From Pollen Grains to Peptides: Coarse-Graining and the Brownian Motion of Molecules

by

JESSE HALL

A dissertation accepted and approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Physics

Dissertation Committee:

Eric Corwin, Chair

Marina Guenza, Advisor

Dan Steck, Core Member

Jayson Paulose, Core Member

Leif Karlstrom, Institutional Representative

University of Oregon
Spring 2025

Dissertation Abstract

Jesse Hall

Doctor of Philosophy in Physics

Title: From Pollen Grains to Peptides: Coarse-Graining and the Brownian Motion of Molecules

The thermally driven fluctuations in shape and orientation of molecules in solvent are of immense interest in biology, chemistry, and medical applications. Studying such complex systems, often involving hundreds of thousands of atoms, pushes the boundaries of computational science. Coarse-graining is a theoretical approach that attempts to reduce the complexity of these systems so that they can be studied more intuitively and simulated faster, ideally without sacrificing statistical accuracy.

Highly accurate and rigorous coarse-graining techniques exist for low-dimensional systems of restricted symmetry, such as the free diffusion of a pollen grain on the surface of water or the bound diffusion of an internal reaction coordinate of a molecule, but high-dimensional site-specific models with mixed bound and unbound structure typically rely on more heuristic approaches.

Here we present new rigorous techniques that allow for the parametrization of highly versatile site-specific coarse-grained models of diffusing molecules, showing how the friction coefficients and potential of mean force can be constructed in a manner consistent with both the internal fluctuations in the molecule's shape and its overall translation and rotation.

This dissertation includes previous published co-authored material.

Acknowledgements

I would like to thank my advisor Marina Guenza, for putting up with my stubbornness over the years and for always being in my corner.

I would like to thank my friend, Bo Ward, for coffee and inspiration when I thought I was done with physics.

I would like to thank my dad, Buz Davy Morris, for showing me how to be kind, and patient, and true.

I would like to thank my mom, Sandra Jean Morris, for sharing her love of books with me, and for showing me how to be strong.

I would like to thank my wife, Mackenzie Leigh Ní Flainn, for growing old with me, for showing me how to be brave, and for all the beauty in the world we've seen together.

And I would like to thank my son, Aidan Elias Rón Ó Flainn, for bringing us all so many smiles in the home stretch.

Contents

List of Figures

1 - INTRODUCTION

I. BROWNIAN MOTION

In 1827, the British botanist Robert Brown conducted a series of observations of pollen grains immersed in water under a one-lens microscope. He became interested in the behavior of the much smaller ejecta particles that came out of the pollen grains themselves when they burst. These were comprised of starch and lipid organelles whose function was not known at the time. He observed that these particles underwent ceaseless random motion in the water. He had no explanation for this motion, and suggested that the particles may be the fundamental constituents of organic matter [1].

Brown later abandoned this hypothesis upon further experimentation, as he found that any matter, organic or inorganic, when reduced to a fine enough powder, moved in much the same manner upon the surface of water. Other microscopists before and after also saw such motion of many kinds of fine matter, but it was Brown's investigations that became the most well-known. Brown was, in the end, unable to account for the motion that came to be named after him.

It would be Albert Einstein, as part of his thesis work published in 1905 [2], who would develop a convincing quantitative model of the motion and its mechanism. At this point the molecular theory of matter was well-known and supported by various lines of evidence, but was still not universally accepted.

Using hydrodynamic and statistical mechanical arguments, Einstein derived a simple relationship between the diffusion coefficient of a particle buffeted by the thermal fluctuations of a solvent and the solvent's viscosity. Today this relationship is known as the Einstein relation:

$$D = \frac{k_B T}{\zeta}. \tag{1}$$

where D is the diffusion constant, ζ the friction coefficient of the brownian particle in its solvent, k_B is the Boltzmann constant, and T the temperature.

This equation accurately modeled the evolution of the mean-squared displacement, with notable experimental confirmation by Jean Baptiste Perrin.

Paul Langevin developed a dynamical description of Brownian motion using a stochastic differential equation [3]

$$m\dot{\vec{v}} = -\zeta\vec{v} + \vec{f}^R(t), \tag{2}$$

where $\vec{v}$ is the velocity of the particle and $\vec{f}^R(t)$ is a random force with zero mean and a

second moment given by

$$\langle f_\alpha^R(t) f_\beta^R(t') \rangle = 2k_B T \zeta \delta_{\alpha\beta} \delta(t - t'). \tag{3}$$

With these developments and many more beyond the scope of this brief introduction, the theory of the Brownian motion of particles was rapidly assembled in the early 20th century. We note that these two equations together could be employed for a rudimentary coarse-graining procedure. If a particle's mass is known, the friction coefficient could be discerned by observing its diffusive statistics on the water's surface, and this parameter then used in the dynamical Langevin equation to model its motion without explicitly modeling the water.

We now turn to the later development, in the 1960s, of a more sophisticated formalism that connected these equations in a much more general manner to the foundations of classical mechanical theory.

II. THE PROJECTION OPERATOR FORMALISM

A rigorous formalism exists for the project of coarse-graining. Beginning from a many-body Hamiltonian, a smaller set of coarse-grained variables can be defined and the complete equations of motion projected onto these variables. This procedure was developed in the 1960s and is known as the Mori-Zwanzig projection operator formalism after Hajime Mori and Robert Zwanzig [4, 5]. The formalism provides rigorous theoretical justification for the use of Langevin equations, as the result of the procedure generically yields a friction-like term proportional to the velocity of the CG coordinates and a noise-like term that depends only on the initial conditions of the system and whose statistics are consistent with the Fluctuation-Dissipation Theorem.

The friction term is in general more complex than the heuristic formulas discussed in the previous section, but it does reduce to them in broad and well-defined regimes.

There exist several choices of projection method which yield different partitions of the coarse-grained forces into a conservative part, a dissipative part, and a "noise" that originates from the initial conditions of the complete system [6]. All of these projections are in principle exact, but given that the noise must be approximated as a random variable, it is desirable for it to have certain properties that not all projections yield. For example, the Mori projection operator always yields a harmonic conservative force, and any effects due to nonlinearity in the system are buried within the noise term. This is often highly undesirable, as it provides no obvious way to parametrize the complex structure of the nonlinearity within the noise. In contrast, the Zwanzig projection operator allows for nonlinear conservative CG forces that correspond precisely to the potential of mean force. Since diffusing molecule maintain reasonably well-defined structure around a moving center of mass, their complete motion

cannot be adequately described by a harmonic potential. As such, it is the Zwanzig operator that we will employ throughout this dissertation.

We will briefly review the principal results of the theory here, largely following the derivation in [7].

Given a Hamiltonian $\mathcal{H}(\vec{z})$ over a fine-grained phase space $\vec{z}$, there is an associated Liouville operator L such that, for any function $\vec{A}(\vec{z}(t))$,

$$\frac{d\vec{A}}{dt} = L\vec{A}(\vec{z}(t)) \tag{4}$$

If $\vec{A}$ is taken to be a collection of one or more coarse-grained variables of interest (such as a reaction coordinate), then the goal of projection operator formalism is to define a projection operator $P_{\vec{\alpha}}$ onto a specific CG state $\vec{\alpha}$ and its orthogonal operator $Q_{\vec{\alpha}} = 1 - P_{\vec{\alpha}}$ such that the right-hand side may be partitioned into projected and orthogonal dynamical forces.

The Zwanzig projection operator mentioned earlier is defined as:

$$P_\alpha F(\vec{z}) := \frac{1}{\Omega(\vec{\alpha})} \int d\vec{z}\, F(\vec{z}) \rho_{eq}(\vec{z}) \delta(\vec{A}(\vec{z}) - \vec{\alpha}) \tag{5}$$

where

$$\Omega(\vec{\alpha}) := \int d\vec{z}\, \rho_{eq}(\vec{z}) \delta(\vec{A}(\vec{z}) - \vec{\alpha}). \tag{6}$$

and $\rho_{eq}(\vec{z})$ is the equilibrium phase-space probability distribution.

Making use of the evolution operator e^{tL} and inserting the identity $P_{\vec{\alpha}} + Q_{\vec{\alpha}} = 1$ into Eq. (4) to split the liouville operators, the exact equation of motion can be reorganized into terms that depend on projected or orthogonal coordinates respectively:

$$\dot{\vec{\alpha}} = P_{\vec{\alpha}} L\vec{\alpha} + T \int_0^t ds\, \Gamma(\alpha(\overrightarrow{t-s}), s) \frac{\partial S}{\partial \vec{\alpha}}(\vec{\alpha}(t-s)) + k_B T \int_0^t ds\, \vec{\nabla}_{\vec{\alpha}} \cdot \Gamma^T(\vec{\alpha}(t-s), s) + \vec{\tilde{R}}(\vec{z}, t). \tag{7}$$

where we have defined the following quantities:

$$\begin{aligned} \vec{\tilde{R}}(\vec{z}, t) &:= e^{tQ_{\vec{\alpha}}L} Q_{\vec{\alpha}} L\vec{A}, \\ \Gamma(\vec{\alpha}, t) &:= \frac{1}{k_B T} P_{\vec{\alpha}} \left(\vec{\tilde{R}}(\cdot, t) \vec{\tilde{R}}(\cdot, 0)^T \right) \\ S(\vec{\alpha}) &:= k_B \ln \Omega(\vec{\alpha}) \end{aligned} \tag{8}$$

The dot-arguments in the definition of Γ indicate fine-grained dependence that is projected out and hence does not propagate to Γ, which is a function only of the coarse-grained state and time.

The quantity S is an entropy arising from the numerous fine-grained states consistent with any given coarse-grained state. The quantity Γ is known as the memory function or kernel, and it will be seen in a moment to be intimately related to drag forces.

This is an abstract equation of motion for any unspecified set of CG variables. In this work our variables of interest will be the center-of-mass positions and momenta of distinct clusters of atoms, such that $\vec{\alpha} = (\vec{r}, \vec{p})$. Upon making this choice, Eq. (7) can be simplified to:

$$\begin{aligned}\dot{\vec{r}} &= M^{-1}\vec{p} \\ \dot{\vec{p}} &= -\frac{\partial V(\vec{r})}{\partial \vec{r}} - \int_0^t d\tau\, \Gamma(\vec{r}(\tau), \vec{p}(\tau), t-\tau)\vec{v}(\tau) \\ &\quad + k_B T \int_0^t d\tau \vec{\nabla}_{\vec{p}} \cdot \Gamma^T(\vec{r}(\tau), \vec{p}(\tau), t-\tau) + \vec{\tilde{R}}(\vec{z}, t)\end{aligned} \tag{9}$$

In Eq. (10) only the last term on the right retains any explicit dependence on the fine-grained phase space rather than the coarse-grained one. Crucially, since it is $Q_{\vec{\alpha}}-$projected, this term has no correlation with any of the CG variables under averaging conditioned on a given state. Given this property, and the fact that it's second moment is often known via the memory function as above, it is often modeled as a random variable $\vec{f}^R(t)$, turning Eq. (10) into a Generalized Langevin equation. These random "noise" forces are generally taken to be Gaussian-distributed, motivated by the central limit theorem and the fact that they are the sum of a great many interactions between the CG coordinates and the large "bath" of orthogonal coordinates.

$$\begin{aligned}\dot{\vec{r}} &= M^{-1}\vec{p} \\ \dot{\vec{p}} &= -\frac{\partial V(\vec{r})}{\partial \vec{r}} - \int_0^t d\tau\, \Gamma(\vec{r}(\tau), \vec{p}(\tau), t-\tau)\vec{v}(\tau) \\ &\quad + k_B T \int_0^t d\tau \vec{\nabla}_{\vec{p}} \cdot \Gamma^T(\vec{r}(\tau), \vec{p}(\tau), t-\tau) + \vec{f}^R(t)\end{aligned} \tag{10}$$

In these coordinates the role of the memory function becomes more clear. The second term on the right-hand side of Eq. (10) accumulates contributions proportional to and opposite the velocity (when Γ is positive) over the history of the trajectory, with the proportionality dependent on the time-lag and the state of the cluster at that time. This history-dependent scheme is more complicated than the more heuristic Stokes' drag described in the previous section, but ultimately still recognizable as a dissipative friction term.

The parametrization of state-dependent memory kernels is an active problem, and the current common practice is to omit the state-dependence and proceed with the state-independent GLE:

$$\begin{aligned}\dot{\vec{r}} &= M^{-1}\vec{p} \\ \dot{\vec{p}} &= -\frac{\partial V(\vec{r})}{\partial \vec{r}} - \int_0^t d\tau\, \Gamma(t-\tau)\vec{v}(\tau) + \vec{f}^R(t)\end{aligned} \tag{11}$$

It is this state-independent GLE that we will treat as the underlying dynamics of our coarse-grained variables in this dissertation. From here another approximation may be made: if the memory kernel decays in an integrable manner, then there exists always a timescale T_M beyond which $\Gamma(t)$ can be approximately treated as proportional to a Dirac delta, such that

$$\Gamma(t) = 2\zeta\delta(t). \tag{12}$$

In this case, Eq. (11) simplifies to

$$\begin{aligned}\dot{\vec{r}} &= M^{-1}\vec{p} \\ \dot{\vec{p}} &= -\frac{\partial V(\vec{r})}{\partial \vec{r}} - \zeta\vec{v}(t) + \vec{f}^R(t)\end{aligned} \tag{13}$$

Thus an object obeying Eq. (11) will, when resolved only on timescales larger than T_M, appear to evolve according to Eq. (13). The latter equation, no longer history-dependent, is known as a Markovian Langevin Equation, while Eq. (11) is fundamentally Nonmarkovian in character. In light of this T_M is referred to as the Markov timescale of the system. We see from Eq. (13) that the Stokes' drag is recovered from the exact projection operator formalism in the case that the memory function is not state-dependent, over timescales long enough that the memory cannot be resolved.

If the mean force term in Eq. (13) is omitted, we have a description of a free particle diffusing in the solvent, recalling the case of Robert Brown's pollen particles. Approximating even further by integrating out the effects of inertia over timescales larger than the ballistic motion yields an "overdamped" equation of motion

$$\zeta\dot{\vec{r}} = \vec{f}^R(t) \tag{14}$$

that generates the idealized Brownian motion of the stochastic systems literature.

The goal of our work is to develop methods to accurately parametrized Eq. (13) for many-bead models of molecules in solution. The development of these methods must occasionally acknowledge less approximate forms of the dynamics to maintain physical consistency. In these cases we will treat Eq. (11) as the "true" equation of motion governing the observed clusters, neglecting the possibility of state-dependence. Including time- and state-dependent memory is a goal of our future work, but we will see that new techniques needed to be

developed even in the Markovian case to handle the full complexity of translating, rotating, and fluctuating molecules in solution.

III. PARAMETRIZATION OF THE GENERALIZED LANGEVIN EQUATION

While the projection operator formalism for coarse-graining was developed decades ago, the difficulty of accurately parameterizing GLEs for a given system remains with us today. For the study of biological molecules, the desire is to project out the detailed motion of most atoms in the solvent and indeed the molecule itself, replacing the intensive MD simulation with a Langevin equation for some reaction coordinate of interest. While the form of the resulting dynamical equation is known from the formalism as described in the previous section, direct calculation of the parameters from their definitions is highly intractable. In particular, the memory kernel is defined in terms of the orthogonal dynamics evolved by $e^{t\hat{Q}_{\vec{\alpha}}L}$, which are not even obtainable from simulated trajectories. In order to parametrize the GLE, then, additional relations between these quantities and observable statistics must be sought.

Much recent work has focused on using short MD simulations to gather statistics from the fine-grained system, which are then used to parametrize a CG Langevin equation [8–11]. While this does not circumvent the need for fine-grained simulations, it does complement them with an analytical model from which further insight can be gained using the same data. Additionally, enhanced sampling methods for recovering physical quantities from MD simulations are an area of great activity, so there remains the possibility that unenhanced fine-grained simulations could be bypassed entirely in the construction of a CG model by this approach.

Langevin equations with and without memory have been successfully paramatrized from simulations for the case of a tagged particle diffusion in solvent, using the Einstein relation or related Volterra equations (this procedure is described in detail in Chapter 1). This case is essentially a microscopic version of the pollen grain problem. Recent approaches have had success in accurately modeling single internal reaction coordinates in this way as well, developing accurate Langevin equations with and without memory, and even including some state-dependence of the memory function [8, 9].

However, studying a single reaction coordinate has disadvantages. While sophisticated techniques have been developed to identify good reaction coordinates, there is no guarantee that a given phenomenon of interest can be well described by a single coordinate. States may be connected by multiple reaction pathways. Couplings between specific coordinates may be of explicit interest, such as rotational-vibrational coupling during binding events. Additionally, any coordinates projected out of the GLE must nonetheless contribute to the

memory function, so low-dimensional models will generally have longer memory than high-dimensional models unless a fortuitious separation of timescales occurs.

In our work we have found that many of the methods employed to parametrize Langevin equations for single reaction coordinates do not scale well when applied to higher dimensional CG models. Indeed, some of these methods break down entirely once the translational and rotational freedom of the entire molecule are included explicitly in the CG model.

In this thesis, we have developed parametrization methods that accomplish several goals at once. First, they explicitly anticipate the translational and rotational symmetry of freely diffusing molecules, allowing these degrees of freedom to be incorporated into the resulting CG models. Second, by avoiding explicit construction of histograms, these methods scale much more effectively to the parametrization of high-dimensional coarse-grained models. This allows a single CG model to be constructed for, say, every atom in a small peptide, or every amino acid in a protein, rather than a single carefully chosen reaction coordinate, provided adequately sampled statistics are available.

