Vendi Sampling For Molecular Simulations: Diversity As A Force For Faster Convergence And Better Exploration

Amey P. Pasarkar[1, 3], Gianluca M. Bencomo[1],
Simon Olsson[2, *], and Adji Bousso Dieng[1, 3, *]

[1]Department of Computer Science, Princeton University
[2]Department of Computer Science and Engineering, Chalmers University
[3]Vertaix
[*]Corresponding authors: adji@princeton.edu and simonols@chalmers.se

July 3, 2023

Abstract

Molecular dynamics (MD) is the method of choice for understanding the structure, function, and interactions of molecules. However, MD simulations are limited by the strong metastability of many molecules, which traps them in a single conformation basin for an extended amount of time. Enhanced sampling techniques, such as *metadynamics* and *replica exchange*, have been developed to overcome this limitation and accelerate the exploration of complex free energy landscapes. In this paper, we propose *Vendi Sampling*, a replica-based algorithm for increasing the efficiency and efficacy of the exploration of molecular conformation spaces. In Vendi sampling, replicas are simulated in parallel and coupled via a global statistical measure, the Vendi Score, to enhance diversity. Vendi sampling allows for the recovery of unbiased sampling statistics and dramatically improves sampling efficiency. We demonstrate the effectiveness of Vendi sampling in improving molecular dynamics simulations by showing significant improvements in coverage and mixing between metastable states and convergence of free energy estimates for four common benchmarks, including Alanine Dipeptide and Chignolin[1].

Keywords: molecular simulations, enhanced sampling, Markov chain Monte Carlo, diversity, Vendi score, chemical physics.

1 Introduction

The exchange between metastable configurations of proteins and nucleic acids is essential to their biological function Henzler-Wildman and Kern (2007). Molecular dynamics (MD) simulations are widely adopted for simulating the transitions between metastable states because they offer full spatial and temporal resolution of molecular systems. These transitions can range from simple localized changes, such

[1] Code for using Vendi Sampling on your simulation will be made available at https://github.com/vertaix/Vendi-Sampling.

as aromatic ring-flips Shaw et al. (2010), to complex global structural rearrangements, including protein folding Noé et al. (2009); Lindorff-Larsen et al. (2011); Piana et al. (2013) and protein-ligand binding Chakrabarti et al. (2022); Plattner and Noé (2015); Buch et al. (2011); Plattner et al. (2017). Experimental techniques that enable the measurement of signals sensitive to the exchange between these states Olsson and Noé (2016); Noé et al. (2011); Buchete and Hummer (2008); Cavanagh et al. (1996) have greatly advanced the study of these metastable states. However, low spatial resolution and ensemble averaging are still challenges that complicate data analyses Opanasyuk et al. (2022).

MD simulations can predict stationary and dynamic correlations, allowing for direct comparison to experimental data. However, a significant limitation of MD arises due to the strong metastability of many molecular systems, which can trap them in a single conformational basin for an extended amount of time Prinz et al. (2011); Hempel et al. (2022); Olsson and Noé (2019). This limitation undermines quantitative comparisons to experiments because statistical sampling comes at a high computational cost. To address this issue, there has been intense research aimed at developing faster algorithms for sampling highly metastable systems at a lower computational cost Henin et al. (2022).

These enhanced sampling or extended ensemble methods rely on two broad strategies: finding a surrogate of the Boltzmann ensemble that rapidly mixes between metastable states or exchanging the state variable with one or multiple ensembles that mix faster Camilloni et al. (2013); Abrams and Bussi (2013); Grubmüller (1995); Sugita and Okamoto (1999); Swendsen and Wang (1986); Laio and Parrinello (2002); Kříž et al. (2017); Šućur and Spiwok (2016). To ensure that the modified ensembles sufficiently overlap with the true Boltzmann ensemble, statistical reweighing techniques such as importance sampling, the weighted histogram analysis method Ferrenberg and Swendsen (1989), the multistate Bennett acceptance ratio Shirts and Chodera (2008), or transition-based methods Wu et al. (2016); Stelzl and Hummer (2017); Galama et al. (2023) are used. These techniques allow for an effective reweighting of the equilibrium statistics but are disadvantaged by the intrinsic need to identify collective variables, perturb macroscopic thermodynamic variables, or alter Hamiltonians, which often rely on manual trial-and-error of numerous potential candidates Henin et al. (2022). The identification of useful collective variables is a field of its own, with multiple application-specific approaches, including for ligand binding Limongelli et al. (2013) and slow dynamics Ribeiro et al. (2018a); Tiwary and Berne (2016); Sultan and Pande (2017). In particular, collective variable estimation has seen a surge in interest in recent years due to the broad accessibility of deep representation learning methods Ribeiro et al. (2018b); Chen and Ferguson (2018); Bonati et al. (2021).

