How Accurately Do Current Force Fields Predict Experimental Peptide Conformations? An Adiabatic Free Energy Dynamics Study

Alexandar T. Tzanov,‡§ Michel A. Cuendet,‡†§ and Mark E. Tuckerman*‡§

†Department of Chemistry, New York University, New York, New York 10003, United States
‡Swiss Institute of Bioinformatics, UNIL Sorge, 1015 Lausanne, Switzerland
§Courant Institute of Mathematical Sciences, New York University, New York, New York 10003, United States

Supporting Information

ABSTRACT: The quality of classical biomolecular simulations is inevitably limited by two problems: the accuracy of the force field used and the comprehensiveness of configuration space sampling. In this work we tackle the sampling problem by carrying out driven adiabatic free energy dynamics to obtain converged free energy surfaces of dipeptides in the gas phase and in solution using selected dihedral angles as collective variables. To calculate populations of conformational macrostates observed in experiment, we introduce a fuzzy clustering algorithm in collective-variable space, which delineates macrostates without prior definition of arbitrary boundaries. With this approach, we calculate the conformational preferences of small peptides with six biomolecular force fields chosen from among the most recent and widely used. We assess the accuracy of each force field against recently published Raman or IR–UV spectroscopy measurements of conformer populations for the dipeptides in solution or in the gas phase.

1. INTRODUCTION

Folding to a specific three-dimensional structure is essential for a protein to function properly. Because metastable conformational states represent kinetic pitfalls that reflect competing and mutually exclusive interactions, it is crucial to characterize not only the native state but also all non-native states. Experimental X-ray or nuclear magnetic resonance (NMR) structures are the usual starting point for studies of relations between the native structure of a protein and its function. The NMR and/or X-ray structures, however, do not provide direct information about the dynamics of the protein, its folding process, or its unfolded state. Experimental characterization of metastable states relies on numerous advanced techniques: NMR,1 infrared/ultraviolet (IR–UV) hole-filling and infrared-induced population transfer spectroscopies,2–9 attenuated total reflection (ATR) spectroscopy,10 and Raman spectroscopy,10 all of which can probe different stages during the folding process. The experimental picture is, however, often incomplete, such as when the folding times are within the same time scale as experiment.

In many cases, computer simulation is a necessary complement to experiment to obtain a more detailed representation of conformational ensembles or of folding processes. Given the complexity of the systems and the large amount of sampling required to explore the associated conformational spaces, classical molecular dynamics (MD) simulation is the method of choice. For the most part, MD force fields (FFs) have been validated through their ability to find the folded states of proteins as a global minimum of the free energy surface. The accuracy of MD simulations to describe unfolded states has yet to be validated against experiment.

The present study represents a step in this direction. We use recent enhanced sampling techniques to extensively explore the conformational space of two small peptides: N-acetylaniline-N′-methylamide (NA-A-MA or alanine dipeptide) and N-acetyltrypophan-N′-methylamide (NA-T-MA) (Figure 1). Though several computational studies have focused on comparing the free energy surfaces (FES) of dipeptides for different FFs,11–13 comparison to experiment requires one additional step. We need to introduce the notion of conformational macrostates and calculate the corresponding macrostate populations to compare to experimental results. Here, we propose a fuzzy clustering technique to delineate conformational macrostates based on the MD data and calculate populations in a robust way for the specific conformations observed in a given experiment.

Comparison of six different FFs (CHARMM27, CHARMM36, AMBER99SB, FUJI, OPLS-AA/L, and GROMOS 54a7, as discussed below) reveals a large degree of variability among conformational preferences predicted by the different models. We compare these results to conformational populations obtained by infrared and Raman spectroscopy for NA-A-MA and NA-T-MA in solution and by IR–UV fluorescence spectroscopy after supersonic expansion for NA-T-MA in the gas phase. This enables a critical assessment of the ability of the six FFs to reproduce experimentally observed conformational ensembles of...
small peptides. In contrast to previous studies, our goal here is not to exhaustively explore the conformational states of all twenty amino acids. Rather, we focus on two representative systems for which experimental data are available and describe, in detail, how we classify the conformational states so that comparison to experiment can be made in the most unbiased way. Having a reliable model for these small systems and a robust methodology for analyzing the conformational states are necessary prerequisites for any meaningful description of unfolded states in larger proteins.

1.1. Conformational States of Peptides. 1.1.1. Definition of Macrostates. Conformations of amino acids are often represented in the Ramachandran map, i.e., the plane spanned by the two backbone dihedral angles ϕ and ψ defined by the atom quadruplets (C–N–Cα–C) and (N–Cα–C–N), respectively. On the Ramachandran map, all non-proline and non-glycine amino acids preferentially occupy similar regions schematically depicted in Figure 2a. These regions define archetypical macrostates that are related to the secondary structure of peptide chains. Most of the conformations fall into two major basins: the helical α basin with αR (right-hand helix structure) and α′, and the extended β basin with the C5 and polyproline II (PⅡ) conformations. Other higher energy macrostates related to loops and turns are αL (left-hand helix), C7ax (γ turn), and C7eq. Figure 2a shows how these preferred conformations are mapped to the (ϕ, ψ) plane, and Table 1 gives archetypal coordinates for the centers of these regions. More recent extensive analysis of high resolution crystal structures suggested that residues may also favor conformations with (ϕ, ψ) outside “classical” regions.

In amino acids with an indole ring side chain (NA-T-MA, melatonin, etc.), the conformations can be described in further detail. The conformation of the backbone with respect to the indole ring plane is denoted with four letters that have the following meaning. The first two letters indicate the conformation of the backbone by its position in the Ramachandran map, as shown in Figure 2a. The last two letters indicate the position of the N-terminus and C-terminus relative to the indole ring: A means anti, P denotes gauche on the pyrrole side, and Φ denotes gauche on the phenyl side. Using this notation, C7eq(AP) represents a C7 equatorial peptide backbone with the N-terminus anti to indole ring and the C-terminus gauche to the indole ring on the pyrrole side of the indole. Examples of these structures for NA-T-MA are shown in Figure 3.

1.1.2. Experimental Approaches. Experimental approaches often utilize dipeptides, i.e., N-acetyl-N′-methylamides, of single
Table 1. Canonical Conformations of Non-proline and Non-glycine Amino Acids in Solution and Their Position in the Ramachandran Plane

<table>
<thead>
<tr>
<th>conformation</th>
<th>φ</th>
<th>ψ</th>
<th>secondary structure</th>
<th>exp NA-A-MA</th>
<th>exp NA-T-MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5</td>
<td>−150</td>
<td>160</td>
<td>β-sheet</td>
<td>0.29</td>
<td>0.44</td>
</tr>
<tr>
<td>Pii</td>
<td>−65</td>
<td>150</td>
<td>polyproline II</td>
<td>0.61</td>
<td>0.54</td>
</tr>
<tr>
<td>αR</td>
<td>−60</td>
<td>−50</td>
<td>right-handed helix</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>α′</td>
<td>−130</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>αI</td>
<td>50</td>
<td>50</td>
<td>left-handed helix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C7 rotation</td>
<td>50</td>
<td>−130</td>
<td>γ turn</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**The (φ, ψ) values indicated for the C5, Pii, and αR conformations are used as archetypes for our clustering method. Experimental values are from ref 23.**

amino acids to gather insights about biomolecular conformation. It was found that three amino acids, tryptophan (Trp), phenylalanine (Phe), and tyrosine, (Tyr), all of which have an ultraviolet chromophore (indole and benzene rings), can exhibit conformation-specific infrared spectra when fluorescence dip infrared spectroscopy is combined with UV–UV spectroscopy.2–4,10,21,52 In particular, NA-T-MA engendered much attention as an object of experimental investigation despite the fact that tryptophan itself is not in abundance in biological systems. Its bulky side chain gives rise to a complex potential energy surface, the detailed characterization of which provides insights that can be carried over to larger systems.