Obtaining representative sampling from MD simulations is itself a considerable challenge and an area of active research, which lies outside the scope of this thesis. Our focus is on expanding the scope of what is possible with coarse-grained Langevin equations given that proper sampling is obtained. The many-bead models we describe here allow for considerable exploratory work to be done after the coarse-graining stage, or to study phenomenon which are not described well by a single coordinate, such as those involving multiple pathways. In particular, they preserve the translational and rotational symmetry of the molecule, permitting study of the interaction of global diffusion and conformational fluctuations. Such coupling may be of interest during binding events in particular.

The paper presented in Chapter 2 details a method for reconstructing Markovian friction coefficients consistent with an underlying memory function. The method involves a Generalized Einstein Relation (GER) that allows for the use of a broader class of correlation functions than previous methods. This allows some freedom in the choice of which lagged correlations will be used in the friction calculation. Such expressions for the friction generically express the memory function or friction coefficient matrix as a ratio of correlation matrices. This extra freedom turns out to be crucial in the case of diffusing molecules with internal degrees of freedom, as we show that the eigenvalue structure of previous expressions [10, 12] leads to numerically unstable ratios in the long time limit. In this paper, we demonstrate how the more general expression can be exploited to eliminate this source of instability and generate stable plateau values for the Markov friction. These friction coefficients are consistent with both the internal and global diffusion of the molecule.

In chapter 3, we turn to the construction of the mean force term in the GLE. Again novel difficulties arise in the case of systems undergoing free rotational diffusion. In this

paper we show that the free energy landscape over a molecule's internal shape coordinates is distorted by a shape-dependent rotational entropy. This distortion is a consequence of coupling between global rotation and internal shape coordinates (rovibrational coupling [13]) due to fictitious forces in the body-fixed frame. In order to construct a CG model with an accurate rotationally symmetric potential from these probability distributions, this entropic contribution must be removed to isolate the effects of the potential of mean force. We derive a formula for the shape-dependence of this entropy, and demonstrate how the CG force field can be corrected for it. Additionally, we show that this correction may be crucial for the subsequent calculation of friction coefficients using this force field, as the errors in lagged force correlations can propagate to significant errors at long time scales even when the instantaneous correction is small.

Together these two methods allow for the complete recovery of the Markovian regime of molecular systems obeying an exactly state-independent GLE, provided appropriate sampling can be obtained. Even in toy models, this is a novel capability that previous methods were unable to provide.

In chapter 4 we turn to a test of these methods on an MD simulation of a real molecule, alanine dipeptide, in explicit solvent. We show that in order to construct a many-bead model with explicit diffusion, the rotational entropy correction to the modeled force is critical, as the resulting friction coefficients return negative eigenvalues if the correction is not performed. Using the methods of chapters 2 and 3 in tandem, however, yields a CG model of alanine dipeptide whose position-position correlations show very good qualitative agreement with the MD simulation, capturing complex features over many cross-correlations. There is significant room for quantitative improvement, however, even at long time scales where the system should be Markovian, and we believe the remaining error has two likely causes. First, alanine dipeptide was chosen because it is a well-studied model system [8, 14, 15], but it has a significant energy barrier to a rare state, and our input MD simulations were unbiased. Enhanced sampling methods lay outside the scope of this work, but incorporating them into these new coarse-graining protocols is a priority of future work. Second, this coarse-graining approach still does not account for state-dependence, which is the only major feature of the full Zwanzig GLE that it does not account for in the Markovian regime. Recovering the full state- and time-dependence of the memory kernel is also a goal of future work.

---

[1] P. Pearle, B. Collett, K. Bart, D. Bilderback, D. Newman, and S. Samuels, American Journal of Physics **78.12**, 1278 (2010).

[2] A. Einstein, Ann. Phys. **19.4**, 371 (1905).

[3] P. Langevin, CR Acad. Sci. Paris **146**, 530 (1908).

[4] H. Mori, Prog. Theor. Phys., Japan **33**, 423 (1965).

[5] R. Zwanzig, (Oxford : New York : Oxford University Press, 2001).

[6] H. Vroylandt and P. Monmarché, J. Chem. Phys. **156**, 244105 (2022).

[7] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, Faraday Discuss. **144**, 301 (2010).

[8] H. S. Lee, S. Ahn, and E. F. Darve, J. Chem. Phys. **150**, 174113 (2019).

[9] P. Ge, Z. Zhang, and H. Lei, Phys. Rev. Lett. **133**, 077301 (2024).

[10] H. Lei, N. A. Baker, and X. Li, Proc. Natl. Acad. Sci. U.S.A **113**, 14183 (2016).

[11] B. Kowalik, J. O. Daldrop, J. Kappler, J. C. F. Schulz, A. Schlaich, and R. R. Netz, Phys. Rev. E **100**, 012126 (2019).

[12] F. Grogan, H. Lei, X. Li, and N. A. Baker, J. Chem. Phys. **418**, 109633 (2020).

[13] E. Bright Wilson Jr., J. C. Decius, and Paul C. Cross, (Dover, 1955).

[14] P. E. Smith, J. Chem. Phys. **111**, 5568–5579 (1999).

[15] J. Apostolakis, P. Ferrara, and A. Caflisch, J. Chem. Phys. **110**, 2099–2108 (1999).

2 - THE GENERALIZED EINSTEIN RELATION

The following is reproduced from *Generalized Einstein Relation for Markovian Friction Coefficients from Molecular Trajectories*, published in volume 134 of *Physical Review Letters* on April 7, 2025, and authored by J.M. Hall and M.G. Guenza.

Coarse-grained (CG) dynamical models of complex systems aim to explicitly describe a reduced set of quantities of interest while only implicitly modeling the remaining degrees of freedom. For example, a protein in solution may be described by the centers-of-mass of the constituent amino acid monomers, and its dynamics described by a Langevin equation, where the solvent and atomistic degrees of freedom are modeled as friction and noise. Such CG models have the potential to provide analytical insight into the mechanisms behind biological function of macromolecules on targeted length scales [1–4] as well as to increase computationally accessible timescales by projecting out high-frequency modes.

The Mori-Zwanzig projection operator method can be used to rigorously derive a Generalized Langevin Equation (GLE) for the time evolution of the CG variables consistent with an underlying atomistic Hamiltonian [5, 6]. Recent years have seen a growing interest in developing GLEs as a practical tool due to the modern accessibility of long trajectories via atomistic Molecular Dynamics (MD) simulations and new methods of parameterizing the equations of motion using statistical correlations from these trajectories [7–9].

Despite these advances and the growing interest in designing fast CG computational methods to simulate complex systems across multiple scales, parametrizing the GLE with high accuracy remains an ongoing challenge. A key task in designing a highly accurate GLE is the calculation of the friction coefficients entering the CG model [10–13]. In this manuscript, we present a solution for the friction coefficients in the Markovian regime by proposing an extended Einstein relation consistent with an underlying memory function. We evaluate the quality of our approach on a simple three-bead model, and compare the accuracy of the recovered friction with that of a conventional method involving the numerical inversion of a Volterra integral equation.

We adopt this simplified model with a known analytical, multidimensional free energy landscape, because the accuracy of the friction coefficient depends in our model on the accuracy of the conservative forces. Extensive literature on deriving diffusion coefficients (friction) from simulation trajectories [14] or experimental data [15–19] of more complex systems, such as protein folding, commonly identifies a single or a pair of reaction (collective) coordinates [20–22]. Rare events are typically analyzed along these coordinates using

enhanced sampling of the free energy landscape [23]. Testing our method on such complex systems would introduce uncertainties on the intramolecular forces, which would cloud the quality of our results [21, 22, 24]. Therefore, we rely on a simple three-bead model with known interactions. This model has a multidimensional free energy landscape, both bound and unbound coordinates, providing a simplified representation of the complexity inherent to the dynamics of a coarse-grained macromolecule.

The proposed model for calculating self and cross friction coefficients provides a significant advantage over traditional methods by allowing flexibility in the selection of vectors entering the Generalized Einstein Relation. This adaptability enables one to resolve instabilities in the friction coefficients by selecting appropriate coordinates.

In the Zwanzig-projected equation there are two functions that must be parameterized. One is the potential of mean force U, which can be obtained from equilibrium statistics. The other is the memory function Γ, which in general is a function of time and the CG positions and momenta [25]. In this work we take the common approximation that this kernel is state-independent, leaving only the time-dependent memory.

If the CG coordinates are taken to be the center-of-mass positions of n distinct clusters of atoms, then the state-independent form of the Zwanzig GLE for cluster i is given by

$$\begin{aligned}\dot{\vec{r}}_i(t) =& \frac{1}{m_i}\vec{p}_i(t) \\ \dot{\vec{p}}_i(t) =& -\frac{\partial U(\vec{r}(t))}{\partial \vec{r}_i} - \sum_j \int_0^t d\tau\, \Gamma_{ij}(t-\tau)\vec{v}_j(\tau) + \vec{f}_i^R(t),\end{aligned} \tag{1}$$

where $\vec{r}_i$ is the position of the i-th center of mass, $\vec{p}_i$ and $\vec{v}_i$ are the corresponding momentum and velocity, and $\vec{f}^R$ is a set of gaussian-distributed random forces whose statistics are related to Γ by the second fluctuation-dissipation theorem [26],

$$\langle f_{i\alpha}^R(t) f_{j\beta}^R(t')\rangle = k_B T \Gamma_{ij}(t-t')\delta_{\alpha\beta}, \tag{2}$$

where T is the temperature and k_B is the Boltzmann constant. The Greek indices indicate Cartesian components and the Kronecker delta reflects our assumption of an isotropic environment.

When the memory function is integrable, as is typical, this time-dependent kernel can be replaced by a constant friction coefficient to model kinetics over sufficiently long timescales, where the behavior is Markovian. This Markovian regime is of particular interest for its mathematical simplicity compared to the full GLE when accurate dynamics over the short-

est timescales are not crucial. This is often the case in the study of diffusive motion in macromolecules, where relaxation of the solvent and other local degrees of freedom can occur much faster than conformational changes of functional interest.

In the Markovian limit, the running integral $G(t)$ of the memory function should converge to a stable plateau yielding the friction coefficients

$$\zeta = \lim_{t\to\infty} G(t) = \lim_{t\to\infty} \int_0^t d\tau\, \Gamma(\tau). \tag{3}$$

The stability and accuracy of this plateau at long lag times is then crucial to developing Markovian CG models from trajectory data. This is especially so in the case of multidimensional models that possess a rich mode structure over a wide hierarchy of timescales. Correlations in the fastest coordinates may completely decay while others are still in a non-Markovian regime, and such coordinates may be coupled via off-diagonal friction through mechanisms such as hydrodynamic interactions. Such cases require an approach that is numerically reliable at long timescales despite the exponential decay of the input time correlation functions (TCFs).

One popular method to directly extract the memory kernel from MD trajectories involves the inversion of a Volterra integral equation of the first kind [11, 14]. The friction coefficients can then be obtained from the total integral of this memory function. We now show that this method can lead to instability in the friction coefficients and propose a new method that yields a far more accurate and stable plateau at long timescales.

By taking the dot product of the force equation in (1) with the initial velocity $\vec{v}_k(0)$ and then taking the mean, one obtains for a multi-particle model in an isotropic solvent [7]

$$M\dot{C}^{vv}(t) = -C^{vU'}(t) - \int_0^t dt\, \Gamma(t-\tau)C^{vv}(\tau), \tag{4}$$

where

$$C_{ij}^{xy}(t) := \langle \vec{x}_i(0) \cdot \vec{y}_j(t) \rangle \tag{5}$$

is the time-lagged correlation function between $\vec{x}_i$ and $\vec{y}_i$, and M is the diagonal matrix of cluster masses. As a projected CG coordinate, $\vec{v}_k$ is uncorrelated with the random noise $\vec{f}_i^R$ for all time lags (Appendix A). Integrating Eq. (4) over time yields

$$MC^{vv}(t) - MC^{vv}(0) = \mathcal{F}^{vU'}(t) - \int_0^t d\tau\, G(t-\tau)C^{vv}(\tau), \tag{6}$$

where $\mathcal{F}^{vU'}(t)$ is defined as

$$\mathcal{F}^{vU'}(t) = -\int_0^t d\tau\, C^{vU'}(\tau) \tag{7}$$

and $G(t)$ is the integral of the memory function, Eq.(3).

Conventionally, the integral in Eq. (6) is discretized in order to solve for $G(t)$. The trapezoid rule can be used for this discretization, but leads to spurious high-frequency oscillations in the solution which must be filtered out [27]. We use, instead, the left-handed rectangle rule, which after some rearranging transforms Eq. (6) into

$$\begin{aligned} G(t) =& \frac{1}{\Delta\tau}\left[\mathcal{F}^{vU'}(t) + MC^{vv}(0) - MC^{vv}(t)\right] C^{vv^{-1}}(0) \\ &+ \sum_{i=1}^{N-1} G(t - i\Delta\tau) C^{vv}(i\Delta\tau) C^{vv^{-1}}(0) \end{aligned} \tag{8}$$

where $N = t/\Delta\tau$. Once the correlation functions on the right-hand side are measured, Eq. (8) can be solved iteratively for $G(t)$ starting from $G(0) = 0$.

Variations of this procedure have been used frequently in the literature to calculate $G(t)$ for many single-dimensional CG models of molecular systems, and occasionally for multi-dimensional models. If a Markovian description is desired, then the limit of $G(t)$ as $t \to \infty$ furnishes the Markovian friction coefficients. If a non-Markovian description is necessary, then $G(t)$ can be differentiated to obtain the memory function, $\Gamma(t)$.

Unfortunately, since TCFs typically decay exponentially and are calculated from finite MD simulations, a low ratio of signal to noise becomes inevitable at long lag times, leading to poor stability of the long-time plateau in the integral. This limits the applicability of the Volterra inversion method to calculating Markovian friction coefficients.

To illustrate the memory extraction problem and the limitations of the procedure described by Eq. (8), we simulate Eq. (1) for a toy molecule. The molecule is formed by three CG sites, interacting with a known internal potential and memory function. The accuracy of the procedure is tested by attempting to recover the known friction coefficients via the integral of the memory function.

The internal potential of the trimer is given by

$$U(l_1, l_2, \theta) = \frac{k_1}{2}(l_1 - l_{1_0})^2 + \frac{k_2}{2}(l_2 - l_{2_0})^2 + \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{9}$$

where l_1, l_2 are the bond lengths along the three-bead chain and θ is the angle between them, and $(l_{1_0}, l_{2_0}, \theta_0)$ is the minimum-energy configuration, which we set equal to $(1, 1, \frac{\pi}{2})$. The spring constants are set to $(k_1, k_2, k_\theta) = (14, 20, 7)$.

The memory function of the trimer is taken to be

$$\Gamma_{ij}(t) = 10e^{-10t/\zeta_{ij}} \tag{10}$$

and the Markovian friction coefficients are

$$\zeta = \begin{bmatrix} 10 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 20 \end{bmatrix} \tag{11}$$

indicating a single off-diagonal interaction between the ends of the trimer. The beads in the trimer are assigned masses $m = (30, 40, 30)$.

The time-correlated noise is generated according to Eq. (2) using the method detailed in [28]. As the sum of many interactions, the noise is taken to be gaussian-distributed, motivated by the Central Limit Theorem. Equation (1) is then simulated using the Euler method with time step $\Delta t = 0.01$ for a total time of $T_{sim} = 10^4$.

In Fig. 2.1, we compare the results of extracting the memory function from the resulting trajectory using equation (8) with the true integral of the known Γ, Eq. 10. While the Volterra inversion method matches $G(t)$ well for short times, it begins to drift at timescales comparable to the Markov timescale we are interested in, and the error becomes progressively worse at longer timescales. Without a stable plateau, there is no unambiguous result for the friction coefficients.

To address these shortcomings of the Volterra inversion method, we propose an alternative approach that circumvents the direct calculation of $\Gamma(t)$ while reliably yielding the plateau value.

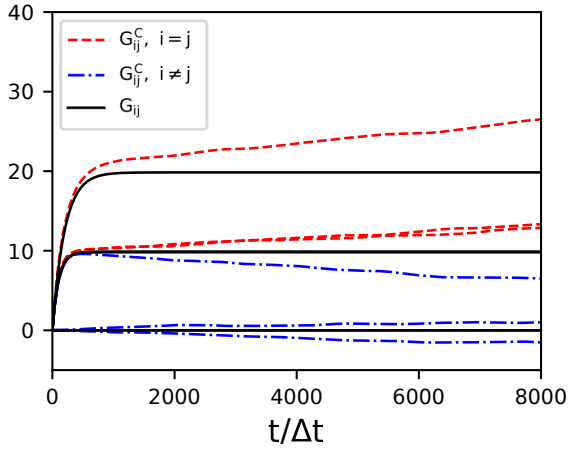

FIG. 2.1. The diagonal (dashed) and off-diagonal (dot-dashed) components of the calculated $G^C(t)$ as obtained via the Volterra inversion method compared to the exact solution $G(t)$ obtained from Eq. 10 (solid).

