponding statistical errors are given in Table S2 of the [supplementary material](#).

IV. CONCLUSIONS

In summary, we show (using REMD atomistic trajectories of explicitly solvated FF peptides) that, in cases when representative molecular conformations (e.g., actual atomistic coordinates) can be identified for the corresponding Markov states, and thus their corresponding collective evolution of atomic positions can be calculated along MD or REMD trajectories, one can use these representative conformations to build a new type of a simple collective variable that can be particularly useful in both the correct assignment of the state from the trajectory and in the subsequent accurate counting of inter-state transition probabilities. In particular, we considered the ubiquitously used root-mean-square deviation (RMSD) of atomic positions,^{35,36} in order to introduce the relative RMSD (RelRMSD) measure as a good reaction coordinate candidate in Markov-based analysis. We apply this method to the analysis of REMD trajectories of amyloid-forming diphenylalanine (FF) peptides—a system with important nanotechnology and biomedical applications due to its self-assembling and piezoelectric properties³⁷—illustrating the use of RelRMSD in extracting from the data the underlying temperature-dependent kinetics, without *a priori* assumptions on the functional form (e.g., Arrhenius or not) of the intrinsic conformational transition rates, using a simple method for direct counting of interstate transitions from REMD data.^{23,24} The RelRMSD analyses also enable a simple and objective assessment of the convergence of the REMD simulations, by allowing the quantitative estimation of statistical errors of the kinetic parameters extracted from the REMD data.^{38–40} The RelRMSD collective variable construction method may be generalized to other collective variables that are able to accurately capture conformational deviations between the underlying Markov states (e.g., dRMSD,⁴¹ the fraction of native contacts,⁴³ etc.).

SUPPLEMENTARY MATERIAL

See [supplementary material](#) for the three supplementary figures and their captions, Figs. S1–S3, and Table S1, as mentioned in the main text.

ACKNOWLEDGMENTS

We wish to thank the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC), and the Biowulf cluster at the National Institutes of Health, United States (<http://biowulf.nih.gov>) for the provision of computational facilities and support. We gratefully acknowledge financial support from the Irish Research Council (IRC) for C.H. and N.-V.B., and from the UCD School of Physics for B.N.

NOMENCLATURE

MD	=	molecular dynamics
REMD	=	replica-exchange molecular dynamics
TBA	=	transition-based assignment
MLPB	=	maximum likelihood propagator-based