In the IR–UV hole-filling method,2–4,7,22 single modes in gas phase NA-T-MA molecules at 423 K are selectively excited using IR radiation before the molecules are cooled by supersonic jet expansion and probed by UV spectroscopy. The population transfer spectroscopy (PTS) technique analyzes the changes in populations of conformers as a function of infrared wavelength to generate populations directly from spectra.22 In refs 2–4 and 22, the three NA-T-MA macrostates appearing in the spectra are labeled with A, B, and C. Using a combination of empirical arguments and molecular modeling, these spectroscopic macrostates were assigned to conformers described by the four-letter codes C5(Φ), C5(AP), and C7 rotation(ΦP), respectively. These four-letter codes can, in turn, be mapped to regions in the four-dimensional space spanned by φ, ψ, and the two dihedral angles χ1 and χ2. The centers of these regions, which we call archetypes, are reported in Table 2. The populations of conformers A, B, and C were measured to be 0.23, 0.40, and 0.37, respectively, in refs 3 and 22. A subsequent experiment by the same authors4 yielded slightly different values of 0.20, 0.50, and 0.30, respectively. The latter values are assumed to be of higher accuracy because of better expansion cooling conditions. Error bars are reported to be of the order of 0.02 for all three numbers.

In a recent landmark study, Grdadolnik et al.23 used the amide III region of the peptide IR and Raman spectra to infer populations for 19 dipeptides in solution. The analysis in this study is restricted to the three major conformers C5, Pii, and αR. The relative intensity of the three bands is fitted in both IR and Raman spectra,24 and the bands are assigned to the three backbone conformers using NMR J-couplings. For NA-A-MA, the populations reported for C5, Pii, and αR were 0.29, 0.61, and 0.10, respectively. For NA-T-MA, the corresponding populations are 0.44, 0.54, and 0.02, with an uncertainty of 0.02 for all (Table 1).

1.2. Molecular Simulation as a Tool for Structure Prediction. Besides the sampling problem, which we address in the following sections, the central question in MD simulations is the suitability of a given FF to study a particular molecular system. In biomolecular systems, establishing agreement between simulated and experimental results or between simulations with different FFs is often highly dependent on the complexity of a system. In the case of protein folding, for example, it is difficult to disentangle effects of the local φ/ψ preference of each amino acid and the effects of long-range and cooperative phenomena. This is the key question for FF transferability to large systems, as it is well-known that amino acids show different conformational propensities when alone, in short peptide chains or in proteins.25,26 Nevertheless, if we wish to obtain a correct answer for the right reason, we must ensure that current FFs are reliable models for the shortest peptide chains, i.e., the dipeptide molecules.

Most standard classical FFs for biomolecular simulations come from four families: CHARMM,27–29 AMBER,30,31 OPLS,32–35 and GROMOS36–38 (see refs 39 and 40 for more complete reviews on FFs). These FFs are nonpolarizable (they use fixed partial charges on atomic centers). All FFs of these families utilize very similar potential energy functions including so-called bonded terms for bond-stretching, angle-bending, and torsions of rotatable bonds, as well as nonbonded terms in the form of Coulomb and Lennard-Jones potentials. Small differences include the specific functional form for torsion potentials and the treatment of 1–4 interactions.

Parameter sets are typically obtained by fitting to quantum mechanical (QM) calculations for small molecules. These initial parameters are refined by following a procedure that varies among FFs but usually involves obtaining averages from long MD simulations, which are fed back to adjust parameters to reproduce experimental macroscopic properties, such as the solvation free energy. Another common approach involves fitting FF energies to a set of QM calculations done on incrementally larger molecular systems. The three main reasons

![Figure 3. Structures of NA-T-MA in the gas phase corresponding to states A, B, and C as assigned in ref 2 (Table 2).](image-url)
for discrepancies among FFs, both from the same family and across families, are (i) differences in QM methodology for the initial parameter derivation, (ii) differences in training sets used, and (iii) differences in parameter tuning strategies.

The quality or accuracy of a FF is usually assessed from its ability to stabilize experimentally observed protein structures or to reproduce other indirect experimental observables reflecting conformational equilibria of peptides or proteins. A difficulty faced by all classical biomolecular FFs has been an improper balance between predicted $\alpha$ and $\beta$ structures. Well-known examples of structural preferences across FFs are the bias of AMBER94 toward $\alpha$-helical structures and AMBER03 toward $\beta$-helical structures and inclination of the AMBER96, AMBER99SB, and CHARMM22 toward $\beta$ structures in explicit solvent. We note that these tendencies can differ in implicit solvent where, for example, AMBER99SB was shown to favor $\alpha$-helical structures. Very subtle modifications to commonly used molecular mechanical potentials, such as changes to the scaling factor for $1\cdots4$ electrostatic interactions, can significantly alter the behavior of those FFs with respect to the stabilizing/destabilizing of protein substructure. Because of this, the primary questions for all classical biomolecular FFs are (i) what is the limit of accuracy of a given functional form? and (ii) how can the parameters be adjusted to achieve optimal reliability? In the following section, we provide background information about the FFs considered in this study.

1.2.1. Overview of the Force Fields Considered in This Study. The first member of the CHARMM FF family was named EF2 and was designed to be used in a vacuum. This FF included both united-atom and all-atom models. The nonbonded terms were derived from dimer interaction energies, and EF2 also included an explicit term for hydrogen bonds. The next generation, CHARMM22, differed significantly in that it included the explicit TIP3P water model and the explicit representation of hydrogen bonds from EF2 was removed. Partial charges were based on a reproduction of QM minimum interaction energies and on distances between small molecule dimers. The van der Waals or Lenard-Jones parameters were transferred from the proteins or based on crystal heats of sublimation. A more recent version of the CHARMM force field used in this study, CHARMM27, includes many improvements for non-protein systems. For proteins, the only difference with CHARMM22 was the addition of an extra term in the potential energy function to account for a joint $\phi$- and $\psi$-angular contributions to the energy, which is the dihedral cross term or CMAP correction, based on high-level QM calculations of the glycine, alanine, and proline dipeptides.