First we rewrite Eq. (6) as

$$MC^{vv}(t) - MC^{vv}(0) = \mathcal{F}^{vU'}(t) - \int_0^t dt' \int_0^{t'} d\tau\, \Gamma(t'-\tau) C^{vv}(\tau). \tag{12}$$

We then define

$$\mathcal{D}^{vv}(t) := \int_0^t d\tau\, C^{vv}(\tau)\ , \tag{13}$$

and multiply both sides of equation (12) by $\mathcal{D}^{vv^{-1}}$ to get

$$\zeta_v(t) = \left[MC^{vv}(0) - MC^{vv}(t) + \mathcal{F}^{vU'}(t)\right] \mathcal{D}^{vv^{-1}}(t) \tag{14}$$

where we have defined

$$\zeta_v(t) := \left[\int_0^t dt' \int_0^{t'} d\tau\, \Gamma(t'-\tau) C^{vv}(\tau)\right] \mathcal{D}^{vv^{-1}}(t)\ . \tag{15}$$

The subscript v recalls the initial value we multiplied across Eq. (1). Since the inner τ integral is a convolution between Γ and C^{vv}, in the limit $t \to \infty$ the double integral separates into two factors

$$\zeta_v(\infty) = \left[\int_0^\infty dt\, \Gamma(t) \int_0^\infty dt\, C^{vv}(t)\right] \mathcal{D}^{vv^{-1}}(\infty) \tag{16}$$

which by the definitions of G and $\mathcal{D}$ simplifies to

$$\zeta_v(\infty) = G(\infty)\ . \tag{17}$$

We have in $\zeta_v(t)$ a quantity distinct from the integral of the memory function that nonetheless has the same limit at long times. Using Eq. (14) the Markovian friction coefficients can be calculated directly from observed TCFs without solving an integral equation while maintaining consistency with the underlying memory function. To identify the Markovian timescale T_M, we calculate $\zeta_v(t)$ for multiple values of t in order to monitor convergence.

In the long time limit, one can understand Eq. (14) as a generalized Einstein relation. If we consider the case of a single molecular center-of-mass diffusing freely such that $U'(\vec{r}) = 0$, we have

$$\zeta_v(\infty) = \frac{mC^{vv}(0)}{\mathcal{D}^{vv}(\infty)} \tag{18}$$

given that $\mathcal{F}^{vU'}(t) = 0$ for all t, and $C^{vv}(t) \to 0$ as $t \to \infty$. By the equipartition theorem, one has $mC^{vv}(0) = 3k_BT$. The usual diffusion coefficient D is identified via $\mathcal{D}^{vv}(\infty) = 3D$, yielding the Einstein relation

$$\zeta_v(\infty) = \frac{k_B T}{D} \ . \tag{19}$$

The proposed formalism, illustrated in Eq. (14), is a convenient method to calculate the friction coefficients in systems described entirely by unbound coordinates, such as the center-of-mass diffusion of a free molecule. When attempting to model the internal dynamics of a molecule, however, a further obstacle arises that requires some generalization of the approach.

For a single bound coordinate, such as an internal coordinate of a coarse-grained macromolecule, the mean-squared displacement saturates to a finite value over time instead of growing linearly. In this case, the diffusion coefficient $\lim_{t\to\infty} \mathcal{D}^{vv}(t) = 0$. When modeling n atom clusters to describe an entire molecule, only the eigenvalue associated with translation of the entire CG molecule has a finite limit, while the remaining $n-1$ eigenvalues go to zero at long times.

The numerator of $\zeta_v(t)$ goes to a finite value at large t (see Appendix A), and $\zeta_v(\infty)$ diverges. The divergence of the Markovian limit in the case of bound coordinates was noted and resolved in Ref. [29] for a different approach to the memory function using a hierarchy of approximations. Here we implement a solution inspired by that approach that also simultaneously handles the unbound translational mode.

The formalism we derived in Eq. (14), can be easily generalized to solve this issue. While we initially chose to take the dot product of Eq. (1) with $\vec{v}_k(0)$, we now generalize this approach by taking the dot product with an arbitrary vector-valued function of the initial CG phase space variables $\vec{g}_k(\vec{r}(0), \vec{p}(0))$. To the extent that the state-independent GLE is valid, this vector $\vec{g}_k$ remains orthogonal to the noise since all CG variables are Zwanzig-projected coordinates (see Appendix A). The resulting generalized integral equation is:

$$M\dot{C}^{gv}(t) = -C^{gU'}(t) - \int_0^t d\tau\, \Gamma(t-\tau) C^{gv}(\tau). \tag{20}$$

Integrating in the same manner as before,

$$\zeta_g(t) = \left[MC^{gv}(0) - MC^{gv}(t) + \mathcal{F}^{gU'}(t) \right] \mathcal{D}^{gv^{-1}}(t), \tag{21}$$

yields a generalized friction coefficient

$$\zeta_g(t) := \left[\int_0^t dt' \int_0^{t'} d\tau\, \Gamma(t'-\tau) C^{gv}(\tau) \right] \mathcal{D}^{gv^{-1}}(t). \tag{22}$$

We note that

$$\zeta_g(\infty) = G(\infty) = \zeta \tag{23}$$

for all choices of $\vec{g}$. In the long time limit, Eq. (21) then provides an even more generalized

form of the Einstein relation, connecting the friction coefficient with ratios of a broader class of correlations resulting from different choices of $\vec{g}$.

We are now free to search for any set of vectors $\vec{g}_k(\vec{r},\vec{p})$ such that $\mathcal{D}^{gv}$ retains a full set of finite eigenvalues in the limit $T \to \infty$. For the case where all positions are bound, [29] were able to obtain a stable Markov limit using correlations equivalent to setting $\vec{g}_k = \vec{r}_k$ here. Inspired by their work we adopt

$$\vec{g}_0 := \begin{bmatrix} \vec{r'}_C/\tau_0 \\ \\ \vec{V} \end{bmatrix} \tag{24}$$

where $\vec{V}$ is the center-of-mass velocity and $\vec{r'}_C/\tau_0$ is the vector of CG position coordinates relative to the center-of-mass, normalized by the unit of time, τ_0. Here, the x, y, z coordinates of the Nth CG site are removed, such that $\vec{r'}_C = (x_1^C, y_1^C, z_1^C, ..., x_{n-1}^C, y_{n-1}^C, z_{n-1}^C)$. The motivation is to associate the bound modes of rotational and internal vibration with the bound position coordinates, $\vec{r}_C$, relative to the center-of-mass position, where one set of coordinates has been removed to maintain the correct vector's size. However, it is essential to include the unbound translational mode of the center of mass. If we simply set $\vec{g} = \vec{r}_C$, the bound position coordinates exclude the translational degrees of freedom, leading to an unwanted zero eigenvalue of $\mathcal{D}^{gv}(t)$.

Finally, we observe that while $\mathcal{D}^{gv}(t)$ has finite eigenvalues as $t \to \infty$ for $\vec{g} = \vec{g}_0$, its derivative $C^{gv}(t)$ now has zero eigenvalues at $t = 0$, as dictated by the equipartition theorem. Given that the first derivative of the numerator in equation (21) is finite, L'Hôpital's rule indicates that the limit as $t \to 0$ becomes infinite (see Appendix A). By adopting $\vec{g} = \vec{g}_0$ we effectively exchanged the long-time instability for a short-time one. Although this does not strictly pose a problem for the calculation of the Markovian friction, it would be preferable to achieve stability at both limits. This issue can be easily resolved through an additional adjustment of our $\vec{g}$-vector,

$$\vec{g} := \begin{bmatrix} \vec{r'}_C/\tau_0 - \vec{v'}_L \\ \\ \vec{V} \end{bmatrix} \tag{25}$$

where $\vec{v'}_L$ are the lab frame velocities where one set of velocity coordinates is removed. Since $C^{vv}(0)$ possesses a full set of finite eigenvalues due to the equipartition theorem, this stabilizes the $t \to 0$ limit (see Appendix A). Hence, Eq. (25) will be our choice for $\vec{g}$ from this point on.

We now calculate $\zeta_g(t)$ via Eqs. (21, 25) using the same trajectory of the trimer as before. In principle the Markovian friction can be obtained by calculating $\zeta_g(t)$ at a single

time $t > T_M$. In practice, however, T_M is generally unknown. In this case $\zeta_g(t)$ can be calculated at multiple sparse values of t and monitored for convergence. In Fig. 2.2 we plot $\zeta_g(t)$ at full resolution to facilitate comparison with $G(t)$.

The results demonstrate a clearly identifiable and stable plateau that agrees well with the long-time limit of the exact $G(t)$.

We have presented an Einstein relation generalized to systems of interacting particles, where the friction coefficients are consistent with the total integral of an underlying memory function. Due to the nature of the Einstein relation as a ratio, it is possible to further generalize to a broader class of correlations than the usual velocity-velocity correlations corresponding to the diffusion coefficient, and all of these ratios yield the same friction coefficients. The generality of this formalism allows us the freedom of selecting appropriate coordinates to resolve instabilities in the memory integral.

This relation could serve as a powerful tool for constructing Markovian CG models. Not only does it yield accurate friction coefficients without explicit calculation of the memory function, but its flexibility allows it to be tailored to the coordinate structure of the system at hand in order to achieve robust numerical stability. The example we have provided demonstrates how site-specific friction coefficients simultaneously consistent with bound internal dynamics and unbound global diffusion in a free molecule can be obtained.

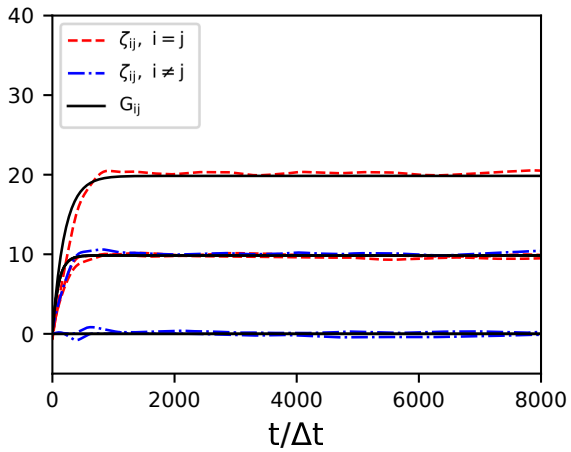

FIG. 2.2. The diagonal (dashed) and off-diagonal (dot-dashed) components of $\zeta_g(T)$ compared with the exact components of $G(T)$ (solid).

BRIDGE

In the preceding paper, the Markov friction coefficients were calculated from an observed molecular trajectory using lagged time correlations. One of the correlations required is that of the mean force against coarse-grained phase space variables. The friction coefficients, then, cannot be calculated in this way until the mean force has itself been parametrized accurately.

In the preceding paper the toy model obeyed a given Generalized Langevin Equation exactly, and hence the true forces from the original simulation could be used to isolate the problem of modeling the friction. In a real system, however, the mean force is a result of the projection, and must be modeled.

There is a rich literature on extracting the mean force and the potential of mean force from observed equilibrium distributions in a wide variety of systems. However in our work we found that the traditional inversion of the Boltzmann factor yield poor results. In this case the chief obstacle was, yet again, the rotational symmetry inherent to a diffusing molecule. Inversion of the Boltzmann factor yields a free energy that is not identical to the potential of mean force itself if there is any shape-dependent entropy present. The entropy corresponding to rotational degrees of freedom is in fact shape-dependent, and so a correction is needed to isolate the potential of mean force and hence obtain an accurate force field.

In the following paper we derive such a correction, and we demonstrate that applying it significantly improves the accuracy of the modeled force, and that the correction can be cruical for the recovery of correct friction coefficients by the method of the preceding paper.

---

[1] K. A. Dill and J. L. MacCallum, Science **338**, 1042 (2012).

[2] J. Copperman, M. Dinpajooh, E. R. Beyerle, and M. G. Guenza, Phys. Rev. Lett. **119**, 158101 (2017).

[3] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).

[4] O. F. Lange and H. Grubmuller **70**, 1294 (2008).

[5] H. Mori, Prog. Theor. Phys., Japan **33**, 423 (1965).

[6] R. Zwanzig, (Oxford : New York : Oxford University Press, 2001).

[7] B. Kowalik, J. O. Daldrop, J. Kappler, J. C. F. Schulz, A. Schlaich, and R. R. Netz, Phys. Rev. E **100**, 012126 (2019).

[8] B. A. Dalton, C. Ayaz, H. Kiefer, and R. R. Netz, Proc. Natl. Acad. Sci. **120**, 31 (2023).

[9] H. S. Lee, S. Ahn, and E. F. Darve, J. Chem. Phys. **150**, 174113 (2019).

[10] J. E. Straub, M. Borkovec, and B. J. Berne, J. Phys. Chem. **91**, 4995 (1987).

[11] B. J. Berne and G. D. Harp, (John Wiley & Sons, Ltd, 1970) Chap. 3, p. 63.

[12] J. O. Daldrop, J. Kappler, F. N. Brunig, and R. R. Netz, Proc. Natl. Acad. Sci. U. S. A. **115**, 5169 (2018).

[13] J. O. Daldrop, B. G. Kowalik, and R. R. Netz, Phys. Rev. X **7**, 041065 (2017).

[14] M. Berkowitz, J. D. Morgan, D. J. Kouri, and J. A. McCammon, J. Chem. Phys. **75**, 2462 (1981).

[15] D. A. N. Foster, R. Petrosyan, A. G. T. Pyo, A. Hoffmann, F. Wang, and M. T. Woodside, Biophys. J. **114**, 1657 (2018).

[16] R. J. de Oliveira, J. Chem. Phys. **149**, 234107 (2018).

[17] F. C. Freitas, A. N. Lima, V. G. Contessoto, P. C. Whitford, and R. J. Oliveira, J. Chem. Phys. **151**, 114106 (2019).

[18] H. Lannon, J. S. Haghpanah, J. K. Montclare, E. Vanden- Eijnden, and J. Brujic, Phys. Rev. Lett. **110**, 128301 (2013).

[19] R. Satija, and D. E. Makarov, J. Phys. Chem. B **123**, 802 (2019).

[20] G. Hummer, New J. Phys. **7**, 34 (2005).

[21] R. B. Best, and G. Hummer, Proc. Natl. Acad. Sci. U. S. A. **107**, 1088 (2010).

[22] R. B. Best, and G. Hummer, Phys. Chem. Chem. Phys. **13**, 16902 (2011).

[23] M. L. Mugnai, and R. Elber, J. Chem. Phys. **142**, 014105 (2015).

[24] M. Hinczewski, Y. von Hansen, J. Dzubiella, and R. R. Netz, J. Chem. Phys. **132**, 245103 (2010).

[25] C. Hijon, P. Espanol, E. Vanden-Eijnden, and R. Delgado-Buscalioni, Faraday Discuss. **144**, 301 (2010).

[26] N. G. Van Kampen, (Elsevier, 2007).

[27] P. Linz, The Computer Journal **12**, 393 (1969).

[28] J. Garcıa-Ojalvo and J. M. Sancho, (Springer-Verlag New York, Inc., 1999) pp. 95–97.

[29] F. Grogan, H. Lei, X. Li, and N. A. Baker, J. Chem.Phys. **418**, 109633 (2020).

[30] T. J. Boerner, S. Deems, T. R. Furlani, S. L. Knuth, and J. Towns, (Association for Computing Machinery, New York, NY, USA, 2023) p. 173–176.

3 - ROTATIONAL ENTROPY

The following is reproduced from a manuscript currently under review for publication, authored by J.M. Hall and M.G. Guenza.

Coarse-grained (CG) models of solvated molecules often rely on the potential of mean force (PMF) to capture the equilibrium conformational distributions. When the gradient of the PMF is used, it also defines the corresponding force field governing the system's dynamics. The PMF is often obtained via the Boltzmann factor from the distribution of molecular configurations observed in computer simulations [1, 2]. While progress has been made on techniques for optimizing the potential for use in Langevin Equations to simulate reaction coordinates, for example through machine learning, much of this work has been focused on low-dimensional models of purely internal degrees of freedom describing the shape of the molecule. Such models do not include the translational and rotational diffusion of the molecule or any possible rovibrational coupling [3–7]. In many biological applications, such as protein–ligand binding, the timescales of slow conformational fluctuations and overall orientational dynamics can be comparable, making it desirable to explicitly model all relevant degrees of freedom for a comprehensive understanding of the process. In this work, we present a formalism to calculate an accurate lab-frame force field with the appropriate rot+trans symmetries by modifying the PMF obtained via inversion of the Boltzmann factor from the distribution of the shape coordinates.

A coarse-grained model of a molecule composed of N point particles has $3N$ total degrees of freedom, $\vec{r}$. However, because the potential energy surface is invariant under global translations and rotations, it depends only on a reduced set of $3N-6$ internal shape coordinates, $\vec{q}$. These coordinates can be defined by removing the effect of translation and rotation through appropriate frame constraints [8]. Despite this reduction, the six global degrees of freedom remain implicitly present in the coarse-grained space and can couple to the internal dynamics, introducing a residual rotational entropy in the free energy landscape over $\vec{q}$. This entropy contribution is not associated with the internal potential $V(\vec{q})$. Therefore, to obtain a force field corresponding solely to $V(\vec{q})$ by inverting the Boltzmann distribution, the shape-dependent rotational entropy must be properly accounted for and removed to isolate the contribution of the potential energy surface.

If we were interested only in a body-fixed coarse-grained description, for example, of a single internal reaction coordinate, then the translational and rotational degrees of freedom would be eliminated through the projection procedure itself, with their influence absorbed

into the memory kernel and noise. It is our decision to keep these symmetries explicit in our CG model that forces us to grapple with their effects on the free energy landscape. This choice, while more complex, enables us to capture global diffusion and its coupling to internal dynamics, providing a foundation for coarse-grained models of processes such as molecular binding.