Machine learning-based approaches, such as Boltzmann generators Noé et al. (2019); Köhler et al. (2020, 2021); Jing et al. (2022), learn surrogate models of the intractable Boltzmann distribution. If these surrogate models allow for exact sample likelihood estimation, highly effective sampling of unbiased equilibrium statistics can be achieved via reweighting or importance sampling. While a promising strategy, current architectural limitations in the deep neural networks used to learn such surrogate models have prevented their broad adoption.

In this paper, we propose a replica-based method called *Vendi Sampling* where multiple copies of a molecular system are simulated in parallel. To enhance the conformational space exploration, the replicas are coupled via a global statistical measure, *the Vendi score* Friedman and Dieng (2022). The Vendi score reflects the instantaneous and internal diversity of the replicas and is a function of the eigenvalues of a Gram matrix computed via a pre-specified kernel. The Gram matrix is constructed by evaluating the kernel between all pairs of simulation replicas. As such, the proposed *Vendi sampling* method does not rely on modulating thermodynamic variables, Hamiltonians, or defining collective variables to enhance sampling. Instead, an extended ensemble is defined, which can drive sampling in infinite dimensional feature spaces, without explicit calculation of collective variables.

Our investigation of Vendi sampling in several challenging systems reveals its ability to rapidly detect free energy minima, particularly in cases where there are large free energy barriers between states. Moreover, Vendi sampling enables the rapid convergence to the unbiased equilibrium statistics, facilitating the computation of observables such as free energy differences. We accomplished these findings using generic kernels, which do not encode slowly relaxing degrees of freedom. However, the current implementation incurs some computational overhead. We anticipate that further research into the development of more efficient kernels and the use of multiple time-stepping schemes Ferrarotti et al. (2014) can further boost the performance of Vendi sampling.

2 Theory

Here we provide some background, describe Vendi sampling, and explain how to recover unbiased observable statistics.

2.1 Boltzmann distribution, observables, and free energies.

At equilibrium, a single molecular system denoted by $\mathbf{x} \in \Omega \subset \mathbb{R}^{3N}$, where N is the number of atoms, has a distribution of conformational states which is equal to the Boltzmann distribution,

$$p(\mathbf{x}) = \mathscr{Z}^{-1} \exp(-\beta U(\mathbf{x})), \tag{1}$$

with inverse temperature β. In MD simulations, we asymptotically generate samples from this distribution.

Given a sample $\{\mathbf{x}_i\}_{i=0}^{M}$ drawn independently and identically from $p(\mathbf{x})$, a state function—or forward-model—corresponding to an experimental observable denoted by $f(\mathbf{x})$, we can compute bulk ensemble averaged observables using Monte Carlo,

$$o_f = \int_\Omega f(\mathbf{x}) p(\mathbf{x})\, \mathrm{d}\mathbf{x} \approx \frac{1}{M} \sum_{i=1}^{M} f(\mathbf{x}_i). \tag{2}$$

Furthermore, we can define the free energy of a state $A \subset \Omega$ as a special case of a stationary observable,

$$F_A = -\log \int_\Omega p(\mathbf{x})\delta(x \in A)\,\mathrm{d}\mathbf{x} = -\log \mathbb{P}(x \in A) \tag{3}$$

where $\mathbb{P}(\cdot)$ denotes the probability of an event. We can similarly express the free energy difference between states $A, B \subset \Omega$ as

$$F_{AB} = -\log \int_\Omega p(\mathbf{x})\delta(x \in A)\,\mathrm{d}\mathbf{x} + \log \int_\Omega p(\mathbf{x})\delta(x \in B)\,\mathrm{d}\mathbf{x} = -\log \frac{\mathbb{P}(x \in A)}{\mathbb{P}(x \in B)}. \tag{4}$$