The most recent development in the CHARMM family of FFs is CHARMM36, which offers an improved representation of the potential energy surface of proteins due to two key modifications. First, the CMAP term was optimized to reproduce NMR scalar coupling data on peptides in solution. Second, the side-chain dihedral potentials were optimized against quantum mechanical energies from dipeptides and NMR data from unfolded proteins. CHARMM36 also includes various previous revisions, including the improvements for the tryptophan side chain. In the first member of the AMBER FFs, introduced in 1994, dihedral parameters were developed by fitting to relative QM energies of rotamers of glycine and alanine. The partial charges were obtained by fitting the electrostatic potential to calculations at the Hartree–Fock 6-31G* level of theory. Because the gas phase energy surfaces of alanine and glycine do not have a minimum in the $\alpha$ region, in the subsequent version, denoted AMBER96, the parameters for $\phi$ and $\psi$ were adjusted to reproduce $\alpha$-helical energies for the alanine tetrapeptide. It was reported, however, that AMBER96 simulations overestimate the $\beta$-strand propensity. The problem was addressed in AMBER99 where the backbone parameters were refined by including 11 structures of the alanine tetrapeptide along with the alanine dipeptide. The AMBER99 FF was later shown to overestimate the $\alpha$-helical conformations. More recently, Simerling and co-workers reparametrized the backbone ($\phi$, $\psi$) dihedral terms of AMBER99 by fitting the energies of multiple conformations of glycine and alanine tetrapeptides from high level ab initio quantum mechanical calculations to obtain the AMBER99SB FF. A further improvement of AMBER99SB called FUJI was reported by Fujitani et al., based on new backbone dihedral parameters derived by fitting torsional energy profiles obtained from very high level ab initio molecular orbital calculations for hydrogen-blocked and methyl-blocked glycine and alanine dipeptides. Note that, at the time the present study was being completed, the latest version of the AMBER package was released with the new AMBER12SB FF that contains corrected backbone and the side-chain torsion potentials.

The OPLS FF was introduced in 1995. In the first version of the FF, the torsional parameters were derived by fitting to rotational energy profiles obtained from ab initio molecular orbital calculations at the RHF/6-31G* level for more than 50 organic molecules and ions. The nonbonded parameters were developed in conjunction with Monte Carlo simulations to obtain thermodynamic and structural properties for 34 pure organic liquids including alkanes, alkenes, alcohols, ethers, acetals, thiois, sulfides, disulfides, aldehydes, ketones, and amides. The reparametrization of OPLS involved refitting the key Fourier torsional coefficients. A set of energies for all amino acids (more than 2000) was generated on the basis of geometry optimization and single-point LMP2/cc-pVTZ-ff calculations. Then $\phi$ and $\psi$ torsion angle parameters were refit to these data. In one case, the van der Waals and charge parameters were also modified on the basis of new liquid-state simulation data. The resulting FF, denoted OPLS-AA/L (L stands for LMP2), is the one employed in the simulations to be reported herein.

The first GROMOS FF was parametrized to reproduce spectroscopic or crystallographic structural data and was followed by a second-generation FF called S3A6 based on hydration properties of amino acid analogues. However, S3A6 was shown to destabilize $\alpha$-helices. This led to a reparametrization of the torsional angle terms based on a large set of crystal structures, which gave rise to the current GROMOS 54A7 version.

1.2.2. Earlier Computational Studies of Amino Acid Conformational Propensities. Dipeptides, and in particular the alanine dipeptide, are common benchmark systems to assess the accuracy and convergence properties of enhanced sampling molecular dynamics methods. In addition, FES of dipeptides have been used in a number of studies to compare different force fields. For example, Hu et al. computed distributions of the alanine and glycine dipeptides in the ($\phi$, $\psi$) plane with five force fields and compared them to similar distributions from QM/MM calculations. Liu et al. produced...
NA-A-MA FESs with AMBER94, AMBER03, OPLSAA, OPLSAA/L, and CHARMM27 and compared them to FESs from DFT calculations (where the FES is approximated using contributions of local vibrational modes, and an implicit solvation model was used). After an initial study\textsuperscript{11} on NA-A-MA, Vymetal et al. recently published an extensive survey\textsuperscript{12} of all amino acids comparing FESs from CHARMM27, AMBER99SB, AMBER03, and OPLS-AA/L FFs using the metadynamics method.\textsuperscript{68}

Some studies went one step further and compared simulation results with conformer populations observed in the database of all known protein structures. For example, Feig\textsuperscript{25} generated FESs for all amino acids using the CHARMM27 FF and compared them to statistics of known structures. More recently, Jiang et al.\textsuperscript{69} used replica exchange to calculate FESs for all amino acids with the OPLS-AA/L and AMBER03 FFs and compared backbone and side-chain conformational propensities to a database of known structures (excluding helix and hairpin regions). The same authors subsequently proposed a knowledge-based correction to the OPLS-AA/L to better reproduce the observed propensities.\textsuperscript{26}

Finally, a few authors attempted to bridge FES calculations with data from experiments, such as NMR or IR spectroscopy, which are able to resolve conformational populations of small peptides. As a very significant example, Best et al.\textsuperscript{20,70} computed FESs for several short peptides using six different force fields and compared the conformation-dependent NMR J-couplings to experiment. Moradi et al.\textsuperscript{71} compared the polyproline II propensities of short peptides calculated with AMBER99SB in implicit solvent to circular dichroism experiments. Recently, Cruz et al.\textsuperscript{72} calculated FESs for all amino acids in solution with OPLS-AA/L and compared populations of the Cs, P\textsubscript{g}, and α\textsubscript{g} conformers with the experimental results from Gradaldiuk et al.\textsuperscript{23} mentioned above.

Note that most computational studies mentioned above that infer conformer populations from the FES\textsuperscript{12,25,67,72} rely on hard rectangular macrostate boundaries in the (ϕ, ψ) plane. However, the dividing line between neighboring states separated by a low, wide free energy barrier is very arbitrary. For example, the separation between the α' and α\textsubscript{g} states is set at φ = −120° in ref 12, whereas it is defined as ψ = −15° in ref 25 for all FFs. This led other authors, for example, the authors of ref 71, to apply clustering techniques to infer macrostate boundaries from the data. Here, we develop a fuzzy clustering method that infers macrostates based on the simulation data, as well as available experimental observables, and assigns shared memberships to conformations in overlapping regions. We further show how relative macrostate populations can be calculated using such partial memberships from biased ensembles resulting from enhanced sampling MD simulations. We employ this approach to study the conformational preferences of two representative dipptide systems across all of the aforementioned FF families. \n
2. METHODS

2.1. Collective Variables and Macrostate Free Energies. Consider a molecular system with N\textsubscript{df} degrees of freedom described by a vector of coordinates r in Cartesian space. We assume that rare events can be characterized by a set of n collective variables (CVs), q(r) \(\equiv (q_1(r), q_2(r), \ldots, q_n(r))\), with \(n \ll N_{\text{df}}\). We define a conformaion \(s\) as a particular value of the CVs, \(s \equiv (s_1, s_2, \ldots, s_n)\). We denote by \(\Omega\) the space spanned by \(s\), with \(\dim(\Omega) = n\). The probability of observing a conformation \(q(r) = s\) in the canonical ensemble is given by

\[
P(s) = \frac{1}{Z} \int_{\mathbb{R}^n} dr \, e^{-\beta U(r)} \prod_{a=1}^{n} \delta(q_a(r) - s_a)
\]

where \(U(r)\) is the potential energy function and \(Z = \int e^{-\beta U(r)} dr\) is the partition function. Here, \(\delta(\cdot)\) is the Dirac delta function and \(\beta = (k_B T)^{-1}\), where \(T\) is the system temperature, and \(k_B\) is Boltzmann’s constant. The FES over the conformational space \(\Omega\) is then

\[
P(s) = -\frac{1}{\beta} \ln P(s)
\]

2.1.1. Conformational Macrostates as Regions of the CV Space. We further define a conformational macrostate \(C\) as an ensemble of conformations occupying a contiguous region \(\Omega(C) \subset \Omega\), which typically maps a significant basin of the FES. If the CVs are a good representation of the physical system, a macrostate should be an ensemble of rapidly interconverting conformations. Different macrostates are separated by barriers on the FES that, again under the assumption of a good choice of CVs, are related to the kinetics of macrostate transitions. The probability of observing (or the population of) macrostate \(C\) is

\[
P(C) = \int_{\Omega(C)} ds \, e^{-\beta F(s)}
\]