I. BODY-FIXED COORDINATES

A freely diffusing molecule may be modeled as N point particles, which can represent atoms or the centers of coarse-grained clusters, for a total of $3N$ spatial coordinates.

The general Hamiltonian for the N particles in Cartesian coordinates is given by

$$H = \frac{1}{2}\vec{p}M^{-1}\vec{p}^{T} + V(\vec{r}). \tag{1}$$

Here M is a mass matrix consisting of a 3×3 diagonal block $m_i I_3$ for each particle i. The vector $\vec{p}$ consists of the xyz-momenta of each of the N particles such that $\vec{p} = (p_{1x}, p_{1y}, p_{1z}, ..., p_{Nx}, p_{Ny}, p_{Nz})$, and likewise $\vec{r}$ represents the lab-frame positions. The potential energy $V(\vec{r})$ will have translational and rotational symmetry and so can be completely described by a reduced set of $3N - 6$ shape coordinates.

As in [8], the six global coordinates corresponding to the overall translation and rotation can be defined via constraint equations on the $3N$ configuration coordinates. The choice of constraints defines the body-fixed frame and fixes the overall position $\vec{R}$ and Euler angles $\vec{\Theta}$ given a configuration $\vec{r}$. An example set of constraints are the Eckart conditions [9], which are linear equations and identify $\vec{R}$ with the center-of-mass of the N particles while defining the $\vec{\Theta}$ such that rotational-vibrational coupling in vacuum is minimized near a reference configuration. Once the constraints have been chosen, a reduced set of $3N-6$ internal shape coordinates $\vec{q}$ may also be chosen to complete the body-fixed description. Again following the approach in [8], the Hamiltonian may be recast, for any choice of constraints, in canonical body-fixed coordinates as

$$H = \frac{1}{2}\vec{P}M^{-1}\vec{P}^{T} + \frac{1}{2}\vec{p}S^{-1}\vec{p}^{T} + \frac{1}{2}\vec{\mathcal{J}}I^{*^{-1}}\vec{\mathcal{J}}^{T} + V(\vec{q}). \tag{2}$$

The first term is the kinetic energy associated with the center-of-mass, where $\vec{P}$ is the total momentum of the molecule and M the total mass. The second term is the kinetic energy associated with the vibrational motion of the shape coordinates $\vec{q}$, which we have the freedom to choose, and S is a mass-weighted metric defined as

$$S_{ij} := \sum_{k,\alpha} m_k \frac{\partial r_{k\alpha}}{\partial q_i} \frac{\partial r_{k\alpha}}{\partial q_j}. \tag{3}$$

(Throughout this manuscript, Greek indices denote Cartesian components.) The third term represents the kinetic energy associated with a generalized angular momentum $\vec{\mathcal{J}}$, corresponding to the total angular momentum less the contribution from vibrations in the body-fixed frame, where

$$\vec{\mathcal{J}} := I^* \vec{\omega} \ , \tag{4}$$

and

$$I^* := I - CS^{-1}C^T \tag{5}$$

with

$$C_{\alpha i} := \sum_k m_k (\vec{r}_k \times \frac{\partial \vec{r}_k}{\partial q_i})_\alpha. \tag{6}$$

The generalized angular momentum may also be expressed entirely in canonical coordinates via

$$\vec{\mathcal{J}} = B(\vec{\Theta})\vec{p}_\Theta - CS^{-1}\vec{p}, \tag{7}$$

where $\vec{p}$ and $\vec{p}_\Theta$ are the canonical momenta associated with $\vec{q}$ and $\vec{\Theta}$ respectively, and

$$B(\vec{\Theta}) := \begin{pmatrix} sin(\chi) & -csc(\theta)cos(\chi) & cot(\theta)cos(\chi) \\ cos(\chi) & csc(\theta)sin(\chi) & -cot(\theta)sin(\chi) \\ 0 & 0 & 1 \end{pmatrix} \tag{8}$$

as in the treatment of rovibrational coupling in [10].

We are free to choose a set of internal coordinates to describe the shape of the molecule. Certain coordinate systems will have inherent advantages and disadvantages. The Bond-Angle-Torsion (BAT) coordinates are popular and have the advantage of being rotation-invariant, and hence calculable in any frame[11–13]. They are fundamentally curvilinear coordinates, however, and we will see that their state-dependent metric, via the S matrix (3), introduces other difficulties. We will instead, in this work, accept the inconvenience of frame-switching and work with a reduced set of rectilinear displacements from a body-fixed reference structure (see Appendix B).

To obtain the distribution of internal displacements, the translational and rotational motion of the molecule must be removed from the lab-frame trajectory. This procedure

is inherently ambiguous, so appropriate frame constraints must be applied. To eliminate translational motion, we fix the center of mass at the origin. To remove rotational motion, we adopt the Eckart frame constraints. [9, 14]. This choice is again motivated by the desire to maintain linearity in the constraint equations, which ensures a constant metric tensor. In contrast, using the quadratic constraints of the principal axes frame would result in a curvilinear metric for the internal coordinates.

II. ROTATIONAL ENTROPY CORRECTION TO THE FORCE FIELD IN A COARSE-GRAINED REPRESENTATION

If the atomistic system is Boltzmann-distributed, then its probability distribution in phase-space is

$$\mathcal{P}^{atom}(\vec{r},\vec{p})d\vec{r}\,d\vec{p} = \frac{1}{Z}e^{-\left(\frac{1}{2}\sum_i \frac{1}{m_i}p_i^2 + V^{atom}(\vec{r})\right)/k_BT} d\vec{r}\,d\vec{p} \tag{9}$$

where $\vec{r},\vec{p}$ are the lab-frame positions and momenta of the atoms. The Boltzmann distribution doesn't strictly apply in the case of unbounded translational diffusion, but the simulated system can be made normalizable by a finite box or periodic boundary conditions and the influence of these constraints made sufficiently small as desired.

We will consider the case where the desired CG variables are the lab-frame centers-of-mass $\vec{r}^L$ of designated clusters of atoms in the molecule, and their corresponding momenta $\vec{p}^L$. We note that the remaining degrees of freedom in each cluster can be represented by the relative Jacobi coordinates $\vec{r}^J,\vec{p}^J$ while preserving the decoupled quadratic form of the kinetic energy. This allows us to trivially integrate out these degrees of freedom, together with those of the solvent and other ignored atoms, yielding the distribution in the CG phase space:

$$\mathcal{P}^L(\vec{r}^L,\vec{p}^L)d\vec{r}^L\,d\vec{p}^L = \frac{1}{Z}e^{-\left(\frac{1}{2}\sum_n \frac{1}{m_n}p_n^{L^2} + V(\vec{r}^L)\right)/k_BT} d\vec{r}^L\,d\vec{p}^L\ , \tag{10}$$

where

$$V(\vec{r}^L) := -k_BT\ ln\left(\int d\vec{r}_{orth}\ e^{-V^{atom}(\vec{r})/k_BT}\right). \tag{11}$$

Owing to translational and rotational invariance, the potential of mean force V for a freely diffusing molecule depends only on the $3N-6$ shape coordinates $\vec{q}$, i.e., $V = V(\vec{q})$.

Our objective is to derive an expression for the observable probability distribution $\mathcal{P}(\vec{q})$ in terms of the potential of mean force $V(\vec{q})$ by integrating out the center-of-mass and rotational (Euler) degrees of freedom from 31.

To this end, we begin by performing a canonical transformation from the lab-frame coordinates to a body-fixed representation, following the approach derived in [8] and summarized

in Section I.

$$(\vec{r^L}, \vec{p^L}) \to (\vec{q}, \vec{p}, \vec{R}, \vec{P}, \vec{\Theta}, \vec{p}_\Theta) \tag{12}$$

where $\vec{q}, \vec{p}$ are the internal shape coordinates and their conjugate momenta, $\vec{R}, \vec{P}$ are the center-of-mass of the CG sites and its momentum, $\vec{\Theta} = (\theta, \phi, \chi)$ are the Euler angles, and $\vec{p}_\Theta = (p_\theta, p_\phi, p_\chi)$ are the conjugate momenta to the Euler angles.

Since the transformation is canonical, Liouville's theorem implies that the phase space volume is preserved, and thus the probability distribution retains a simple volume element:

$$d\vec{r}^L \, d\vec{p}^L \to d\vec{R} \, d\vec{P} \, d\vec{\Theta} \, d\vec{p}_\Theta \, d\vec{q} \, d\vec{p}. \tag{13}$$

Again following [8], the Hamiltonian is now expressed in the body-fixed coordinates by Eq. (2).

The Hamiltonian's independence from $\vec{R}$ implies a uniform distribution over center-of-mass coordinates, which may thus be trivially integrated out. Moreover, the center-of-mass momentum decouples from the remaining degrees of freedom and can likewise be integrated out, yielding:

$$\mathcal{P}^0(\vec{\Theta}, \vec{p}_\Theta, \vec{q}, \vec{p}) = \frac{1}{Z} e^{-\left(\frac{1}{2}\vec{p}S^{-1}\vec{p}^T + \frac{1}{2}\vec{\mathcal{J}}{I^*}^{-1}\vec{\mathcal{J}}^T + V(\vec{q})\right)/k_BT} \, . \tag{14}$$

A Gaussian integration over the internal shape momenta yields:

$$\int d\vec{p} \, e^{-\left(\frac{1}{2}\vec{p}S^{-1}\vec{p}^T\right)/k_BT} = k_BT\sqrt{(2\pi)^{3N-6}}\sqrt{(|S(\vec{q})|)} \, . \tag{15}$$

After incorporating the constants into the normalization, the expression becomes:

$$\mathcal{P}^1(\vec{\Theta}, \vec{p}_\Theta, \vec{q}) = \frac{1}{Z}\sqrt{|S(\vec{q})|} e^{-\left(\frac{1}{2}\vec{\mathcal{J}}{I^*}^{-1}\vec{\mathcal{J}}^T + V(\vec{q})\right)/k_BT} \, . \tag{16}$$

Due to the simultaneous dependence of $\vec{\mathcal{J}}(\vec{\Theta}, \vec{p}_\Theta)$ on $\vec{\Theta}$ and $\vec{p}_\Theta$, there is no simple way to integrate out these coordinates. Instead, we perform a coordinate transformation from the canonical momenta $\vec{p}_\Theta$ to the angular velocity $\vec{\omega}$. Rewriting the first term of the exponent in terms of $\vec{\omega}$ gives us

$$\frac{1}{2}\vec{\mathcal{J}}{I^*}^{-1}\vec{\mathcal{J}}^T = \frac{1}{2}\vec{\omega}I^*\vec{\omega}^T \, . \tag{17}$$

Since the transformation is noncanonical, care must be taken with the volume element, which transforms according to the determinant of the Jacobian as

$$d\vec{p}_\Theta = |J| d\vec{\omega} \ , \tag{18}$$

where

$$J = \frac{\partial \vec{p}_\Theta}{\partial \vec{\omega}}. \tag{19}$$

We find this derivative by inverting Eq. (7) to express $\vec{p}_\Theta$ in terms of the remaining coordinates and $\vec{\omega}$:

$$\vec{p}_{\vec{\Theta}} = B^{-1}(\vec{\Theta}) \left(I^*(\vec{q})\vec{\omega} + C(\vec{q})S^{-1}(\vec{q})\vec{p} \right). \tag{20}$$

It follows that

$$\frac{\partial \vec{p}_\Theta}{\partial \vec{\omega}} = B^{-1}(\vec{\Theta}) I^*(\vec{q}) \ , \tag{21}$$

and the Jacobian determinant is

$$\begin{aligned} |J| &= \left| \frac{\partial \vec{p}_\Theta}{\partial \vec{\omega}} \right| \\ &= \left| B^{-1}(\vec{\Theta}) I^*(\vec{q}) \right| \\ &= \left| B^{-1}(\vec{\Theta}) \right| \left| I^*(\vec{q}) \right|. \end{aligned} \tag{22}$$

Upon evaluating the determinant of $B^{-1}(\vec{\Theta})$, we find

$$|J| = sin(\theta) \left| I^*(\vec{q}) \right|. \tag{23}$$

Thus, after the coordinate transformation, the probability distribution becomes

$$\mathcal{P}(\vec{\Theta}, \vec{\omega}, \vec{q}) = \frac{1}{Z} sin(\theta) |I^*(\vec{q})| \sqrt{|S(\vec{q})|} \, e^{-\left(\frac{1}{2}\vec{\omega} I^* \vec{\omega}^T + V(\vec{q})\right)/k_B T}. \tag{24}$$

The dependence on ϕ and χ has dropped out, and the $sin(\theta)$ term integrates to a constant, leaving

$$\mathcal{P}(\vec{\omega}, \vec{q}) = \frac{1}{Z} |I^*(\vec{q})| \sqrt{|S(\vec{q})|} \, e^{-\left(\frac{1}{2}\vec{\omega} I^* \vec{\omega}^T + V(\vec{q})\right)/k_B T}. \tag{25}$$

Finally, performing the Gaussian integral over $\vec{\omega}$ yields the final expression for the probability distribution:

$$\mathcal{P}(\vec{q}) = \frac{1}{Z}\sqrt{|I^*(\vec{q})|}\sqrt{|S(\vec{q})|}\, e^{-V(\vec{q})/k_BT}. \tag{26}$$

We find that the resulting distribution of shape coordinates, expressed via their potential of mean force, deviates from what a naive application of the Boltzmann factor would predict. The landscape is modified by the $\vec{q}$-dependence of both I^* and S. Notably, this result holds for arbitrary choices of shape coordinates and frame constraints.

The shape-dependence of S is merely a geometric artifact introduced by the use of curvilinear shape coordinates and/or nonlinear frame constraints. As discussed in Appendix B, this dependence can be removed by an appropriate choice of coordinates and constraints. For the remainder of this work, we assume such a choice has been made, rendering S a constant that may be absorbed into the normalization.

By contrast, the shape-dependence of I^* is an inherent consequence of the rotational entropy associated with a given conformation and must be evaluated.

The connection to rotational entropy becomes more apparent upon rewriting the probability distribution as

$$\mathcal{P}(\vec{q}) = \frac{1}{Z}\, e^{-\left(V(\vec{q}) - TS_R(\vec{q})\right)/k_BT}, \tag{27}$$

where the rotational entropy S_R associated with the coarse-grained shape $\vec{q}$ is defined by

$$S_R(\vec{q}) := k_B\, ln(A\sqrt{|I^*(\vec{q})|}), \tag{28}$$

and

$$A := \frac{\pi^{1/2}}{\sigma}\left(\frac{8\pi^2 e k_B T}{h^2}\right)^{3/2}, \tag{29}$$

is a normalization constant. The symmetry number σ accounts for the symmetry of the reference conformation, taking values greater than one for molecules with rotational self-symmetry, and unity otherwise [15].

For a rigid body, I^* reduces to a constant matrix equal to the standard moment of inertia, yielding the familiar rigid-body expression for the rotational entropy of the reference conformation $\vec{q} = 0$.[15] For flexible molecules, the conformation-dependent moment of inertia must be replaced by the generalized tensor of Eq.5. Accordingly, the potential of mean force

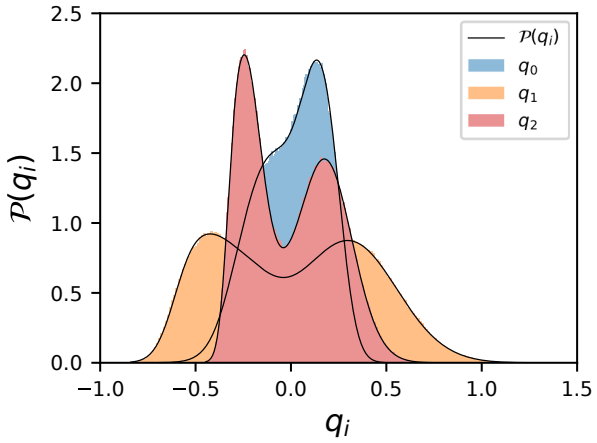

FIG. 3.1: Histograms (transparent) of the shape coordinate distributions obtained from simulations of the model trimer, overlaid with the corresponding marginal distributions predicted by a 10-component Gaussian Mixture Model (GMM) fit (solid lines).

$V(\vec{q})$ relates to the probability distribution $\mathcal{P}(\vec{q})$ as follows:

$$V(\vec{q}) = -k_B T \, ln(\mathcal{P}(\vec{q})) + k_B T \, ln(A\sqrt{|I^*(\vec{q})|}). \tag{30}$$

The generalized force conjugate to the ith shape coordinate is therefore

$$-\frac{\partial V(\vec{q})}{\partial q_i} = k_B T \, \frac{1}{\mathcal{P}(\vec{q})} \frac{\partial \mathcal{P}(\vec{q})}{\partial q_i} - \frac{k_B T}{2} \, \frac{1}{\left|I^*(\vec{q})\right|} \frac{\partial}{\partial q_i} \left|I^*(\vec{q})\right|. \tag{31}$$

The first term is obtained from an analytical fit to the probability distribution, as described below. During numerical calculation, it may be useful to replace the derivative of the determinant in the second term with the derivative of a matrix using Jacobi's formula:

$$-\frac{\partial V(\vec{q})}{\partial q_i} = k_B T \, \frac{1}{\mathcal{P}(\vec{q})} \frac{\partial \mathcal{P}(\vec{q})}{\partial q_i} - \frac{k_B T}{2} Tr\big(I^{*^{-1}}(\vec{q}) \frac{\partial I^*(\vec{q})}{\partial q_i}\big). \tag{32}$$

III. PARAMETERIZATION OF THE FORCE FIELD FOR A MODEL TRIMER

Representative samples from the distribution $\mathcal{P}(\vec{q})$ can be obtained from long, ergodic simulations of a molecule, for example through Molecular Dynamics (MD) simulations of a protein in solvent. To study the coarse-grained dynamics of that molecule, one needs to define the CG sites by grouping together a number of atoms. Then, by integrating out the

fast degrees of freedom using a projection operator expressed in the space of the coarse-grained coordinates, one derives a Generalized Langevin Equation (GLE) that guides the CG dynamics of the molecule. [16, 17]

Over sufficiently long time intervals, where the protein dynamics is uncorrelated with the solvent dynamics, the GLE reduces to a Markovian Langevin equation in the position, $\vec{r}_i(t)$, and velocity, $\vec{v}_i(t) = \vec{p}_i(t)/m_i$, of each CG site, i. The resulting Langevin equation,

$$\begin{aligned}\dot{\vec{r}}_i(t) =& \frac{1}{m_i}\vec{p}_i(t),\\ \dot{\vec{p}}_i(t) =& -\frac{\partial V(\vec{r}(t))}{\partial \vec{r}_i} - \sum_j \zeta_{ij}\vec{v}_j(t) + \vec{f}_i^R(t),\end{aligned} \tag{33}$$

describes the time evolution of the CG variables in the field of the fast variables projected out. The eliminated degrees of freedom affect the dynamics through the form of the effective potential $V(\vec{r})$, the random forces $\vec{f}_i^R$, and the friction coefficients ζ_{ij}.