2.1.1 Replicated systems and expanded ensembles

In replicated systems, N copies of the same system, $\mathscr{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$, are simulated simultaneously. Each copy is referred to as a '*replica*.' In the simplest case, when the replicas are uncoupled, they can be simulated independently as their joint distribution factorizes as the product of the different marginal distributions over each replica,

$$p_{\text{uncoupled}}(\mathscr{X}) = \prod_{i=1}^{N} p_i(\mathbf{x_i}). \tag{5}$$

Here the densities $p_i(\cdot)$ can either be identical or different. One common choice for rendering the densities different is by using different temperatures β_i for each of the replicas, as is done in replica-exchange and parallel tempering methods Swendsen and Wang (1986); Marinari and Parisi (1992); Sugita and Okamoto (1999).

In the case where the replicas are not independent, we call this the *coupled case*, the replicas exchange information via an additional function, $\pi(\cdot)$, which depends on all or parts of the replicas. The corresponding joint distribution can be written as

$$p_{\text{coupled}}(\mathscr{X}) = \pi(\mathscr{X}) \prod_{i=1}^{N} p_i(\mathbf{x_i}). \tag{6}$$

This approach is extensively used to impose experimental constraints on molecular simulations, e.g. by enforcing an averaged experimental observable to match an average computed across multiple replicas. Hummer and Köfinger (2015); Cavalli et al. (2013); Olsson and Cavalli (2015); Lindorff-Larsen et al. (2005); Best and Vendruscolo (2004); Pitera and Chodera (2012); Roux and Weare (2013); Camilloni et al. (2013)

2.2 Vendi Sampling

Here we introduce a coupled replica approach which we refer to as *Vendi Sampling*. In Vendi sampling replicas are coupled using a statistical measure of diversity, the Vendi Score Friedman and Dieng (2022). The goal is to encourage the replicas to cover different regions of conformation space, thereby rapidly discovering metastable states and allowing faster convergence of stationary ensemble properties, such as free energy differences.

The Vendi Score. The Vendi score Friedman and Dieng (2022) is an interpretable diversity metric that quantitatively describes how diverse a collection of items is. It is the effective number of dissimilar elements in the collection being evaluated for diversity. Given a collection of N items, the maximum possible value of the Vendi Score is N, when all items in the collection are uniquely distinct from each other. The measure of similarity between the items in the collection is an input to the Vendi Score, which allows for a flexible specification of any form of similarity. More specifically, the Vendi Score of a collection of samples $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_M}\}$ is defined as the Shannon entropy of the eigenvalues of a similarity matrix,

$$\mathrm{VS}(\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}) = \exp\left(-\sum_{i=1}^{M} \lambda_i \log \lambda_i\right). \tag{7}$$

Here λ_i is the i'th eigenvalue of $\mathbf{K}$, the matrix induced by a user-defined similarity function $k(\cdot, \cdot)$. $\mathbf{K}$ is a Gram matrix such that

$$\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \qquad 0 < i < M, 0 < j < M. \tag{8}$$

Vendi Sampling. In Vendi sampling, the M samples correspond to M replicas of a molecular system, and the resulting Vendi Score is the coupling function. We combine eqs. 6 and 7,

$$p_{\mathrm{vendi-coupled}}(\mathscr{X}) \propto \mathrm{VS}(\mathscr{X}) \prod_{i=1}^{N} p_i(\mathbf{x_i}). \tag{9}$$

Note that the joint distribution of the replicas is no longer normalized as the Vendi Score is not a probability distribution function. However, Vendi sampling doesn't require normalization of the joint distribution of the replicas. More specifically, we use the *Vendi energy function* implied by the density $p_{\mathrm{vendi-coupled}}$,

$$U_{\mathrm{vendi-coupled}}(\mathscr{X}) = \sum_{i=1}^{N} u_i(\mathbf{x_i}) + \lambda_i \log \lambda_i, \tag{10}$$

where $u_i(\cdot) \triangleq \beta_i U_i(\cdot)$ is a unitless energy and λ_i is as described earlier. We use the Vendi energy function to simulate the system via overdamped Langevin dynamics (Tuckerman, 2010),

$$\frac{\mathrm{d}\mathscr{X}(t)}{\mathrm{d}t} = -\nabla U_{\mathrm{vendi-coupled}}(\mathscr{X}(t))/\gamma + \sqrt{2D}\mathrm{d}W \tag{11}$$