Note that in eq 3, no normalization constant is necessary because \(\int_{\Omega} e^{-\beta F(s)} = 1\), due to the normalization in eq 1. The free energy of macrostate \(C\) at temperature \(T\) is defined as

\[
F(C) = -\frac{1}{\beta} \ln P(C)
\]

According to standard thermodynamics, \(F(C)\) can be decomposed as \(F(C) = E(C) - TS(C)\), where \(E(C)\) is the enthalpy and \(S(C)\) is the entropy of macrostate \(C\). The overall balance between enthalpic and entropic contributions determines the stability of a macrostate at the given \(T\). Thus, it is possible for a given macrostate to be entropically stabilized despite an unfavorable enthalpic component. In terms of the FES, this means that a shallow but broad basin can represent a more stable macrostate than a deep but very localized minimum. In particular, the value of \(F(s)\) at the FES minimum is not the relevant quantity to determine the stability of a given macrostate, and eq 4 must be used instead.

2.1.2. Driven Adiabatic Free Energy Dynamics Enhanced Sampling Method. The PES of biomolecular systems is typically rough, with many local minima and barriers, which implies a wide distribution of time scales for the motions in these systems. Over the years, numerous methods have been developed to enhance sampling as well as to obtain dynamic information on various molecular systems.\textsuperscript{73,74} In particular, temperature accelerated molecular dynamics (TAMD)\textsuperscript{75} and diven-adiabatic free energy dynamics (AFED)\textsuperscript{76–78} are two similar methods that stem from an earlier approach, adiabatic free energy dynamics (AFED).\textsuperscript{76–78} The AFED enhanced sampling method imposes an adiabatic separation between a set of collective variables (CVs) and the remaining degrees of freedom of a system, after a coordinate transformation that makes the CVs explicit dynamical variables. In TAMD and dAFED, however, the adiabatic separation is applied to a set of extended variables tightly coupled to the CVs, rather than the CVs themselves. Explicit coordinate transformations are thus no longer required, rendering TAMD
and dAFED much easier to implement with almost no loss of efficiency. The full description of the TAMD and dAFED approach is given elsewhere, and in the following we summarize only the main aspects of the dAFED approach.

If we replace the product of δ functions in eq 1 with the limit of a product of Gaussian functions, we bring about terms in the exponential that can be viewed as harmonic potentials with harmonic constants \(k_1, \ldots, k_s\) that keep the collective variables \(q_1(\mathbf{r}), \ldots, q_s(\mathbf{r})\) close to the extended variables \(s_1, \ldots, s_s\). These extended variables are assigned momenta \(p_1, \ldots, p_s\) and fictitious masses \(m_1, \ldots, m_s\) and are evolved as independent dynamical variables. If high values are assigned to these masses, the motion of the \(s\) variables is slow with respect to the time scales of the physical system, and the \(s\) variables become adiabatically decoupled from the physical variables. Under these conditions, \(s\) evolves on the potential of mean force, eq 2, generated by the averaged interactions with the physical system at inverse temperature \(\beta\), according to the effective adiabatic Hamiltonian for \(s_s\)

\[
H_{\text{ad}}(p_s|s) = \sum_{\alpha=1}^{s} \frac{p_{s\alpha}^2}{2m_{s\alpha}} + F_{\beta}(s)
\]

(5)

If the adiabatic decoupling is effective, we can couple the extended system to a heat bath at temperature \(T_s > T\) while keeping the amount of heat transferred to the physical system limited. In this way, the extended variables can slowly but effectively drive the system across free energy barriers of height up to a few \(k_s T_s\) which results in enhanced sampling along the directions of the CVs. It can be shown that under the high-temperature and adiabatic conditions, the probability distribution \(P_{\text{ad}}(s)\) sampled from the adiabatic dynamics leads to a good approximation of the exact free energy energy profile at temperature \(T_s\),

\[
F_{\beta}(s) = -k_s T_s \ln P_{\text{ad}}(s)
\]

(6)

This approximation holds under the assumptions that the adiabatic separation is effective and that the coupling of \(s\) with \(q(\mathbf{r})\) is tight. These requirements can, in principle, be approached by choosing appropriate values for masses \(m_1, \ldots, m_s\) and coupling constants \(k_1, \ldots, k_s\), which are arbitrary parameters of the simulation. We note that under the same assumptions, the FES can also be reconstructed from the mean force exerted by the physical system on \(s\), a key element in the recent uniﬁed free energy dynamics (UFED) method that combined dAFED with metadynamics. The technical details of the simulations carried out in this study can be found in Appendix A.

2.2. Clustering Approach To Map Macrostates. If we have a set of \(N\) canonically distributed samples in CV space, e.g., collected during an MD or enhanced sampling simulation, we can estimate the probability of a conformational macrostate \(C\) at temperature \(T\) by counting the number of samples falling within \(\Omega(C)\) and dividing by \(N\). In comparison with experiment, a different normalization might be needed if some parts of the conformational space \(\Omega\) do not contribute to the experimental observable. This happens, for example, if conformer populations are inferred from measurements using a spectroscopic technique in which some conformations do not show any signature. These invisible conformations are de facto considered to be out of \(\Omega\).

The appropriate normalization constant is \(N_{\text{exp}}\), which is the number of samples falling into all \(M_{\text{exp}}\) conformational macrostates \(C_1, \ldots, C_{M_{\text{exp}}}\) observed in the experiment.

2.2.1. Macrostate Populations in the Adiabatic Ensemble. In the context of dAFED, the samples \(s_1, \ldots, s_N\) are obtained at temperature \(T_s\). We recently showed how to calculate the ensemble average of any observable at temperature \(T\) from a TAMD or dAFED simulation. If we are interested in populations, the relevant observable is the indicator function of macrostate \(C_i\), \(I_{\Omega(C)}(s)\), where \(I_{\Omega(C)} = 1\) if \(s \in \Omega(C)\) and 0 otherwise. This observable needs to be reweighted by a factor \(e^{−β[F_{\beta}(s) − F_{\beta}^{\text{obs}}]}\), with \(F(s)\) the dAFED FES calculated using eq 6. Thus, we can estimate the probability of observing macrostate \(C\) as

\[
P(C) = \frac{1}{N_{\text{exp}}} \sum_{i=1}^{N} I_{\Omega(C)}(s_i) e^{-\beta[F_{\beta}(s) − F_{\beta}^{\text{obs}}]}
\]

(7)

Many previous studies of small peptides have used predefined regions of the Ramachandran surface to deﬁne macrostates. Each of these regions was delimited by maximum and minimum values for \(\phi\) and \(\psi\) and had a rectangular shape. Using arbitrary regions of \(\Omega\) to deﬁne conformational macrostates can be problematic when different molecules are studied. For example, a fixed region \(\Omega(C)\) might not correctly encompass the free energy basin corresponding to macrostate \(C\) for all molecules. In addition, results can be sensitive to the precise location of the arbitrary boundary between two macrostates. This calls for data-driven approaches to identify macrostates, such as a grouping of conformations in the \((\phi, \psi)\) plane using a clustering algorithm, as was done in ref 71.