The friction coefficients are obtained from the time integral of the memory kernel. In recent work, we demonstrated that these coefficients can be computed as a ratio of time correlation functions, highlighting that an accurate representation of the friction requires an equally accurate force-field parameterization [18]. Consequently, to rigorously evaluate the accuracy and dynamical impact of the rotational entropy correction in coarse-grained (CG) models, it is essential to employ a system where all quantities entering the Langevin equation are known exactly.

As a test system, we consider a model trimer composed of three beads with masses $m = (3, 4, 3)$ and Markovian friction coefficients $\zeta = (10, 10, 20)$. All the simulation quantities are in reduced units [19, 20]. The internal potential governing the trimer conformation is given by

$$\begin{aligned}U(l_1, l_2, \theta) =& \frac{k_1}{2}(l_1 - l_{1_0})^2 + \frac{k_2}{2}(l_2 - l_{2_0})^2\\ &+ \frac{k_\theta}{2}[(\theta - \theta_0)^2(\theta - (\pi - \theta_0))^2 - b(\theta - \frac{\pi}{2})^2],\end{aligned} \tag{34}$$

where l_1, l_2 are the bond lengths and θ is the angle between them. The trimer thus has a double well in the bond angle symmetric around $\theta = \pi/2$. The well locations are defined by $(l_{1_0}, l_{2_0}, \theta_0) = (1, 1, \pi/3)$. The other well parameters are chosen as $(k_1, k_2, k_\theta, b) = (40, 40, 28, 1.5)$.

The stochastic noise is modeled as Gaussian-distributed, consistent with the Central Limit Theorem, which justifies this approximation in real systems where the random force arises from the cumulative effect of numerous solvent collisions. The temperature is set to $k_BT = 5$

via by the second fluctuation-dissipation theorem [21]:

$$\langle f^R_{i\alpha}(t) f^R_{j\beta}(t') \rangle = 2k_B T \zeta_{ij} \delta(t - t') \delta_{\alpha\beta}. \tag{35}$$

Using this model system, we investigate the influence and quantitative impact of the rotational entropy correction on both the reconstruction of the full force field and the calculation of friction coefficients. We simulate Equation (33) using the Euler-Maruyama method with time step $\Delta t = 0.01$ for a total time of $T_{sim} = 10^4$.

To calculate the force field, we need to transform the trajectory into body-centered dynamics. Thus, after simulating Eq. (33) in the laboratory frame, we derive the shape coordinates in the corresponding Eckart frame. We construct a reference structure $\vec{r}_0$ by aligning the configuration of one of the wells, $(l_1 = 1, l_2 = 1, \theta = \pi/3)$, into a principal axes frame. Using this reference, we then compute the time series of Eckart-frame Euler angles from the laboratory-frame trajectory via the quaternion-based method described in [22]. The shape coordinates $\vec{q}$ are defined as the first $3N - 6$ elements of $P(\vec{r}_B - \vec{r}_0)$, where the $\vec{r}_B$ are the body-fixed positions of the beads and P is the permutation matrix

$$P_{ij} = \begin{cases} \delta_{ij}, & i \notin (3N-6,\ 3N-5) \\ \delta_{i+1,j} & i = 3N-6 \\ \delta_{i-1,j} & i = 3N-5 \end{cases} \tag{36}$$

With this permutation matrix the last bead N is omitted from the shape coordinates entirely, bead $(N-1)$ retains only an x-coordinate in the body-fixed frame, and bead $(N-2)$ retains its x- and y-coordinates. This staggered set of internal displacements prevents degeneracies in the calculation of $\left(\frac{d\vec{r}}{d\vec{q}}\right)$. More details are reported in Appendix B.

In order to generate the force field described above, an analytical estimate of $\mathcal{P}(\vec{q})$ must be fit to the samples from the simulation. We model $\mathcal{P}(\vec{q})$ from this data with a Gaussian Mixture Model (GMM) fit using the Expectation-Maximization (EM) algorithm implemented in the Python package Scikit-Learn [**?**].

Figure 3.1 shows the observed histograms for the trimer's three shape coordinates, overlaid with the corresponding marginals of the GMM fit $\mathcal{P}(\vec{q})$ constructed using 10 components. The close agreement between the GMM and the empirical distributions for each coordinate suggests that 10 Gaussians provide a reasonably accurate representation. This will be further validated in IV through direct comparison of the resulting force field.

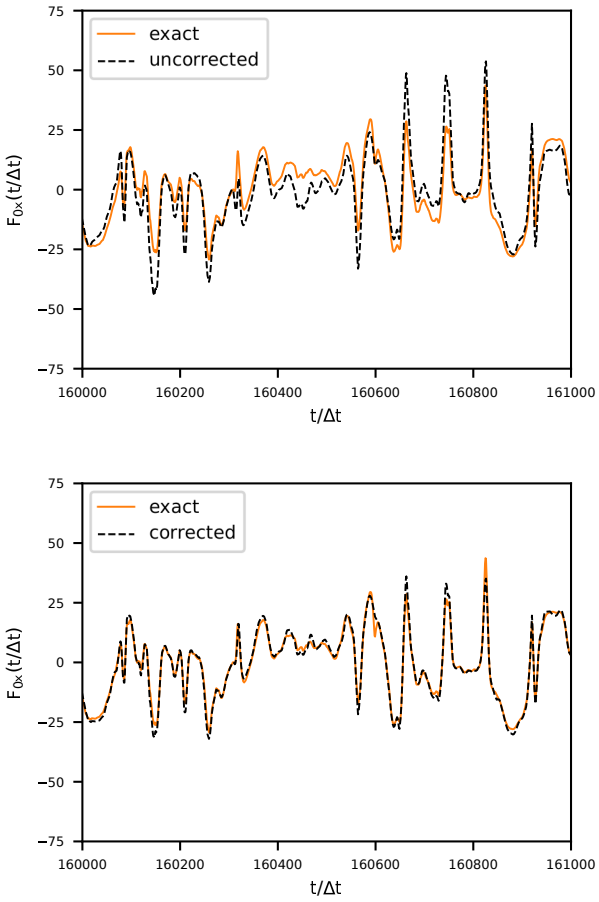

FIG. 3.2: Representative segment of the x-component force time series acting on a single bead of the trimer. The exact force from simulation (solid line) is compared with the reconstructed force obtained from the gradient of the Gaussian Mixture Model (GMM) fit (dashed line), shown both without (top panel) and with (bottom panel) the rotational entropy correction.

IV. RESULTS

Given a fitted form of $\mathcal{P}(\vec{q})$, obtained, for instance, using a Gaussian Mixture Model, the gradient of the distribution can be evaluated and substituted into Eq. 32 to compute a time series of generalized forces along the sampled shape trajectories. These generalized forces are then mapped back to the full set of lab frame forces by applying the inverse transformation procedure outlined in Appendix B.2. To isolate the effect of the rotational entropy correction, we compare the force field calculated with and without the second term in Eq. 32.

The top panel of Fig. 3.2 shows the time series of forces reconstructed from the GMM fit, excluding the rotational entropy correction, plotted against the exact forces from the

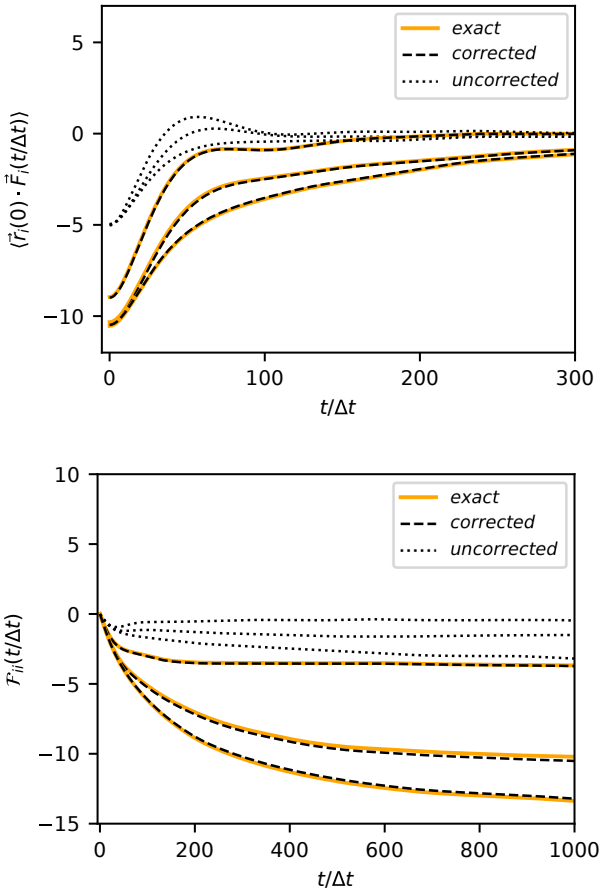

FIG. 3.3: The top panel displays the lagged position–force correlation function for each bead of the model trimer, computed using the exact simulation forces (solid line), the uncorrected gradient of the Gaussian Mixture Model (GMM) fit (dotted line), and the corrected gradient incorporating the rotational entropy term as in Eq. 32 (dashed line). The corresponding running integrals $\mathcal{F}_{ii}(t/\Delta t)$ of these correlation functions are shown in the bottom panel using the same line styles.

simulation over a representative segment of the trajectory. Although the overall trends are qualitatively captured, significant discrepancies remain in the finer details of the force profile.

In the bottom panel of Fig. 3.2, we show the results of the full force reconstruction, now including the rotational entropy correction. The agreement with the exact forces improves markedly, capturing both the overall trends and finer features of the time series.

To evaluate the practical significance of the rotational entropy correction, we turn now to calculate its influence on the reconstruction of time correlation functions (TCFs).

The first TCF we consider is the lagged position–force correlation, which directly contributes to the calculation of the Markovian friction coefficients [18]. As such, the accuracy of the reconstructed friction depends critically on the fidelity of the reconstructed force field.

In Fig. 3.3, we compare observed position–force TCFs for the trimer with the corresponding TCFs computed using the modeled force field, both with and without the rotational entropy correction. Notably, accurate recovery of the TCF is achieved only when the rotational correction is included.

The deviation observed at small $t/\Delta t$ in Fig. 3.3 underscores the importance of accounting for rotational distortion in the free energy landscape when aiming to accurately capture short-time dynamical behavior. Notably, certain methods for parametrizing friction coefficients or memory kernels in coarse-grained Langevin models rely on the time integral of this correlation function as an input [18, 23]. In the bottom panel of Fig. 3 we see that the error in the total integral of the position-force correlation incurred by ignoring the rotational correction is considerable. As a result, inaccuracies at short times can propagate into the estimation of parameters that govern long-time dynamics. These findings suggest that multiple aspects of CG model construction, based on observed statistics, may be sensitive to the effects of rotational entropy.

To further investigate this possibility, we compute the friction coefficients of the trimer using the Generalized Einstein Relation described in Reference [18]. The calculation is performed using three different force fields: the reference force field obtained directly from the simulation trajectory, and two reconstructed force fields derived from the GMM fit to the probability distribution—with and without the inclusion of the rotational entropy correction (see Eq. 32).

As shown in Fig. 3.4, the force field reconstructed from the GMM with the rotational entropy correction reproduces the true friction coefficients $\zeta = (10, 10, 20)$. Both exhibit rapid convergence to well-defined plateaus at short times, consistent with the Markovian character of the simulation. In contrast, the friction functions $\zeta_g(t)$ obtained from the uncorrected GMM display a slower convergence to their asymptotic limits, indicating artificial memory effects, and ultimately plateau at incorrect values.

V. DISCUSSION

Recovering the effective intramolecular force field from atomistic simulations is an essential step in modeling complex macromolecular systems by coarse-graining. In this article, we show that correcting for rotational entropy is, in general, essential for accurately recovering the underlying intramolecular force field from the distribution of internal coordinates in molecules undergoing global translational and rotational diffusion. If left unaccounted for, the rotational-vibrational coupling distorts the apparent free energy landscape (FEL), leading to inaccuracies in the inferred potential of mean force (PMF) and the resulting force

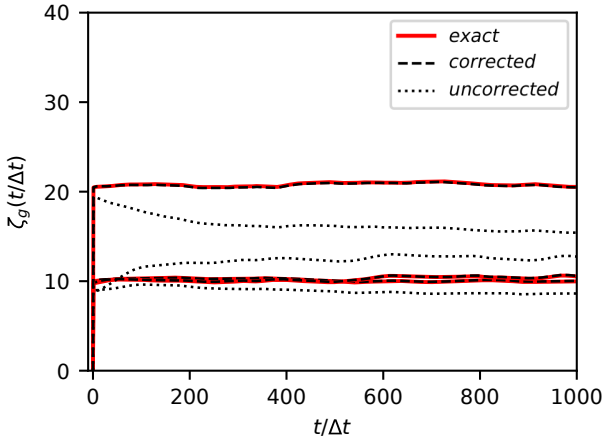

FIG. 3.4: Time-dependent diagonal friction coefficients $\zeta_g(t/\Delta t)$ computed using the method of Ref. [18], based on the exact force field (solid line), the Gaussian Mixture Model (GMM) fit with the rotational entropy correction (dashed line), and the uncorrected GMM fit (dotted line).

field.

To address this, we derived an explicit correction to the free energy landscape that isolates and removes the rotational entropy contribution, yielding a PMF consistent with the true internal energetics of the molecule. This correction is expressed in terms of the generalized moment of inertia tensor $I^*(\vec{q})$, which naturally encodes the dependence of rotational entropy on molecular shape. The corrected PMF can then be used to reconstruct accurate forces and friction coefficients within coarse-grained Langevin dynamics models.

Using a well-controlled toy model—a trimer with analytically specified potential energy, mass, and friction parameters—we validated this approach by comparing the reconstructed force fields and dynamical statistics to known quantities. We showed that inclusion of the rotational entropy correction yields significant improvements in the recovery of the force time series, time correlation functions, and friction coefficients. In contrast, neglecting the correction may lead to observable artifacts, such as spurious memory effects and mis-estimated friction values, even in a simple Markovian setting.

These results underscore that the shape-dependence of rotational entropy is not a negligible artifact but a physically meaningful contribution that can influence both short- and long-time dynamical observables.

By establishing a rigorous formalism and validating it on a tractable model, this work lays the foundation for applying these corrections in more complex settings. In future work, we aim to extend this framework to atomistic molecular dynamics simulations of biomolecules in solution, where the potential of mean force is not known a priori. This will allow us to

quantify the practical significance of rotational entropy effects in systems of biological and chemical interest and to assess the conditions under which such corrections are necessary for reliable CG modeling.

BRIDGE

The Mori-Zwanzig projection operator formalism indicates that, when the orthogonal term is approximated as a gaussian noise, there are two functions that must be obtained in order to fully paramaterize the Zwanzig Generalized Langevin Equation; the potential of mean force, or the corresponding mean force field, which depends on the CG positions, and the memory function, which depends on the CG positions, momenta, and time. In the case of approximate state-independence and over Markovian timescales, the latter may be replaced by a matrix of friction coefficients.

In Chapter 2 we described a method for obtaining these friction coefficients. Now we have supplemented that with a method for obtaining an accurate mean force field in the case of rotationally symmetric systems that are free to diffuse.

In the case of the friction coefficients we found that accurate retrieval required careful consideration of the eigenvalue structure associated with the mix of bound and unbound coordinates due to the translation symmetry.

In this chapter, we have found that the rotational symmetry of the molecule introduces a shape-dependent entropy that must be corrected for. Additionally, we have found Gaussian Mixture Models to be a useful and scalable tool for parametrizing multi-dimensional landscapes without the explicit construction of histograms.

These tools together allow us to retrieve all of the input parameters of a toy model molecule obeying an excact Markovian Langevin equation, a novel capability for a model with these symmetries. The toy model was crucial in verifying the accuracy of each method, as the parametrization of friction coefficients is entangled with that of the force field and exact values were required as reference. We now turn to an application of these methods to a real molecule.

---

[1] B. Roux, Computer Physics Communications **91**, 275 (1995).

[2] D. Trzesniak, A.-P. E. Kunz, and W. F. van Gunsteren, Chemphyschem **8**, 162 (2007).