- ¹G. E. Moore, *IEEE Solid-State Circuits Soc. Newsletter* **11**, 33 (2006).
- ²K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, *Science* **334**, 517 (2011).
- ³M. Shirts and V. S. Pande, *Science* **290**, 1903 (2000).
- ⁴B. E. Husic and V. S. Pande, *J. Am. Chem. Soc.* **140**, 2386 (2018).
- ⁵Z. Cournia, T. W. Allen, I. Andricioaei, B. Antonny, D. Baum, G. Brannigan, N. V. Buchete, J. T. Deckman, L. Delemotte, C. del Val, R. Friedman, P. Gkeka, H. C. Hege, J. Henin, M. A. Kasimova, A. Kolocouris, M. L. Klein, S. Khalid, M. J. Lemieux, N. Lindow, M. Roy, J. Selent, M. Tarek, F. Tofoleanu, S. Vanni, S. Urban, D. J. Wales, J. C. Smith, and A. N. Bondar, *J. Membrane Biol.* **248**, 611 (2015).
- ⁶S. Piana, K. Lindorff-Larsen, and D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17845 (2012).
- ⁷R. B. Best, *Curr. Opin. Struct. Biol.* **22**, 52 (2012).
- ⁸N. V. Buchete and G. Hummer, *J. Phys. Chem. B* **112**, 6057 (2008).
- ⁹L. Martini, A. Kells, R. Covino, G. Hummer, N.-V. Buchete, and E. Rosta, *Phys. Rev. X* **7**, 031060 (2017).
- ¹⁰Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.* **314**, 141 (1999).
- ¹¹A. E. Garcia and K. Y. Sanbonmatsu, *Proc. Natl. Acad. Sci. U. S. A.* **99**, 2782 (2002).
- ¹²Q. Lu, J. Kim, J. D. Farrell, D. J. Wales, and J. E. Straub, *J. Chem. Phys.* **141**, 18C525 (2014).
- ¹³A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsovvelaminov, *J. Chem. Phys.* **96**, 1776 (1992).
- ¹⁴E. Marinari and G. Parisi, *Europhys. Lett.* **19**, 451 (1992).
- ¹⁵T. Nagai, G. A. Pantelopulos, T. Takahashi, and J. E. Straub, *J. Comput. Chem.* **37**, 2017 (2016).
- ¹⁶N. V. Buchete and G. Hummer, *Phys. Rev. E* **77**, 030902(R) (2008).
- ¹⁷B. Strodel, C. S. Whittleston, and D. J. Wales, *J. Am. Chem. Soc.* **129**, 16005 (2007).
- ¹⁸N. V. Buchete, *Biophys. J.* **103**, 1411 (2012).
- ¹⁹A. T. Frank and I. Andricioaei, *J. Phys. Chem. B* **120**, 8600 (2016).
- ²⁰L. S. Stelzl and G. Hummer, *J. Chem. Theory Comput.* **13**, 3927 (2017).
- ²¹M. Carballo-Pacheco and B. Strodel, *Protein Sci.* **26**, 174 (2017).
- ²²D. De Sancho and R. B. Best, *J. Am. Chem. Soc.* **133**, 6809 (2011).
- ²³C. T. Leahy, R. D. Murphy, G. Hummer, E. Rosta, and N.-V. Buchete, *J. Phys. Chem. Lett.* **7**, 2676 (2016).
- ²⁴C. T. Leahy, Ph.D. thesis, University College Dublin, School of Physics, Belfield, Dublin 4, Ireland, 2013, p. 107.
- ²⁵C. T. Leahy, A. Kells, G. Hummer, N.-V. Buchete, and E. Rosta, *J. Chem. Phys.* **147**, 152725 (2017).
- ²⁶D. van der Spoel and M. M. Seibert, *Phys. Rev. Lett.* **96**, 238102 (2006).
- ²⁷M. Oliveberg, Y. J. Tan, and A. R. Fersht, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8926 (1995).
- ²⁸F. Zhu, *J. Chem. Phys.* **146**, 124128 (2017).
- ²⁹G. Hummer and A. Szabo, *J. Phys. Chem. B* **119**, 9029 (2015).
- ³⁰B. E. Husic, K. A. McKiernan, H. K. Wayment-Steele, M. M. Sultan, and V. S. Pande, *J. Chem. Theory Comput.* **14**, 1071 (2018).
- ³¹F. Noé and F. Nüske, *Multiscale Model. Simul.* **11**, 635 (2013).
- ³²F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, *J. Chem. Theory Comput.* **10**, 1739 (2014).
- ³³C. R. Schwantes and V. S. Pande, *J. Chem. Theory Comput.* **9**, 2000 (2013).
- ³⁴G. Pérez-Hernández, F. Paul, T. Giorgino, G. D. Fabritiis, and F. Noé, *J. Chem. Phys.* **139**, 015102 (2013).
- ³⁵W. Kabsch, *Acta Crystallogr., Sect. A* **32**, 922 (1976).
- ³⁶E. A. Coutsiias, C. Seok, and K. A. Dill, *J. Comput. Chem.* **25**, 1849 (2004).
- ³⁷A. Kholkin, N. Amdursky, I. Bdikin, E. Gazit, and G. Rosenman, *ACS Nano* **4**, 610 (2010); C. M. Kelly, T. Northey, K. Ryan, B. R. Brooks, A. L. Kholkin, B. J. Rodriguez, and N.-V. Buchete, *Biophys. Chem.* **196**, 16 (2015); K. Ryan, J. Beirne, G. Redmond, J. I. Kilpatrick, J. Guyonnet, N.-V. Buchete, A. L. Kholkin, and B. J. Rodriguez, *ACS Appl. Mater. Interfaces* **7**, 12702 (2015).
- ³⁸E. Rosta and G. Hummer, *J. Chem. Phys.* **131**, 165102 (2009).
- ³⁹J. D. Doll and P. Dupuis, *J. Chem. Phys.* **142**, 024111 (2015).
- ⁴⁰J. D. Doll, N. Plattner, D. L. Freeman, Y. Liu, and P. Dupuis, *J. Chem. Phys.* **137**, 204112 (2012).
- ⁴¹K. Nishikawa and T. Ooi, *J. Theor. Biol.* **43**, 351 (1974).
- ⁴²C. Magis, P. Di Tommaso, and C. Notredame, *Nucleic Acids Res.* **41**, W358 (2013).
- ⁴³R. B. Best, G. Hummer, and W. A. Eaton, *Proc. Natl. Acad. Sci.* **110**, 17874 (2013).
- ⁴⁴D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, *J. Comput. Chem.* **26**, 1701 (2005).