In the present study, we developed a fuzzy clustering approach to deﬁne states in the two- or four-dimensional CV space directly from a biased DAFED ensemble. The number \(M\) of clusters and the locations of the cluster centers \(c_1, \ldots, c_M\) are determined by the clustering algorithm (see below). For each sample point \(s\), the algorithm provides a fractional membership weight \(u_{ik} \in [0, 1]\) for cluster \(C_k\) with \(\sum_{k=1}^{M} u_{ik} = 1\). This defines an \(N \times M\) membership matrix \(U\). In this context, the cluster or macrostate probability eq 7 becomes

\[
P(C_k) = \frac{1}{N_{\text{exp}}} \sum_{i=1}^{N} u_{ik} e^{-\beta[F_{\beta}(s) − F_{\beta}^{\text{obs}}]}
\]

(8)

The normalization constant \(N_{\text{exp}}^{u}\) is the weighted sum of all membership coefﬁcients related to the experimentally observed macrostates and is given by

\[
N_{\text{exp}}^{u} = \sum_{k=1}^{M} \sum_{i=1}^{N} u_{ik} e^{-\beta[F_{\beta}(s) − F_{\beta}^{\text{obs}}]}
\]

(9)

For the \(M_{\text{exp}}\) archetypical macrostates, the populations \(P(C_k)\), normalized as in eq 8, can be directly compared to experiment.

2.2.2. Modified Fuzzy Gustafson–Kessel Clustering Algorithm. To be used as predictive tools, clustering algorithms should be able to identify the underlying clustering structure of a data set. Most of the popular distance-based clustering algorithms such as K-means or fuzzy C-means (FCM) have several drawbacks which prevent direct use on MD simulation data.

First, the FCM and K-means algorithms work well only for spherical-shaped clusters because their objective function is based on the Euclidian distance between samples and cluster centers. There are many algorithms to solve this problem, such as the Gustafson–Kessel algorithm (see below). However, shape-aware distance-based algorithms are still not appropriate for MD data because they tend to assign improper membership.
to members of small clusters in the vicinity of large clusters and therefore distinguish poorly between clusters that have the same size but different densities.

Second, proper classification requires the number of clusters to be known in advance or to be estimated directly from data set before the actual clustering begins. Approaches aiming to solve this problem, which involve simple, explicit testing of several cluster models that are compared to each other using a predefined quality measure, include the well-known Xie–Beni index,\textsuperscript{83} Arlot’s v-fold penalization algorithm,\textsuperscript{84} the gap clustering method,\textsuperscript{85} the resampling algorithm,\textsuperscript{86} and many others.

Third, clustering algorithms are usually employed as inference methods to determine the number \( M \), the centers \( c_1, \ldots, c_M \), and the shapes of clusters from the data. In the present context, however, we need to determine populations for the particular set of \( M_{\text{arch}} \) macrostates assigned by experiment. These macrostates are centered around special positions \( a_{\phi_1}, \ldots, a_{\phi_M} \) in the CV space, which we call archetypes. For example, for dipeptides in solution (see ref 23), the archetypes are the centers of the CS, P\( \psi \), and \( \alpha_R \) regions in the \((\phi, \psi)\) plane (defined following ref 29) and are given in Table 1 and are shown in Figure 2a. For NA-T-MA in the gas phase, the experimental macrostates of interest are A, B, and C and the archetypes are the centers of the corresponding regions in the 4-dimensional space \((\phi, \psi, \chi_1, \chi_2)\) as indicated in Table 2.

![Figure 2c shows an example of clustering using the modified FGK method with archetypes C5, P\( \psi \), and \( \alpha_R \) as given in Table 1 for NA-A-MA in solution with the OPLS-AA/L FF. The populations in this case are 17%, 62%, and 21%, respectively. To verify the robustness of our clustering approach with respect to the particular definition of the archetypes, we repeated the clustering analysis 300 times using archetypes randomly perturbed within half of the cluster radius from the original archetypes. Figure 4 shows a close-up in the upper left corner of the Ramachandran plane and a few representative perturbed archetypes for the CS, P\( \psi \), and \( \alpha_R \) states. The crisp centers obtained after clustering with perturbed archetypes and the original archetypes are almost identical, with maximum deviations within 5°. The corresponding populations all fall within 2% of the populations obtained with the original archetypes (Figure 4 and Table S2, Supporting Information). This demonstrates that our calculated populations are robust with respect to the choice of the archetypes.

### 3. RESULTS AND DISCUSSION

For NA-A-MA and NA-T-MA with the six FFs, CHARMM27 (C27), CHARMM36 (C36), AMBER99SB (A99SB), FUJI, OPLS-AA/L, and GROMOS 54a7 (G54a7), we ran dA-FEED simulations as described above. The resulting FESs, similar to those of Figure 2b, are shown in Figures S1–S3 of the Supporting Information. For NA-T-MA, we show projections of the four-dimensional FESs in the \((\phi, \psi)\) plane and in the \((\chi_1, \chi_2)\) planes. Overall, we find good qualitative agreement with published FESs of dipeptides in solution\textsuperscript{12,13,64,66,69,72} and in the gas phase.\textsuperscript{65} We also verified that the FESs obtained with either Gromacs 4.5.5\textsuperscript{68} using the modified PLUMED plugin\textsuperscript{69} or PINY_MD\textsuperscript{90} for the same systems exhibit negligible differences (data not shown). The Supporting Information also provides tables with detailed information on free energy minima for each system and each FF.

In the present work, however, the focus is not on the FESs themselves, but on macrostate populations that can be

| Table 2. Archetype Conformations of NA-T-MA in the Gas Phase and Their Ramachandran and \( \chi_1, \chi_2 \) Backbone Dihedral Angles\textsuperscript{44} |
|-----------------------------------------------|--------|--------|--------|--------|--------|
| macrostate | conformation | \( \phi \) | \( \psi \) | \( \chi_1 \) | \( \chi_2 \) | exp population |
| A           | CS(\( \Lambda\Phi \)) | -65     | 160    | -160   | 80     | 0.20  |
| B           | CS(\( \alpha R \Phi \)) | -150    | 160    | -160   | -80    | 0.50  |
| C           | G7\( \Phi \Phi \Phi \) | -60     | 60     | 80     | 80     | 0.30  |

*Experimental populations are indicated according to ref 4.*

To obtain population estimates comparable to experimental, our analysis method needs to assign a cluster to each of the archetypes, even if the standard clustering algorithm would not spontaneously place a cluster in this region. Therefore, we must introduce the archetypes as a priori information into the clustering scheme and enforce the presence of cluster centers \( v_1, \ldots, v_{M_{\text{arch}}} \) in the vicinity of the archetypes. The exact position, shape, and total number of clusters is still freely determined by the algorithm. With this modification, the clustering algorithm is no longer strictly an inference method but is rather a probe to interrogate the data on populations at the predefined locations of the archetypes.

To address the three points raised above, we use the modified fuzzy Gustafson–Kessel (FGK) algorithm,\textsuperscript{82} proposed in ref 87, which takes into account both the distance to cluster centers and the local density of samples.\textsuperscript{87} To enforce the presence of a cluster center in the vicinity of each archetype, an extra term is added to the penalty function. Finally, the adapted FGK is coupled to an iterative optimization scheme that finds the optimal number of clusters for each particular data set. The details of our clustering algorithm are described in Appendix B.