[3] O. F. Lange and H. Grubmüller, J. Chem. Phys. **124**, 214903 (2006).

[4] J. E. Straub, B. J. Berne, and B. Roux, J. Chem. Phys. **93**, 6804 (1990).

[5] R. B. Best and G. Hummer, J. Phys. Chem. B **113**, 9004 (2009).
[6] R. B. Best and G. Hummer, Proc Natl Acad Sci U S A **107**, 1088 (2010).
[7] H. S. Lee, S. Ahn, and E. F. Darve, J. Chem. Phys. **150**, 174113 (2019).
[8] X. Chapuisat, A. Nauts, and G. Durand, Chemical Physics **56**, 91 (1981).
[9] C. Eckart, Phys. Rev. **47**, 552 (1935).
[10] E. Bright Wilson Jr., J. C. Decius, and Paul C. Cross, (Dover, 1955).
[11] C. E. Chang, M. J. Potter, and M. K. Gilson, J. Phys. Chem. B **107**, 1048 (2003).
[12] C. E. Chang, M. J. Potter, and M. K. Gilson, J. Phys. Chem. B **110**, 7083 (2006).
[13] S. Hikiri, T. Yoshidome, and M. Ikeguchi, Journal of Chemical Theory and Computation **12**, 5990 (2016).
[14] J. D. Louck and H. W. Galbraith, Reviews of Modern Physics **48**, 69 (1976).
[15] D. A. McQuarrie, (University Science Books, 2000).
[16] H. Mori, Prog. Theor. Phys., Japan **33**, 423 (1965).
[17] R. Zwanzig, (Oxford : New York : Oxford University Press, 2001).
[18] J. M. Hall and M. G. Guenza, Phys. Rev. Lett. **134**, 147101 (2025).
[19] M. P. Allen and D. J. Tildesley, (Oxford University Press).
[20] E. Helfand and S. A. Rice, J. Chem. Phys. **32**, 1642 (1960).
[21] N. G. Van Kampen, (Elsevier, 2007).
[22] S. V. Krasnoshchekov, E. V. Isayeva, and N. F. Stepanov, J. Chem. Phys. **141**, 234114 (2014).
[23] C. Yang and J. Hou, IAENG International Journal of Applied Mathematics **43**, 185 (2013).

4 - MODELING THE MARKOVIAN DYNAMICS OF ALANINE DIPEPTIDE

The following is reproduced from an advanced draft of a manuscript for publication elsewhere. The material is co-authored by Jesse Hall and Marina Guenza.

Coarse-grained models describe complex physical systems in terms of a small number of coordinates, aiming to statistically reproduce the dynamics of these coordinates while treating the remaining degrees of freedom only implicitly. For example, a protein in solvent may be described by the motion of the alpha carbons in each amino acid, and the solvent composed of water molecules modeled only implicitly by a drag force and a fluctuating noise, as in a Langevin equation [1]. Such models generally have the advantages of analytical simplicity and computational accessibility relative to their fine-grained counterparts. Achieving their construction without a loss in accuracy is an area of ongoing research.

In particular, there has been much recent activity in the development of parametrization methods for Markovian and Generalized Langevin Equations consistent with the unerlying Mori-Zwanzig projection operator formalism. These models have the potential to be highly accurate, and their formal rooting in statistical-mechanical theory allows for well-defined, controlled approximations.

Most of these treatments have, however, focused on modeling a small number of reaction coordinates in a molecule, often just one [2–6]. High-dimensional CG models may sometimes be desirable. They allow for the study of a molecule's rich mode structure, and permit the study of complex multidimensional pathways without a priori knowledge of good reaction coordinates to describe them. Additionally, high-dimensional models can incorporate the translational and rotational symmetry of the molecule, allowing the study of interactions between global and internal dynamics, which may be important in certain binding events.

In previous work we have demonstrated the ability to recover friction coefficients from the observed trajectories of toy molecules obeying a Generalized Langevin Equation [7]. In Chapter 3 we have seen how the mean force of a tou molecule may be recovered. Most of the obstacles to be overcome in those works consisted of a careful treatment of the rotational symmetry.

In this work we apply these methods for the first time to the trajectory of a real molecule, alanine dipeptide, obtained from an unbiased Molecular Dynamics (MD) simulation with explicit sovent. We construct a site-specific 10-bead model, with an internal 24-dimensional energy landscape and a 10×10 matrix of Markovian friction coefficients. We demonstrate the stability of the Markovian plateau obtained from the Generalized Einstein Relation

as described in [7], and we demonstrate the necessity of the rotational entropy corrections introduced in Chapter 3 of this thesis. Because no enhanced sampling is used, certain barriers are not ideally sampled and there is much room for improvement in the model. Even so, we find the qualitative agreement in the time correlation functions between the MD and CG models remarkable.

Our main goal is to evaluate whether two previously published parametrization techniques, one for the friction coefficients and one for the potential of mean force, constitute significant improvement in the construction such a high-dimensional model for a real molecule. We will examine the stability of the friction coefficients obtained via the generalized Einstein relation with hybrid correlations published in [7], and we will compare the results of using a pmf corrected for rotational entropy or left uncorrected.

I. MOLECULAR DYNAMICS SIMULATION

Alanine dipeptide in TIP3P water solvent [8] was simulated in GROMACS 2020.4 [9] using the AMBER99SB-ILDN force field [10]. The ions Na^+ and Cl^- were added to achieve a salt concentration of 0.1M. The peptide was relaxed in the presence of the solvent and ions using a steepest-descent algorithm. Two 1ns equilibration runs were performed in the NVT and NPT ensembles respectively at a timestep of 2fs. Both were performed at a temperature of 298K and the latter at a pressure of 1atm, both using a Berendsen thermostat. For each of ten replicas, 10ns of equilibration was performed and discarded using the velocity-rescaling thermostat, again at 298K, with a timestep of 1fs. The final ten production runs were 490ns each under the same conditions, for a total simulation time of 4.9μs. Hydrogen bond lengths were constrained in all simulations. Simulations were performed on the Talapas High-Performance Computing cluster at the University of Oregon.

II. DISTRIBUTION OF INTERNAL COORDINATES

Alanine dipeptide consists of 22 atoms, 12 of which are hydrogen atoms. In this treatment, we will join every hydrogen atom to its heavy bonded partner to form a united atom, and take the center-of-mass of the resulting cluster as a CG site. This results in a $N = 10$ bead model of alanine dipeptide, with 30 degrees of freedom.

In order to retain translational and rotational symmetry, the potential of mean force for an N-bead CG model must depend only on the $3N - 6$ (in this case 24) internal degrees of freedom of the system.

As discussed in Chapter 3, the internal shape coordinates and the body-fixed frame used to calculate them may be chosen carefully to greatly simplify the expression for the effects

of rotational entropy on their distribution. As such, we will use the same scheme for the internal coordinates deployed in that work.

Let $\vec{r}$ be the lab-frame positions of the N beads. Then let $\vec{r}_B$ be the body-fixed positions as defined by the Eckart frame with reference structure $\vec{r}_0$. There is freedom in the choice of $\vec{r}_0$, but it will be convenient for it to be chosen such that the fluctuations in the displacements from $\vec{r}_0$ are zero-centered. We achieve this by choosing the initial conformation, reoriented into a principal axes frame, as an arbitrary $\vec{r}_0'$, then doing a trial removal of translation and rotation from the entire trajectory $\vec{r}(t)$. The calculation of the Eckart frame Euler angles from $\vec{r}(t)$ and $\vec{r}_0'$ is performed using the quaternion algorithm in [11]. The resulting $\vec{r}_B'$ now has a well-defined average conformation. Orienting this $\langle \vec{r}^{B'} \rangle$ into a principal axes frame, we choose this conformation as our reference structure $\vec{r}_0$.

Once this $\vec{r}_0$ has been constructed, it is used to perform the final translation and rotation removal from $\vec{r}(t)$. This results in a body-fixed trajectory $\vec{r}_B(t)$ such that the conformational fluctuations are centered around $\vec{r}_0$.

In this work, as in Chapter 3, we define the internal shape coordinates $\vec{q}$ to be the first $3N-6$ elemements of $P(\vec{r}_B - \vec{r}_0)$ where P is the permutation matrix

$$P_{ij} = \begin{cases} \delta_{ij}, & i \notin (3N-6,\ 3N-5) \\ \delta_{i+1,j} & i = 3N-6 \\ \delta_{i-1,j} & i = 3N-5 \end{cases} \tag{1}$$

With this permutation matrix the last bead N is omitted from the shape coordinates entirely, bead $(N-1)$ retains only an x-coordinate in the body-fixed frame, and bead $(N-2)$ retains its x- and y-coordinates. This staggered set of internal displacements prevents degeneracies in the calculation of the matrix $\left(\frac{d\vec{r}}{d\vec{q}}\right)$, which will be needed in the next section.

To construct the CG force, we will need to calculate the gradient of the probability distribution $\mathcal{P}(\vec{q})$ for any given $\vec{q}$. This requires an analytical fit to the observed $\mathcal{P}(\vec{q})$. Due to the high-dimensional nature of the distribution, any methods involving the explicit construction of a histogram are impractical. We will instead, as in Chapter 3, attempt to construct a Gaussian Mixture Model (GMM) for $\mathcal{P}(\vec{q})$ directly from the trajectory data. We use the Expectation-Maximization (EM) algorithm to construct an approximate Maximum Likelihood Estimation for the parameters of the GMM modeling $\mathcal{P}(\vec{q})$ as the sum of a given number n of gaussians.

The number n of gaussians is a hyperparameter which must be chosen before the EM algorithm is initiated, and so to choose a good value for n we must perform many fits with different values of n and then check some other measure to discern the best fit. Simply increasing n almost always improves the final value of the likelihood, even when overfit-

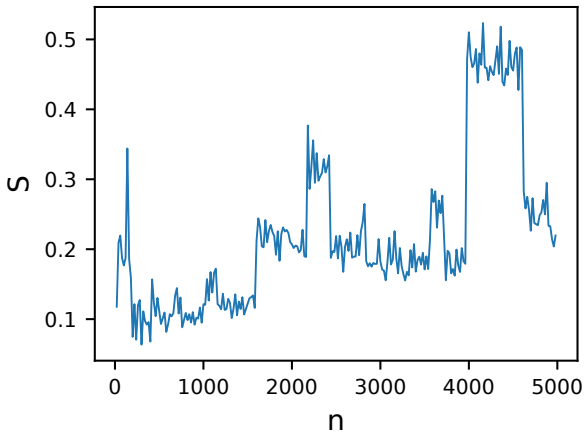

FIG. 4.1: The equipartition scores S obtained from Gaussian Mixture Models of varying component number n. A minimum score is achieved at $n = 270$.

ting is occurring. Instead the measure we choose is the error in satisfying the generalized equipartition theorem

$$\langle q_i \frac{\partial \mathcal{H}}{\partial q_j} \rangle = k_B T \delta_{ij} \tag{2}$$

over the original trajectory. Since the body-fixed coordinate system is obtained from the lab frame coordinates via a canonical transformation, the above equation implies that

$$-k_B T \left\langle q_i \frac{1}{\mathcal{P}(\vec{q})} \frac{\partial \mathcal{P}(\vec{q})}{\partial q_j} \right\rangle = k_B T \delta_{ij} \tag{3}$$

which can be checked using the fit for $\mathcal{P}(\vec{q})$.

We define a score $\mathcal{S}$ such that

$$\mathcal{S} := \left| -k_B T \left\langle \vec{q} \frac{1}{\mathcal{P}(\vec{q})} \frac{\partial \mathcal{P}(\vec{q})}{\partial \vec{q}}^T \right\rangle - k_B T \mathbb{I} \right| \tag{4}$$

where the vertical bars represent the Frobenius norm. A particular fit is considered good when this score is small as calculated on the observed trajectory. We note that it is trivially equal to zero if the samples are instead drawn from the GMM itself.

In Fig. 1 we plot the scores of a range of fits, with the number n of gaussian components ranging from 10 to 5000. We find that the best score is achieved with a relatively small number of gaussians, and the agreement with the equipartition theorem generally worsens beyond $n = 270$, possibly due to overfitting. This strategy is highly heuristic, but we have by experience found that minimizing this score leads to greatly improved friction coefficients later in the coarse-graining process. Force fields with a high $\mathcal{S}$ score almost invariably yield

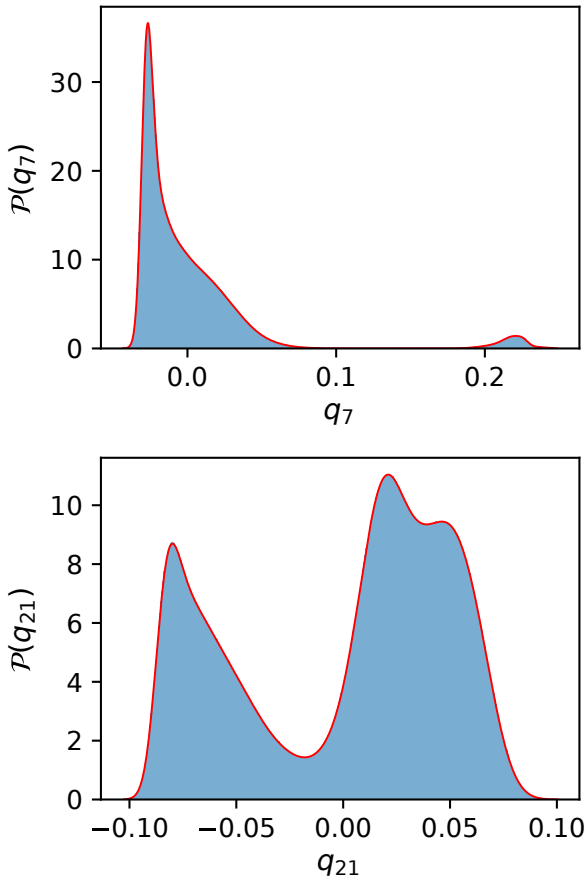

FIG. 4.2: Histograms (transparent) of shape coordinates obtained from the MD simulation of alanine dipeptide, overlaid with the corresponding marginal distributions predicted by a 270-component Gaussian Mixture Model (GMM) fit (solid lines).

friction matrices with unphysical negative eigenvalues. A more rigorous understanding of how to select a particular GMM would be beneficial.

In Fig. 2 we plot histograms of select shape coordinates from the MD simulation, together with the curves of marginal distributions calculated from the $n = 270$ GMM. The visual agreement in regions of high probability is very good across all 24 internal coordinates. A few hundred gaussians seem largely adequate to describe the basins of even a high-dimensional molecular landscape.

III. THE POTENTIAL OF MEAN FORCE AND ROTATIONAL ENTROPY

Once a GMM for $\mathcal{P}(\vec{q})$ has been selected as described in the previous section, the potential of mean force must be isolated from distortions of the free energy landscape due to shape-

dependent rotational entropy. As we demonstrated in Chapter 3, the potential of mean force can be isolated by the formula

$$V(\vec{q}) = -k_B T\, ln(\mathcal{P}(\vec{q})) + k_B T\, ln(A\sqrt{|I^*(\vec{q})|}). \tag{5}$$

where the generalized moment of inertia I^* is defined as

$$I^* := I - CS^{-1}C^T \tag{6}$$

with

$$C_{\alpha i} := \sum_k m_k(\vec{r}_k \times \frac{\partial \vec{r}_k}{\partial q_i})_\alpha. \tag{7}$$

$$S_{ij} := \sum_{k,\alpha} m_k \frac{\partial r_{k\alpha}}{\partial q_i}\frac{\partial r_{k\alpha}}{\partial q_j}. \tag{8}$$

and I the standard moment of inertia matrix. (Throughout this manuscript, Greek indices will indicate Cartesian components.)

The force field corresponding to $V(\vec{q})$ is then obtained from $\mathcal{P}(\vec{q})$ via

$$-\frac{\partial V(\vec{q})}{\partial q_i} = k_B T\, \frac{1}{\mathcal{P}(\vec{q})}\frac{\partial \mathcal{P}(\vec{q})}{\partial q_i} - \frac{k_B T}{2} Tr\big(I^{*^{-1}}(\vec{q})\frac{\partial I^*(\vec{q})}{\partial q_i}\big). \tag{9}$$

where Jacobi's formula has been used for the derivative of the determinant. This formula is valid for shape coordinates and frame constraints such that $\frac{\partial \vec{r}}{\partial \vec{q}}$ is a constant. In Appendix B we demonstrate that this holds true for the reduced displacements relative to an Eckart frame reference structure that we have previously described.

From this formula a time series of generalized forces can be obtained from the observed $\vec{q}$-trajectory. From these $3N-6$ generalized forces, the numerical procedure described in Appendix B allows for the recovery of the full $3N$ lab-frame forces on each CG bead. This is accomplished by enforcing conditions of zero net torque and net force on the intramolecular forces consistent with the required translational and rotational symmetry of the potential. The numerical procedure to accomplish this is detailed in Appendix B.