- ⁴⁵B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, *J. Chem. Theory Comput.* **4**, 435 (2008).
- ⁴⁶R. W. Pastor, B. R. Brooks, and A. Szabo, *Mol. Phys.* **65**, 1409 (1988).
- ⁴⁷M. Parrinello and A. Rahman, *J. Chem. Phys.* **80**, 860 (1984).
- ⁴⁸B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, *J. Comput. Chem.* **30**, 1545 (2009).
- ⁴⁹R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, *J. Chem. Theory Comput.* **8**, 3257 (2012).
- ⁵⁰W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- ⁵¹A. Patriksson and D. van der Spoel, *Phys. Chem. Chem. Phys.* **10**, 2073 (2008).
- ⁵²W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graph* **14**, 33 (1996).
- ⁵³R. Zwanzig, *J. Stat. Phys.* **30**, 255 (1983).
- ⁵⁴C. Schütte, A. Fischer, W. Huisinga, and P. Deuffhard, *J. Comput. Phys.* **151**, 146 (1999).
- ⁵⁵B. L. De Groot, X. Daura, A. E. Mark, and H. Grubmüller, *J. Mol. Biol.* **309**, 299 (2001).
- ⁵⁶Y. Levy, J. Jortner, and R. S. Berry, *Phys. Chem. Chem. Phys.* **4**, 5052 (2002).
- ⁵⁷W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou, *J. Phys. Chem. B* **108**, 6582 (2004).
- ⁵⁸D. S. Chekmarev, T. Ishida, and R. M. Levy, *J. Phys. Chem. B* **108**, 19487 (2004).
- ⁵⁹S. Sriraman, I. G. Kevrekidis, and G. Hummer, *J. Phys. Chem. B* **109**, 6479 (2005).
- ⁶⁰J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, *Multiscale Model. Simul.* **5**, 1214 (2006).
- ⁶¹W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy, *J. Phys. Chem. B* **113**, 11702 (2009).
- ⁶²A. M. Berezhkovskii, F. Tofoleanu, and N.-V. Buchete, *J. Chem. Theory Comput.* **7**, 2370 (2011).
- ⁶³A. M. Berezhkovskii, R. D. Murphy, and N.-V. Buchete, *J. Chem. Phys.* **138**, 036101 (2013).
- ⁶⁴G. S. Buchner, R. D. Murphy, N. V. Buchete, and J. Kubelka, *BBA-Proteins Proteomics* **1814**, 1001 (2011).
- ⁶⁵J. W. Wagner, J. F. Dama, A. E. P. Durumeric, and G. A. Voth, *J. Chem. Phys.* **145**, 044108 (2016).
- ⁶⁶H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, *J. Chem. Phys.* **81**, 3684 (1984).