The macrostate populations obtained from our method should not depend on local variations of the archetype positions, the choice of which involves a large degree of arbitrariness. Figure 2c shows an example of clustering using the modified FGK method with archetypes C5, P\( \psi \), and \( \alpha_R \) as given in Table 1 for NA-A-MA in solution with the OPLS-AA/L FF.
compared to experiment. We applied the modified FGK clustering algorithm described in Appendix B to automatically identify conformational macrostates from simulation data, as shown, for example, in Figure 2b. In the Supporting Information, Figures S1–S3 similarly show the shape of the fuzzy clusters in the ($\phi$, $\psi$) plane for all systems in this study. On the basis of the fuzzy clusters, we calculated populations for the experimentally observed macrostates C5, PII, and $\alpha_R$ for peptides in solution, or A, B, and C for NA-T-MA in the gas phase using eq 8. The resulting populations are represented in Figures 5 and 6, and the corresponding numerical values can be found in Tables S2, S9, and S16 of the Supporting Information.

In the two following sections, we will discuss the results in detail for each system.

3.1. Dipeptides in Solution. We found our dAFED FESs to be in good agreement with the metadynamics FESs published by Vymetal et al.11,12 for NA-A-MA in solution with AMBER99SB, OPLS-AA/L, and CHARMM27. For CHARMM27, our dAFED free energy minima correspond well to those tabulated in ref 64 using other methods. Our FES with OPLS-AA/L is also in good agreement with that reported by Cruz et al.72 For all systems in solution, all FESs give FESs that are qualitatively similar. Notable differences are that the FUJI and GROMOS 54a7 FESs do not exhibit a local minimum in the $\alpha'$ region and show a less pronounced $\alpha_L$ basin. Instead of the $\alpha_L$ state, GROMOS 54a7 favors the neighboring C7eq state, which is almost absent in all other FESs. For NA-A-MA in solution, the experiments of Grdadolnik et al.23 found the largest population of backbone conformations to be in the PII region with minor contributions of the C5 macrostate and only less than 10% population in the $\alpha_R$ conformation. As shown in Figure 5a, our simulations confirm this trend only for the CHARMM36, FUJI, and GROMOS 54a7
FFs. CHARMM27 strongly overestimates the \( \alpha_8 \) population, denoting an exaggerated helical propensity for alanine, a tendency that has been noted in earlier studies.\(^{20,91} \) AMBER99SB overestimates the population of the C5 state, a problem that is partially addressed by the FUJI corrections to the AMBER99SB dihedral parameters.

The summary of calculated populations of NA-T-MA in solution and the comparison to experiment is shown in Figure 5b. Experimentally, the effect of the tryptophan side chain on the backbone conformations in NA-T-MA translates into a stabilization of the C5 macrostate at the expense of both P1 and \( \alpha_8 \). AMBER99SB largely overemphasizes this trend, with a dominating C5 conformer. On the other hand, CHARMM27 overestimates the \( \alpha_8 \) macrostate, as in the case of NA-A-MA. CHARMM36 partially corrects this defect and restores the right ordering of macrostates, although \( \alpha_8 \) is still over-represented. The GROMOS 54a7 populations are only marginally influenced by the bulky side chain, compared to NA-A-MA. Among all FFs, FUJI and OPLS-AA/L most accurately reproduce the experimental populations of NA-T-MA conformers. Overall, we note that all FFs tend to overestimate the \( \alpha_8 \) conformation with respect to the extremely low population observed experimentally.

We emphasize again that the populations shown in Figure 5 do not depend on prior definitions of arbitrary boundaries between macrostates, which can artificially skew the populations. For example, Vymetal et al.\(^{12} \) placed the separation between macrostates, which can arti
don't depend on prior de

The populations of the A, B, or C macrostates given by the clustering method and weighted summation using eq 7 are shown in Table S16 (Supporting Information) and illustrated in Figure 6, together with deviations from the experimental values.\(^{3} \) For all FFs, populations rank in the order B > C > A, which seems to contradict the ordering of the corresponding free energy minima mentioned above. For example, CHARMM36 free energy minima are at 0.0, 2.18, and 1.66 kcal/mol for B, C, and A, respectively. However, the resolution of this apparent paradox lies in the sizes of the clusters. Indeed, the C macrostate is consistently more voluminous than the A macrostate, as shown in Table 3. In the example of CHARMM36, the radius of the C cluster is 65% larger than that of the A cluster, which amounts to a more than 7 times larger volume for a four-dimensional spherical cluster. This makes C entropically more stable than A, which inverts the ranking on the basis of the free energy minima alone. This is an illustration of the fact that the value of the free energy at the FES minimum is not a suitable indicator to rank macrostates.

Figure 6b shows the aggregated prediction score of each FF for NA-T-MA in the gas phase, calculated according to eq 10. Overall, the behavior of all FFs is much more consistent in the gas phase than in solution, all FFs predicting qualitatively the same ranking of states A, B, and C. For NA-T-MA in the gas phase, the least accurate FFs are the ones of the former generation, CHARMM27 and AMBER-99SB. OPLS-AA/L gives the most faithful reproduction of experimental values, followed by the FUJI, which shows that the reparametrization\(^{60} \) improves the FFs’ performance.

### 3.2. NA-T-MA in the Gas Phase

Projections of the FESs from our simulations are shown in Figure S3 in the Supporting Information. The free energy minima in the four-dimensional CV space were assigned to four-letter conformers and to A, B, or C macrostates according to Table 2. Detailed information about free energy minima on the four-dimensional FES are given in Tables S10–S15 in the Supporting Information. We note that the FES minima for the A, B, or C macrostates consistently appear in the order B < A < C for all FFs, as shown in Table 3 as well.

One interesting observation is that, consistently among all force fields, the \( C_{7eq}^{\alpha} \) conformer with lowest free energy is \( C_{7eq}^{\beta} \) and not \( C_{7eq}^{\gamma} \). The latter was assigned by Dian et al.\(^{2} \) to the C state because of spectroscopic considerations and because \( C_{7eq}^{\gamma} \) was found to correspond to a lower potential energy minimum. The same authors, however, found a high free energy for the \( C_{7eq}^{\gamma} \) in a normal mode calculation based on DFT within the B3LYP approximation but dismissed that value, invoking the lack of accuracy of the normal mode approach. The fact that our enhanced-sampling calculations unambiguously confirm \( C_{7eq}^{\beta} \) to be lower in free energy than \( C_{7eq}^{\gamma} \) reopens the case for the assignment of the C state. In the following, however, we stick to the assignment of states A, B, and C proposed by Dian et al.\(^{2,3} \) to preserve consistency of the definitions.

### Table 3. Characteristic Cluster Radii As Defined by Eq B.4 for NA-T-MA in a Gas Phase vs Calculated FE Minima

<table>
<thead>
<tr>
<th>FF</th>
<th>A (deg)</th>
<th>B (deg)</th>
<th>( \alpha_8 ) (deg)</th>
<th>C (deg)</th>
<th>A (kcal/mol)</th>
<th>B (kcal/mol)</th>
<th>( \alpha_8 ) (kcal/mol)</th>
<th>C (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHARMM27</td>
<td>10.09</td>
<td>9.53</td>
<td>12.76</td>
<td>4.69</td>
<td>3.15</td>
<td>8.10</td>
<td>7.66</td>
<td>8.57</td>
</tr>
<tr>
<td>CHARMM36</td>
<td>2.33</td>
<td>2.26</td>
<td>4.23</td>
<td>4.90</td>
<td>4.19</td>
<td>8.12</td>
<td>2.32</td>
<td>3.84</td>
</tr>
<tr>
<td>AMBER99SB</td>
<td>2.0</td>
<td>9.53</td>
<td>12.76</td>
<td>4.69</td>
<td>3.15</td>
<td>8.10</td>
<td>7.66</td>
<td>8.57</td>
</tr>
</tbody>
</table>

Macrostate C, although higher on the FES, is systematically larger than macrostate A, resulting in populations ranking B > C > A (Figure 6).
of the AMBER99SB dihedral angles brings significant improvement also in the gas phase.