IV. FRICTION COEFFICIENTS

In a many-bead CG model of a molecule, the site-specific friction coefficients are responsible for the kinetics of both the internal shape changes and the overall translational and rotational diffusion of the molecule. In previous work [7] we have shown how Markovian frictions coefficients simultaneously consistent with all of these behaviors on long timescales can be obtained in a manner consistent with an underlying memory function. This method made use of a Generalized Einstein Relation (GER)

$$\zeta_g(t) = \left[MC^{gv}(0) - MC^{gv}(t) + \mathcal{F}^{gU'}(t)\right] \mathcal{D}^{gv^{-1}}(t), \tag{10}$$

where

$$C_{ij}^{xy}(t) := \langle \vec{x}_i(0) \cdot \vec{y}_j(t) \rangle \tag{11}$$

$$\mathcal{F}^{vU'}(t) = -\int_0^t d\tau\, C^{vU'}(\tau) \tag{12}$$

$$\mathcal{D}^{vv}(t) := \int_0^t d\tau\, C^{vv}(\tau)\ , \tag{13}$$

In this relation, the long-time limit of ζ_g yields the Markovian friction coefficients for any choice of $\vec{g}$

$$\zeta_g(\infty) = G(\infty) = \zeta \tag{14}$$

but it was found that due to the simultaneous presence of bound internal coordinates and unbound center-of-mass diffusion, the particular choice

$$\vec{g} := \begin{bmatrix} \vec{r}'_C - \vec{v}'_L \\ \\ \vec{V} \end{bmatrix} \tag{15}$$

yielded far better numerical stability than most choices due to the eigenvalue structure of its correlations at short and long times. Here $\vec{V}$ is the center-of-mass velocity and $\vec{r}'_C$ is the vector of CG position coordinates relative to the center-of-mass with the x, y, z coordinates of the Nth CG site removed, such that $\vec{r}'_C = (x_1^C, y_1^C, z_1^C, ..., x_{n-1}^C, y_{n-1}^C, z_{n-1}^C)$, and similarly for $\vec{v}'_L$.

Due to the presence of $\mathcal{F}$ in the numerator, the mean force must be reconstructed before the friction coefficients can be obtained, and the accuracy of the friction is critically dependent on the accuracy of the modeled force. In Chapter 3, we found in a toy model that

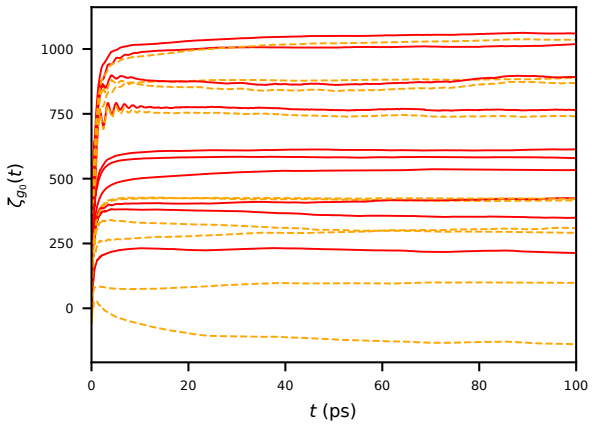

FIG. 4.3: The diagonal components of the friction obtained using the entropically corrected (solid) and uncorrected (dashed) force fields.

the rotational entropy correction was required for reasonably accurate reconstruction of the friction.

Once an estimated time series for the mean force $\vec{F}(\vec{q}(t)) = -\frac{\partial V(\vec{q}(t))}{\partial \vec{q}}$ has been constructed, the lagged time correlation $C^{gU'}(t)$ can be obtained.

In Fig. 4.3, we show the resulting $\zeta_g(t)$ using the entropy-corrected and uncorrected force fields. We see that the approach does indeed yield stable plateau values in both cases. However, as shown in Fig. 4.4, the uncorrected forces yield a Markov friction matrix with negative eigenvalues. This is not usable in a CG model as the trajectory would accumulate energy rather than dissipate it. Interestingly, it is the low-friction modes that are most significantly perturbed by the rotational effects, while the greatest eigenvalues are largely unaltered. The corrected force field, on the other hand, yields a positive definite friction matrix. Additionally, the off-diagonal components of the friction matrix are far more symmetric when using the corrected forces. In Fig. 4.5 we can see that the corrected force field yields a far more structured friction matrix.

These results are striking, and indicate significant effects from rotational-vibrational coupling in real molecules. The correction itself is a very small fraction of the instantaneous force, yet its effects alter the force TCF at long times and affect many eigenvalues of the friction matrix significantly. Our previous work only demonstrated the necessity of this correction in the case of a particular toy model. Here we see that they remain crucial in the case of a small peptide in explicit solvent.

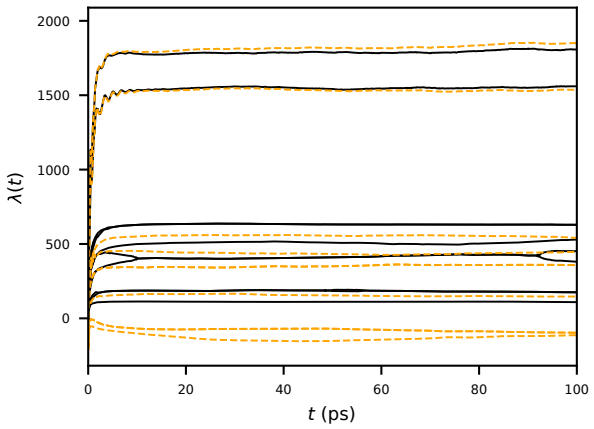

FIG. 4.4: The eigenvalues of the friction obtained using the entropically corrected (solid) and uncorrected (dashed) force fields.

V. COARSE-GRAINED SIMULATION

With a force field and friction coefficients in hand, a Langevin equation for the 10-bead model of alanine dipeptide can be constructed and simulated in the lab-frame.

The Markovian Langevin equation of the CG model is given by

$$\begin{aligned} \dot{\vec{r}}(t) &= M^{-1}\vec{p}(t) \\ \dot{\vec{p}}(t) &= -\frac{\partial V(\vec{q}(t))}{\partial \vec{r}} - \zeta\vec{p}(t) + f^R(t) \end{aligned} \tag{16}$$

The cross-bead correlations in the noise force are obtained from the second Fluctuation-Dissipation Theorem via the matrix square-root of ζ such that

$$\langle f^R_{i\alpha}(t) f^R_{j\beta}(t')\rangle = 2k_B T\zeta_{ij}\delta_{\alpha\beta}\delta(t-t'). \tag{17}$$

The masses of the beads are set equal to the total mass of their corresponding cluster of atoms. The drag force at each time step is calculated using the full friction coefficient matrix obtained from the GER.

The force field we obtained directly yields $3N-6$ generalized forces in the body-fixed frame as a function of $\vec{q}$. In order to propagate the simulation, then, the body-fixed $\vec{q}$-coordinates must be calculated at each frame. Fromt these the generalized forces are calculated, and then the method from Appendix B is used to expand them into a full set of $3N$ forces on the beads in the body-fixed frame. Finally the rotation is undone to yield the lab-frame forces.

This frame-switching scheme slows down the simulation, and could be avoided if an entropy-corrected force field were developed in terms of rotationally invariant coordinates

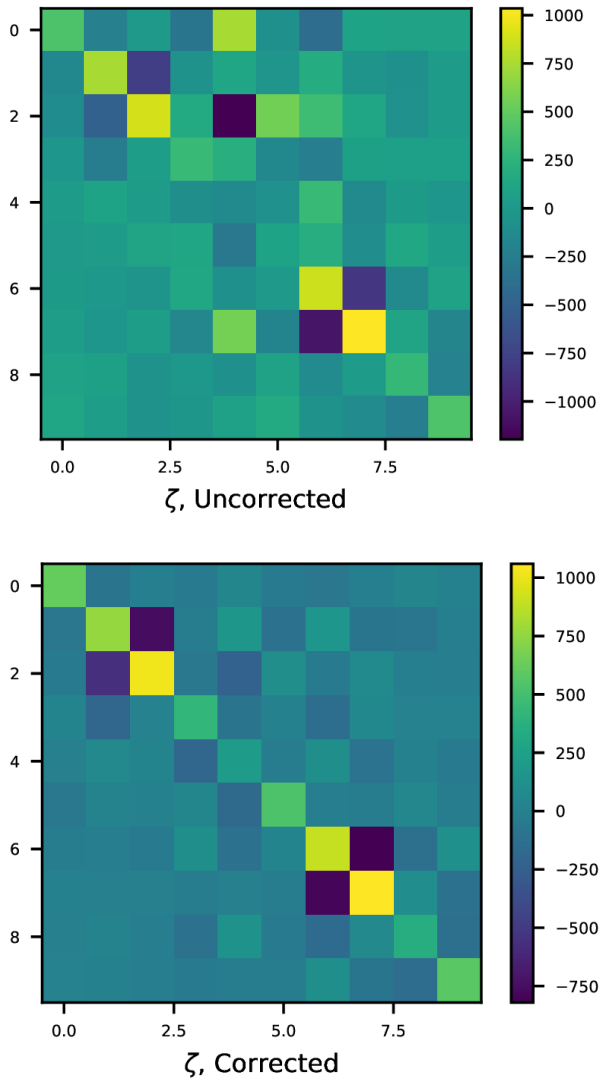

FIG. 4.5: The friction matrices drawn from $\zeta_{g_0}(t)$ at t =100ps using the entropically uncorrected (top) and corrected (bottom) force fields.

such as the popular Bond-Angle-Torsion (BAT) coordinates. However, as described in Chapter 3, such coordinates result in a complicated nonlinear metric which must be corrected for in the force field, so there are trade-offs involved. In this work we accept the inefficiency of frame-switching as the focus of our study is the accuracy of the coarse-graining procedure itself.

The coarse-grained model was used to simulate 100 replicas of duration 50ns at a timestep 1fs. The first 1ns of each trajectory was discarded as equilibration, leaving a total production simulation time of 4.9μs.

VI. RESULTS

Lagged auto- and cross-correlations $C_{ij}(t)$ in the shape coordinates $\vec{q}$ were calculated for both the MD and CG trajectories, such that

$$C_{ij}(t) := \langle q_i(0) q_j(t) \rangle \tag{18}$$

The correlations were calculated for each replica and the results averaged over all the replicas at each time lag. To account for the lack of convergence in finite simulations, margins of error were calculated by obtaining the standard deviation across the replicas for each set and then arriving at the expected standard deviation of the mean by the formula

$$\sigma_n = \sigma / \sqrt{n}. \tag{19}$$

The error bars plotted in Fig. 4.6 represent two such standard deviations.

Some representative comparisons between the TCFs of the MD and CG simulations are plotted in Fig. 4.6. The qualitative agreement is remarkable for auto- and cross-correlations throughout the molecule, with pronounced transitions in decay rates occurring at the correct time lags and correlation amplitudes. Some complex short-time features in the MD TCFs are captured remarkably well by the CG model even when they occur on sub-Markovian timescales. Additionally, many of the coordinates exhibit extremely good quantitative agreement between the CG and MD correlation functions at all observed timescales.

The CG model, however, does not quantitatively reproduce the long-time decay rates of some coordinates, with a tendency for the most correlated motions to decay too slowly in a manner that does not fall within the apparent margin of error.

VII. DISCUSSION

Prior coarse-grained models of alanine dipeptide have modeled only a small number of purely internal reaction coordinates [6]. Our approach is, to our knowledge, unique in its ability to capture the dynamics many internal degrees of freedom along with global translational and rotational symmetry of the molecule in a manner explicitly consistent with an underlying GLE. We have found that the qualitative agreement in the dynamics is exceptional, but that there is significant room for improvement in the quantitative accuracy of the dynamics of certain highly correlated motions.

It is our belief that since the MD simulations used as inputs here were unbiased, a major source of the discrepancy is likely due to a failure to obtain adequate sampling of the largest free energy barrier in the system, which is known to be on the order of $10 k_{BT}$ for solvated

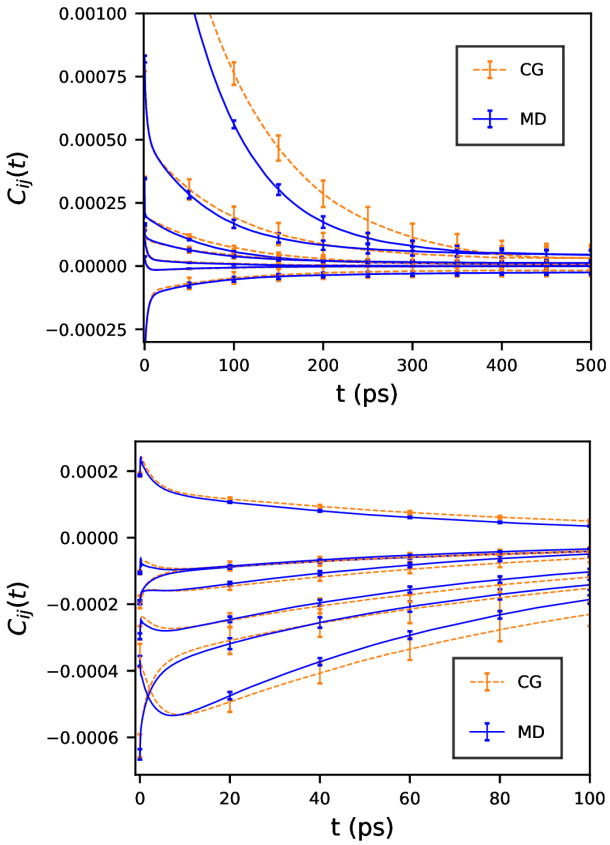

FIG. 4.6: A selection of representative comparisons between $\vec{q}$-coordinate auto- and cross-correlations in the MD (solid lines) and CG (dashed lines) models. The top panel depicts the pairs $(i,j) \in \{(0,0),(18,18),(0,3),(11,12),(14,23),(3,14),(21,21)\}$ while the bottom panel depicts the pairs $(i,j) \in \{(15,21),(14,21),(2,14),(3,20),(1,22),(0,8),(2,11)\}$

alanine dipeptide [12, 13]. While the frequently sampled basins of the free energy landscape are well described by our Gaussian Mixture Model, relatively small errors on poorly sampled saddle points can have large effects on transition timescales.

The incorporation of enhanced sampling methods into this recently developed coarse-graining procedure lies beyond the scope of this manuscript. Here we have been primarily concerned with the demonstration of two points; first, the stability of the Markovian plateau obtained via the Generalized Einstein Relation in the case of a real molecule in explicit solvent; and second, the necessity of rotational entropy corrections when developing force-fields for CG models that include the free rotation of the molecule.

We find the results to be quite promising, and future work will be focused on incorporating enhanced sampling techniques into this coarse-graining procedure to improve the quantitative agreement in the dynamics. Most enhanced sampling techniques are designed with a small

number of reaction coordinates in mind, and so may need adjustment for practical use in high-dimensional coarse-graining.

This work represents an advancement in the capabilities of coarse-grained models that could have applications in the study of multi-dimensional phenomena in diffusing molecules. The demonstration of significant dynamical implications of rovibrational coupling in a small peptide is also of interest, as previous studies have failed to identify rovibrational coupling in proteins using equilibrium statistics [14]. These tools lay groundwork for further study of such effects.

---

[1] E. Beyerle and M. Guenza, The Journal of Chemical Physics **154.12** (2021).

[2] O. F. Lange and H. Grubmüller, J. Chem. Phys. **124**, 214903 (2006).

[3] J. E. Straub, B. J. Berne, and B. Roux, J. Chem. Phys. **93**, 6804 (1990).

[4] R. B. Best and G. Hummer, J. Phys. Chem. B **113**, 9004 (2009).

[5] R. B. Best and G. Hummer, Proc Natl Acad Sci U S A **107**, 1088 (2010).

[6] H. S. Lee, S. Ahn, and E. F. Darve, J. Chem. Phys. **150**, 174113 (2019).

[7] J. M. Hall and M. G. Guenza, Phys. Rev. Lett. **134**, 147101 (2025).

[8] P. Mark and L. Nilsson, The Journal of Physical Chemistry A **105**, 9954 (2001).

[9] Lindahl, Abraham, Hess, and van der Spoel, Zenodo https://doi.org/10.5281/zenodo.4054996 (2020).

[10] J. Ponder and D. Case, Adv. Prot. Chem. **66**, 27 (2003).

[11] S. V. Krasnoshchekov, E. V. Isayeva, and N. F. Stepanov, J. Chem. Phys. **140**, 154104 (2014).

[12] P. E. Smith, J. Chem. Phys. **111**, 5568–5579 (1999).

[13] J. Apostolakis, P. Ferrara, and A. Caflisch, J. Chem. Phys. **110**, 2099–2108 (1999).

[14] M. B. Hamaneh, L. Zhang, and M. Buck, Biophysical journal **101**, 196 (2011).

5 - CONCLUSION

In Chapter 2, we described a generalization of the Einstein Relation to allow calculation of Markovian friction coefficients using a broad class of observable lagged correlations. The flexibility of this expression was exploited in order to develop a numerically stable expression for the friction coefficients in the case of models with mixed symmetry, such as unbound translational diffusion and bound rotational and vibrational diffusion.

In Chapter 3, we develop an expression for the distribution of shape coordinates in systems with rotational symmetry. This expression allows the potential of mean force in such models to be isolated from the effects of rotational entropy on the free energy landscape. Only with the results of Chapters 2 and 3 together are we able to fully recreate a toy model with translational and rotational freedom from its trajectory.

Finally, in Chapter 4, we demonstrated these techniques in a realistic system, solvated alanine dipeptide. The resulting coarse-grained model demonstrated remarkable qualitative agreement, and excellent quantitative agreement across many degrees of freedom. It was also shown that incorporating the rotational entropy correction was vital to recovering a usable many-bead model, as the friction coefficients yielded negative eigenvalues otherwise. Residual error in the most highly correlated motions were likely due to inadequate sampling over a high barrier in the syste.