4. CONCLUSIONS

MD and enhanced sampling simulations of protein systems rely on the ability of the FF to faithfully model the conformational propensities of amino acids, both in the folded and in the disordered states. Recent experiments have been able to report the conformational preferences of dipeptides in solution or in the gas phase that can be used as direct benchmarks to evaluate the performance of FFs. We performed enhanced sampling MD simulations of the NA-A-MA and NA-T-MA peptides in solution and of NA-T-MA in the gas phase with six recent and widely used FFs, CHARMM27, CHARMM36, AMBER99SB, FUJI, OPLS-AA/L, and GROMOS 54a7. We employed a fuzzy clustering approach to map macrostates from the simulation data. The algorithm removes the requirement to define arbitrary macrostate boundaries and provides an optimal state partitioning given the simulation data. For the macrostates corresponding to experimental observables we derived populations that we used to assess FF accuracy.

From the methodological point of view, we showed that using the free energy values at the FES minima are not appropriate to rank conformational macrostates. Instead, the free energy of the entire macrostate should be evaluated, as we have seen with the ordering of the A, B, and C states in the gas phase. In addition, we showed that defining arbitrary hard state boundaries to calculate populations can lead to ranking inversions in the case of close macrostates separated by a low free energy barrier. We note that the present approach relies on macrostate assignments provided in the experimental studies and can thus be seen as somewhat indirect. In future research, it would be interesting to bypass the macrostate assignment step altogether and compare the simulated ensemble of conformations directly to the experimental spectra. This would, however, require a model that can reliably predict Raman or IR spectra from a given molecular conformation. Alternatively, a more direct comparison could be made on the basis of the $J$-couplings, which are simple functions of the torsion angles.

For dipeptides in solution, FFs show a great degree of variability in the predicted conformational populations, in particular in NA-T-MA where the bulky tryptophan side chain has a notable influence on backbone conformation. On the other hand, in the gas phase all FFs consistently reproduce the correct macrostate ranking and obtain a larger prediction score, which is surprising given the fact that these FFs were originally parametrized in solution. Overall, taking into account solution and gas phase results, we find that within the CHARMM and AMBER families the recent CHARMM36 and FUJI reparameterizations bring significant improvements over previous versions. The FUJI and OPLS-AA/L FFs show the best performance on the systems studied here. These two cases are interesting because they show that relying on very high-level QM calculations for parametrizations is advantageous. They also show that there is still room for improvement within the standard FF potential energy function without necessarily introducing additional complexity through terms such as the CMAP correction.

APPENDIX A. COMPUTATIONAL DETAILS

Model dipeptides, NA-A-MA and NA-T-MA, were built and geometrically optimized with SPARTAN08. Simulations with the CHARMM27, CHARMM36, AMBER99SB, and OPLS-AA/L FFs were carried out using the PYN_MG program. Additional simulations with CHARMM27, CHARMM36, FUJI, and GROMOS 54a7 were carried out with the GROMACS 4.5.5 software and the PLUMED 1.3 free energy plugin. The FF input files for CHARMM36 and FUJI were obtained personally from the main developers of the FFs. The input files for GROMOS 54a7 were downloaded from the Automatic Topology Builder website of A. E. Mark et al. (http://compbio.biosci.uq.edu.au/ath/index.py?tab=forceField_tab). Due to the different implementations, there might be slight differences compared to these FFs in their respective original software package. Both in the gas phase and in solvent. The collective variables for this study are the two backbone dihedral angles $\phi$ and $\psi$ and, for NA-T-MA, two side-chain dihedral angles, $\chi_1$ and $\chi_2$, as indicated in Figure 1. Following the dAFED scheme, high masses and temperature were assigned to the extended variables. In all simulations we chose uniform masses, $m_\phi = m_\psi = m_\chi_1 = m_\chi_2$, which we call $m$. For NA-A-MA, we chose $T_s = 600$ K and $m = 400$ amu-deg$^{-2}$Å$^{-2}$. For the other peptides, we chose $T_s = 800$ K and $m = 600$ amu-deg$^{-2}$Å$^{-2}$. The harmonic coupling constant was $5.4 \times 10^3$ (kcal/mol)/deg$^2$ for all simulations. These choices for the dAFED parameters were shown in to yield accurate FESs in our previous work on these systems.

Peptides in the gas phase were equilibrated for 500 ps at $T = 300$ K for NA-A-MA and $T = 423$ K for NA-T-MA, which corresponds to the temperature at which the IR–UV spectroscopy measurements were conducted. For the solution phase simulations in PINY_MD, each peptide was solvated in a periodic cubic box of 860 TIP3P water molecules. The systems were equilibrated at constant volume (NVT) for 200 ps, then at constant pressure (NPT) for 50 ps, and again at NVT for 500 ps. The SHAKE93 and RATTLE94 algorithms were used to constrain the geometry of water molecules to a relative tolerance of $10^{-6}$. The electrostatics were treated with the particle mesh Ewald summation technique, and the short-range forces were switched off at 12 Å. In PINY_MD, the NVT ensemble was generated by coupling a generalized Gaussian moment thermostat (GGMT) to each degree of freedom in the system (massive thermostatting). The equations of motion were integrated using the multiple time scale r-RESPA algorithm to exploit the various time scales. Three r-RESPA time steps were used: all intramolecular forces (covalent and ionic) were integrated using five steps per full MD step, and the dAFED harmonic coupling term was calculated using five steps per intramolecular force step. Production runs were up to 100 ns, with an outermost r-RESPA time step of 1 fs. Atomic coordinates were saved every 100 steps.

Simulations in Gromacs used similar specifications, except for some parameters due to the differences in implementation. In particular, no massive thermostatting is available in Gromacs. Instead, molecules were assigned to 204 groups, which were thermostated independently with Nosé–Hoover chains. In addition, no RESP/ scheme was used, but all bond lengths were constrained and a time step of 2 fs was used. The dAFED extended variables were coupled to a generalized Gaussian moment thermostat.

The FESs were obtained by calculating a normalized probability distribution function $P(s)$ from multidimensional nonuniform histograms with bin sizes ranging from 9$^\circ$ to 2.3$^\circ$. The two-dimensional surfaces were smoothed using a modified
local polynomial regression filter. We first attempted to use the classical Nadaraya—Watson kernel, but it produced significant shrinkage of the smoothed surface due to nonuniform data mesh. That is why an algorithm that works like a low bypass filter and does not produce shrinkage has been developed and implemented.

To estimate statistical uncertainties in our FES and macrostate populations, we used the nonparametric bootstrap bias corrected and accelerated (BCA) method with 2000 bootstrap samples per trajectory. Tables S17–S19 in the Supporting Information show the standard error (SE) and BCA 95% confidence intervals (CIs) for all macrostate free energies and populations reported here. For these populations, the CIs are in most cases within ±2 percentage points of the reported values in most cases.

**APPENDIX B. CLUSTERING ALGORITHM**

**B.1. Modified Fuzzy Gustafson–Kessel Clustering Algorithm**

Our clustering algorithm is based on the original fuzzy Gustafson–Kessel (FGK) algorithm. Our modifications seek to address the problems of (1) properly distinguishing between clusters with similar sizes but different densities, (2) robust discrimination of a small cluster corresponding to a low-populated experimental macrostate when it lies in close proximity to a larger one, and (3) maximum density in a vicinity of a cluster center. The original FGK algorithm is based on a fuzzy covariance matrix of each cluster corresponding to a low-dimensional feature space is

\[ d_k^2 = \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^m (s_i - v_k) (s_i - v_k)^T \]  

(B.1)

where the \( u_{ik} \) are elements of the membership matrix \( U \) and \( m \) is the “fuzzification” factor. The associated (squared) Mahalanobis distance in the \( n \)-dimensional feature space is

\[ d_k^2 \equiv d^2(s, v_k) = ||P_k^{1/2}(s_i - v_k)^T P_k^{1/2}||^2 = (s_i - v_k)^T \]  

(B.2)

If instead of \( P_k \) one chooses the identity matrix, the distance becomes the Euclidean distance. In the original FGK algorithm, the objective function is

\[ J(X, U, V) = \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^m d_k^2 \]  

(B.3)

Here \( X = (s_1, ..., s_M) \) is a set of samples, and \( V \) is the list of crisp centers (cluster prototypes).

In the modified FGK algorithm, the crisp centers are replaced by crisp volumes as suggested in ref 87, such that the samples that are very close to a cluster center automatically get membership 1. The volume of a cluster is defined as the determinant of the covariance matrix given by eq B.1. The effective cluster radius corresponding to this (nonspherical) volume is

\[ R_k = \sqrt{|P_k|^{1/n}} \]  

(B.4)

We define the modified Mahalanobis distance

\[ \tilde{d}_k^2 = \max(0, d_k^2 - R_k^2) \]  

(B.5)

and the objective function for the modified FGK becomes

\[ \tilde{J}(X, U, V) = \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^m \tilde{d}_k^2 \]  

(B.6)

The second modification in the algorithm is to add a penalty function that maintains crisp centers close to the position of the \( M_{exp} \) experimentally observed macrostates (archetypes). We define the distance between crisp center \( v_k \) and archetype \( a_i \) as

\[ D_k^2 = ||v_k - a_i||^2 \]  

(B.7)

The modified objective function takes the form

\[ \tilde{J}(X, U, V, A) = \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^m \tilde{d}_k^2 + \kappa \sum_{k=1}^{M_{exp}} D_k^2 \]  

(B.8)

Here \( \kappa \) is the parameter that determines the importance of the archetypes in the penalty function, with respect to the data-dependent forces that drive the optimization of the cluster centers \( v_k \). Note that the clusters linked to archetypes can never disappear in the cluster number optimization procedure described below.

At each step the elements of the weighted membership matrix \( \tilde{U} \) are calculated by the formula

\[ \tilde{u}_{ik} = \left[ M \sum_{j=1}^{M} \left( \frac{\tilde{d}_{kj}^2}{d_{ij}^2} \right)^{1/(m-1)} \right]^{-1} \]  

(B.9)

**B.2. Optimal Number of Clusters**

During the course of the modified FGK clustering algorithm above, we can merge clusters that overlap. As an overlap criterion, we calculate the similarity of \( S_{kl} \) values between any pair of clusters by pairwise comparison of their membership values,

\[ S_{kl} = \frac{\sum_{i=1}^{N} \min(\tilde{u}_{ik}, \tilde{u}_{il})}{\min(\sum_{i=1}^{N} \tilde{u}_{ik}, \sum_{i=1}^{N} \tilde{u}_{il})} \]  

(B.10)

Given a similarity threshold parameter \( t \), if \( S_{kl} < t \), we merge the clusters by summing the membership coefficients

\[ \tilde{u}_{ik} = \tilde{u}_{ik} + \tilde{u}_{il} \]  

(B.11)

In our approach, we optimize the number of crisp centers by an iterative procedure that starts with different initial numbers of clusters. We then assess the final result of the modified FGK algorithm above using a modified version of the Xie–Beni index. The original Xie–Beni index is defined as

\[ XB = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^m \tilde{d}_k^2 \]  

\[ \min_{k,l} (d_{kl}^2) \]  

(B.12)

For clusters with crisp volumes of various sizes generated by the modified FGK algorithm of eq B.1, the denominator should depend not only on cluster centers but also on variability of a whole cluster CV. To incorporate variability, we use the average distance from the center as a metric for cluster variability.

\[ CV_k = \frac{\sum_{i=1}^{N} \tilde{u}_{ik} d_k^2}{\sum_{i=1}^{N} \tilde{u}_{ik}} \]  

for \( k = 1, ..., M \)  

(B.13)

Then the dissimilarity between two clusters \( k \) and \( l \) is

\[ \Delta_{kl} = \frac{d_{kl}^2}{CV_k \times CV_l} \]  

(B.14)

Then the modified Xie–Beni quality criterion becomes

\[ X_{\text{B}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^m \tilde{d}_k^2 \]  

\[ \sum_{k=1}^{M} \Delta_{kl} \]  

(B.15)
The simplest approach to find the optimal number of clusters is just to test systematically the quality of clusterings with increasing/decreasing number of cluster centers. Instead, we use the following search algorithm. We start with the $M_{\text{up}}$ archetypes at their initial position, and we add additional randomly generated cluster centers to reach a population of $M$ clusters, where $M > 4$. Then we generate four extra cluster populations $M_1, \ldots, M_4$ with the formula:

$$M_i = 2M \quad M_2 = M + 1 \quad M_3 = M - 1 \quad M_4 = M/2$$

(B.16)

We optimize each population using the modified GFK algorithm described in eq B.1, pick the cluster population with largest value of the modified index $XB$, and substitute $M$ with this number. The procedure repeats until the number of crisp centers does not change.

### ASSOCIATED CONTENT

#### Supporting Information

Tables are given that include quantitative information about the position and height of PES minima for each of the systems considered in this study: numerical values for the macrostate populations obtained from the clustering approach and eq 8 are represented in Figures 5 and 6; the statistical uncertainty is related to both PES minima and macrostate populations for each system. Figures showing PESs and clusters obtained for all systems and force fields are represented in the middle and lower panels of Figure 2. This material is available free of charge via the Internet at http://pubs.acs.org.

### AUTHOR INFORMATION

**Corresponding Author**

*M. E. Tuckerman; e-mail, mark.tuckerman@nyu.edu.

**Author Contributions**

These authors contributed equally to this work.

**Notes**

The authors declare no competing financial interest.

### ACKNOWLEDGMENTS

We are grateful to Hideaki Fujitani for providing the Gromacs input files for the FUJI force field. We also thank Robert Best for sending us the Gromacs input files for the CHARMM36 force field. Some of the computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. M.A.C. acknowledges support from the Swiss National Science Foundation grant PA00P2_129092 for part of this work. M.E.T. acknowledges support from the US National Science Foundation grants CHE-1012545 and CHE-1301314.

### ADDITIONAL NOTE

“Note that Figure 7 (middle) of ref 64 is in fact the FES for Charmm22, not forCHARMM27 as reported.

### REFERENCES