These techniques are complete in the sense that they can accurately and completely recover the parameters of a system that precisely obeys a state-independent GLE. This is an achievement in that previous methods did not scale well to high-dimensional projected spaces, and were inappropriate to systems with a mix of free and bound coordinates.

Such a foundation indicates several obvious routes for future work.

The Zwanzig-projected GLE used here should generally include state-dependence as well as time dependence. These are currently the only major features of the complete formal GLE that are missing from the models presented in this dissertation.

Additionally, while the procedures described here may be complete in the sense that they accurately recover parameters from accurate statistics, they do not account for the difficulties in obtaining accurate statistics from unbiased MD simulations in the first place. This is made quite clear by the difficulty in modeling the high energy barrier in alanine dipeptide. In order for these methods to become practical tools in biophysics, they must be able to incorporate enhanced sampling methods. Most enhanced sampling methods are suitable primarily for low-dimensional reaction coordinate spaces, and so further work may be required in order to develop a protocol that incorporates enhanced sampling techniques into our coarse-graining strategies.

Taken together, these methods comprise a marked advancement in the capabilities of

coarse-grained models of diffusing molecules. Such models permit the investigation of physical phenomena for which analytically rigorous tools have not previously existed, such as rovibrational coupling in such systems. We look forward to future applications of our work, which has brought us to the cusp of understanding the complex fluctuations of diffusing molecules a little better.

APPENDIX A

SUPPLEMENTAL MATERIAL FOR GENERALIZED EINSTEIN RELATION FOR MARKOVIAN FRICTION COEFFICIENTS FROM MOLECULAR TRAJECTORIES

This Appendix consists of the Supplemental Material published with *Generalized Einstein Relation for Markovian Friction Coefficients from Molecular Trajectories* in volume 134 of *Physical Review Letters* on April 7, 2025, and authored by J.M. Hall and M.G. Guenza.

I. ORTHOGONALITY OF NOISE

The Generalized Langevin Equation (GLE) can be rigorously justified using the Mori-Zwanzig projection operator formalism. Here we follow the treatment using the Zwanzig operator [1, 2], which results in a GLE with nonlinear potential of mean force. Our aim is to justify the orthogonality of the GLE noise to all vectors $\vec{g}(\vec{r},\vec{p})$, where $\vec{r}$ and $\vec{p}$ are the coarse-grained (CG) coordinates, in the state-independent approximation.

Given a many-body Hamiltonian $\mathcal{H}(\vec{z})$, where $\vec{z}$ is the microscopic phase space, the Zwanzig projection operator $P_{\vec{\alpha}}$ is defined as the conditional average

$$P_{\vec{\alpha}}F = \frac{1}{\Omega(\vec{\alpha})}\int d\vec{z}F(\vec{z})\rho^{eq}(\vec{z})\delta(\vec{A}(\vec{z})-\vec{\alpha}) \tag{1}$$

where $\rho^{eq}(\vec{z})$ is the equilibrium probability distribution of $\vec{z}$, $\vec{A}(\vec{z})$ are the collection of CG variables (in our case $\vec{r}(\vec{z}),\vec{p}(\vec{z})$), and $\Omega(\vec{\alpha})$ is the coarse-grained entropy:

$$\Omega(\vec{\alpha}) = \int d\vec{z}\rho^{eq}(\vec{z})\delta(\vec{A}(\vec{z})-\vec{\alpha}). \tag{2}$$

The dirac delta makes the Zwanzig projection a conditional average over microscopic configurations $\vec{z}$ consistent with the CG state $\vec{A}(\vec{z})=\vec{\alpha}$. Note that $P^2_{\vec{\alpha}} = P_{\vec{\alpha}}$, as required of a projection operator.

We note that in the case that $F(\vec{z}) = G(\vec{A}(\vec{z}))H(\vec{z})$, $G(A(\vec{z}))$ can come out of the integral due to the dirac delta and we have

$$P_{\vec{\alpha}}\, G(\vec{A}(\vec{z}))H(\vec{z})) = G(\vec{\alpha})P_{\vec{\alpha}}(H(\vec{z})). \tag{3}$$

The orthogonal projection operator is defined as

$$Q_{\vec{\alpha}} := 1 - P_{\vec{\alpha}}. \tag{4}$$

Upon alternating projections, we have

$$\begin{aligned} P_{\vec{\alpha}} Q_{\vec{\alpha}} F(\vec{z}) &= P_{\vec{\alpha}}(1 - P_{\vec{\alpha}}) F(\vec{z}) \\ &= P_{\vec{\alpha}} F(\vec{z}) - P_{\vec{\alpha}}^2 F(\vec{z}) \\ &= 0. \end{aligned} \tag{5}$$

Following Zwanzig, the noise term in the GLE is defined as

$$\begin{aligned} \vec{\vec{R}}(t, \vec{z}) &:= exp(t\, Q_{\vec{A}(\vec{z})} L) Q_{\vec{A}(\vec{z})} L \vec{A}(\vec{z}) \\ &= Q_{\vec{A}(\vec{z})} L\, exp(t\, Q_{\vec{A}(\vec{z})} L) \vec{A}(\vec{z}) \end{aligned} \tag{6}$$

where L is the Liouville operator associated with $\mathcal{H}(\vec{z})$. This noise depends explicitly on the microscopic variables $\vec{z}$, but has statistics governed by the CG variables according to the Second Fluctuation-Dissipation theorem and so is modeled as a random variable in the GLE. In the case where the memory kernel is state-independent, this noise $\vec{\vec{R}}(t, \vec{z})$ is equivalent to $\vec{f}^R(t)$ in the main text.

Because the noise can be written with a Q-projection in front, its conditional average over the CG variables is zero by Eq. 5.

$$P_{\vec{A}} \vec{\vec{R}}(t, \vec{z}) = 0, \tag{7}$$

where we have dropped the explicit $\vec{z}$-dependence of $\vec{A}$ for clarity. Combining this result with Eq. 3, we find that the noise is also orthogonal to any function of the CG variables in the sense that

$$P_{\vec{A}} \left(G(\vec{A}(\vec{z})) \widetilde{R}_j(t, \vec{z}) \right) = 0 \tag{8}$$

for all $G(\vec{A}(\vec{z}))$ and all j.

In general the statistics of $\vec{\vec{R}}(t, \vec{z})$ are state-dependent and so this result only holds for conditional averages where the CG variables $\vec{A}$ are held constant. If the memory function (and consequently the noise) happens to be state-independent, then Eqs. (7) and (8) hold even over unconditional averages. So, to the extent that the state-independent GLE is valid,

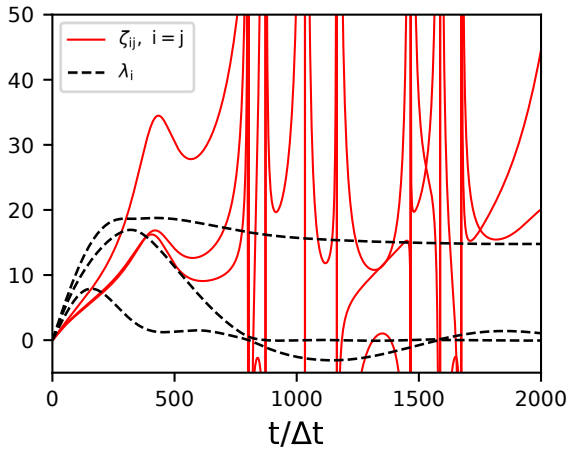

FIG. A1. The diagonal components of the calculated $\zeta_v(t)$ (solid) alongside the eigenvalues λ_i of $\mathcal{D}^{vv}(t)$ (dashed).

we have

$$\langle G(\vec{A}(\vec{z}))f_j^R(t)\rangle = 0, \tag{9}$$

which justifies dropping the noise term in the integral equations of the main text when multiplying the GLE by any vector of the form $\vec{g} = \vec{g}(\vec{r}, \vec{p})$.

II. ALTERNATIVE SELECTIONS FOR THE G-VECTOR

In the main text it is claimed that $\zeta_v(t)$ diverges in the limit $T \to \infty$ if $\vec{g} = \vec{v}$ is used in the generalized Einstein relation. Here we show how this is the case. Since the velocity autocorrelations decay to zero as $t \to \infty$, the numerator of $\zeta_v(\infty)$ simplifies to

$$\zeta_v(\infty)\mathcal{D}^{vv}(\infty) = MC^{vv}(0) + \mathcal{F}^{vU'}(\infty). \tag{10}$$

The diagonal of the first term is equal to twice the average kinetic energies of the particles. From the definition of $\mathcal{F}(t)$, we also have

$$\mathcal{F}^{vU'}(\infty) = -\int_0^\infty d\tau\, C^{vU'}(\tau) = C^{rU'}(0), \tag{11}$$

for off-diagonal components as well as the diagonals, under the assumption that positions and intramolecular forces are uncorrelated over arbitrarily long timescales. In the limit $t \to \infty$,

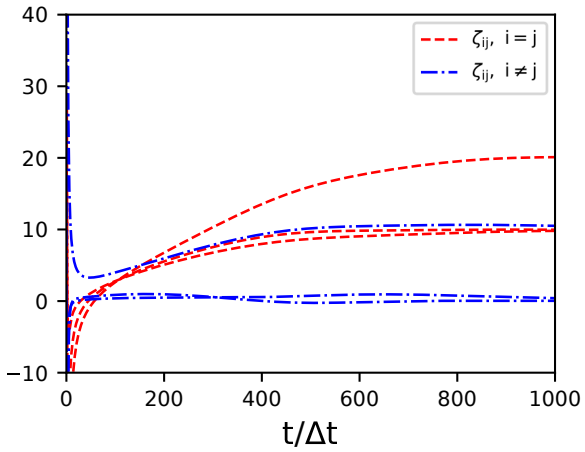

FIG. A2. The diagonal (dashed) and off-diagonal (dot-dashed) components of $\zeta_{g_0}(t)$

then, all eigenvalues of the numerator of $\zeta_v(t)$ are finite, while all eigenvalues of $\mathcal{D}^{vv}(t)$ associated with internal motion go to zero at large lag times, except for the one associated with global translation. In addition, the velocity-velocity correlations oscillate when underdamped modes are present, leading to instantaneous zero eigenvalues. The resulting $\zeta_v(t)$ is plagued with artificial singularities, as seen in Fig. A1, and any estimate of the plateau becomes wholly unreliable.

In contrast, the choice

$$\vec{g}_0 := \begin{bmatrix} \vec{r'}_C \\ \\ \vec{V} \end{bmatrix} \tag{12}$$

stabilizes the plateau at large t. In practice this is a suitable choice of $\vec{g}$ for the recovery of Markovian friction coefficients. However, this choice leads to the appearance of a singularity in the $t \to 0$ limit as discussed in the main text, and seen here in Fig. A2, which is nonetheless undesirable.

---

[1] R. Zwanzig, (Oxford : New York : Oxford University Press, 2001).

[2] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, Faraday Discuss. **144**, 301 (2010).

APPENDIX B

SUPPLEMENTAL MATERIAL FOR COARSE-GRAINED FORCE FIELDS VIA ROTATIONAL ENTROPY CORRECTIONS TO FREE ENERGY LANDSCAPES OF DIFFUSING MOLECULES

This appendix consists of the Supplemental Material from the manuscript currently under review for publication, authored by J.M. Hall and M.G. Guenza, which is the subject of Chapter 3.

I. CALCULATION OF $\frac{\partial \vec{r}}{\partial \vec{q}}$

The calculation of the matrix $\frac{\partial \vec{r}}{\partial \vec{q}}$ is generally nontrivial, but a simple procedure exists in the case of linear frame constraints and rectilinear internal coordinates $\vec{q}$. In this case the matrix is constant and can be constructed directly from the constraint matrix through the procedure outlined here.

For the class of body-fixed frames defined by linear equations of constraint, we define the $6 \times 3N$ constraint matrix E (suggesting an Eckart frame as a salient example) [1]:

$$E\vec{\Delta r} = 0 \tag{1}$$

For example, the first three rows of Eq. 1 can be taken to be the equations setting the center-of-mass to zero, and the last three rows may be the Eckart constraints.

When we go to the body-fixed frame, our internal coordinates are reduced by six. In order to calculate $\frac{\partial \vec{r}}{\partial \vec{q}}$, we must know how these six variables are fixed by the remaining $3N - 6$. First, we use a permutation matrix P to rearrange the vector $\vec{r}$ such that the coordinates to be removed are at the end:

$$EP \begin{bmatrix} \vec{\Delta r}_K \\ \vec{\Delta r}_R \end{bmatrix} = 0 \tag{2}$$

Where the $\vec{r}_K$ are all of the variables that will be *kept* free and the $\vec{r}_R$ are the constrained variables to be *removed* in favor of the frame origin and Euler angles. We then partition the matrix EP:

$$\begin{bmatrix}(EP)_K & (EP)_R\end{bmatrix} \begin{bmatrix}\vec{\Delta r}_K \\ \vec{\Delta r}_R\end{bmatrix} \tag{3}$$

This partitioned form can be expanded:

$$(EP)_K \vec{\Delta r}_K + (EP)_R \vec{\Delta r}_R = 0 \tag{4}$$

Now we rearrange and solve for $\Delta\vec{r}_R$ in terms of $\Delta\vec{r}_K$.

$$\vec{\Delta r}_R = -(EP)_R^{-1}(EP)_K \vec{\Delta r}_K \tag{5}$$

Now that we have equation 5, we can express all of our position coordinates in terms of only the internal degrees of freedom that remain after our constraints are applied:

$$\begin{aligned}\vec{\Delta r} &= P \begin{bmatrix}\vec{\Delta r}_K \\ \vec{\Delta r}_R\end{bmatrix} \\ &= P \begin{bmatrix}I_{3N-6} \\ -(EP)_R^{-1}(EP)_K\end{bmatrix} \vec{\Delta r}_K\end{aligned} \tag{6}$$

If the $\vec{q}$ coordinates are linear in the body-fixed displacements from the reference structure, then

$$\begin{aligned}\vec{q} &= Q\Delta\vec{r}_K \\ \implies \Delta\vec{r}_K &= Q^{-1}\vec{q}\end{aligned} \tag{7}$$

Plugging this into equation (137),

$$\begin{aligned}\vec{r} - \vec{r}_0 &= P \begin{bmatrix}I_{3N-6} \\ -(EP)_R^{-1}(EP)_K\end{bmatrix} Q^{-1}\vec{q} \\ \implies \left(\frac{\partial\vec{r}}{\partial\vec{q}}\right)^T &= P \begin{bmatrix}I_{3N-6} \\ -(EP)_R^{-1}(EP)_K\end{bmatrix} Q^{-1}\end{aligned} \tag{8}$$

Where $\vec{r}_0$ is the constant reference structure. The transpose is ultimately arbitrary and simply due to the prior definition the indices of $(\frac{\partial\vec{r}}{\partial\vec{q}})$. In all of our work, we have simply chosen $Q = I$.

Here we see that in the case of linear frame constraints E and shape coordinates $\vec{q}$ linear in the body-fixed displacements $\Delta\vec{r}_k$, the resulting $(\frac{\partial\vec{r}}{\partial\vec{q}})$ (and hence S) is constant. This special case will be used to simplify the force field derived in the next section.

II. NUMERICAL CALCULATION OF LAB-FRAME FORCES

The force field equation defined in the Main text allows us to calculate the generalized force field $-\frac{\partial V}{\partial \vec{q}}$ on the $3N-6$ shape coordinates. In order to simulate the complete motion of the molecule with translational and rotational diffusion in the lab frame, however, we need the $3N$ Cartesian forces $-\frac{\partial V}{\partial \vec{r}}$.

We can relate the generalized shape forces to the complete set of forces via the chain rule,

$$\frac{\partial V}{\partial \vec{q}} = \left(\frac{\partial \vec{r}}{\partial \vec{q}}\right) \frac{\partial V}{\partial \vec{r}} \tag{9}$$

but since $\left(\frac{\partial \vec{r}}{\partial \vec{q}}\right)$ is a rectangular $(3N-6)\times(3N)$ matrix, this relationship cannot be inverted to recover $\frac{\partial V}{\partial \vec{r}}$.

We can additionally require that the sum of the internal molecular forces cannot induce a net force or a net torque by Newton's Third Law. These constraints furnish the additional six equations required to complete the inversion:

$$\begin{aligned} \sum_i \vec{F}_i &= 0 \\ \sum_i \vec{r}_i \times \vec{F}_i &= 0. \end{aligned} \tag{10}$$

Combining these constraints into a single matrix equation, we have

$$\begin{bmatrix} \left(\frac{\partial \vec{r}}{\partial \vec{q}}\right) \\ \mathcal{T} \\ \mathcal{N}(\vec{r}) \end{bmatrix} \frac{\partial V}{\partial \vec{r}} = \begin{bmatrix} \frac{\partial V}{\partial \vec{q}} \\ 0 \\ 0 \end{bmatrix} \tag{11}$$

where

$$\begin{aligned} \mathcal{T}_{\alpha,3i+\gamma} &:= \delta_{\alpha\gamma} \\ \mathcal{N}_{\alpha,3i+\gamma}(\vec{r}) &:= \sum_\beta \epsilon_{\alpha\beta\gamma} r_{i\beta}. \end{aligned} \tag{12}$$

Here ϵ is the Levi-Civita symbol. The block matrix on the left-hand side is now square

and of full rank, and can be inverted to solve for $\frac{\partial V}{\partial \vec{r}}$ given the generalized forces $\frac{\partial V}{\partial \vec{q}}$.

---

[1] C. Eckart, Phys. Rev. **47**, 552 (1935).