where $\mathrm{d}W$ denotes a Weiner process, D is a diffusion constant and γ is a friction coefficient. Each replica evolves over time according to the stochastic differential equation

$$\frac{\mathrm{d}\mathbf{x_i}(t)}{\mathrm{d}t} = -\left[\nabla u_i(\mathbf{x_i}(t)) + \nu f_{\mathrm{vendi}}(\mathbf{x_i})\right]/\gamma + \sqrt{2D}\mathrm{d}W \tag{12}$$

where $f_{\mathrm{vendi}}(\mathbf{x_i}) = -\frac{\partial}{\partial \mathbf{x_i}} \log \mathrm{VS}(\mathscr{X})$ is the '*Vendi force*' for $\mathbf{x_i}$—a repulsive force applied to $\mathbf{x_i}$ to drive it away from the other replicas—and ν is an associated learning rate. In our experiments, we explore two types of Langevin sampling to simulate the combined system (eq. 9): we use overdamped Langevin dynamics (Brownian dynamics) for the model systems (the Prinz potential and the double well) and Langevin dynamics for the molecular systems.

The variable ν is a hyperparameter that can be subject to various annealing schemes. In systems that use linear annealing, we perform a fast hyperparameter search to find an optimal schedule. In particular, we run the Vendi sampler for a small number of steps (2000 steps) and then pick the rate that yields a sampler with maximal value for the heuristic $\mathrm{VS}(\mathscr{X}) - 2\langle U(\mathscr{X})\rangle$, where $\mathscr{X}$ describes samples from the learned distribution, $\langle U(\mathscr{X})\rangle$ represents the average energy of those samples, and VS is their corresponding Vendi Score.

Recovering unbiased samples. After simulating an ensemble of coupled replicas for T time steps $\{\mathscr{X}(0), \ldots, \mathscr{X}(T)\}$ we can reweigh these trajectories into an unbiased ensemble by resampling using the time-dependent weights

$$w_t = \mathrm{VS}(\mathscr{X}(t))^{-1}, \tag{13}$$

which will yield asymptotically unbiased samples from the joint distribution eq. 5.

In the case of annealing, we use the above weights only while the Vendi force is applied, and use weight 1 for all subsequent samples.

2.3 Kernels

The *Vendi force* applied to each replica at every simulation step is contingent on the kernel matrix **K**, as it encodes the diversity across replicas.

Kernels are inner-products in a feature space, $\mathscr{V}$,

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle_{\mathscr{V}} = \phi_i^\top \phi_j \tag{14}$$

where the latter equality holds if $\mathscr{V}$ is a vector space. In general, the dimension of $\mathscr{V}$ is infinite. This gives rise to the the celebrated 'Kernel trick' in machine learning Schölkopf et al. (2002), where data is embedded in to a high-dimensional space to allow for linear separation. In the context of molecular simulations, the kernel can be interpreted as a way of comparing high-dimensional collective variables simultaneously, side-stepping the need to engineer the collective variables. Below we detail the kernel used in the molecular applications we studied.

Invariant Kernel for Molecular systems. As we are considering MD simulations at equilibrium without exchange of matter or energy with the environment, and without interactions with an external field, sampling the conformational space of the system is invariant to global rotations and translations of the molecular frame. Consequently, we require a kernel $k(\cdot, \cdot)$ that is invariant to those symmetries. Since we would like to quantify some metric for distance within the space of possible molecular conformations, removing these symmetries in the space of all possible molecular configurations is a necessary requirement for a diversity metric that is accurate and efficient.

We use the Gaussian kernel, also called a Radial Basis Function (RBF) kernel,

$$k(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right) \tag{15}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are input vectors, σ is a hyperparameter that determines the width of the kernel, and $\|\cdot\|$ denotes the Euclidean norm. Given some replicas $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, we can ensure translational invariance by forming a new set $\tilde{\mathbf{X}}$ that is mean-centered,

$$\tilde{\mathbf{X}} = \mathbf{X} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i. \tag{16}$$

There are several ways to find an invariant basis function $\boldsymbol{\Phi}(\cdot)$, which we will need to use with our replicas before applying the RBF kernel. We use a procedure outlined previously to find rotation matrices $\boldsymbol{U}_i$ which align all replicas onto a common frameJaini et al. (2021). These transformed coordinates are in turn used to evaluate the kernel.

3 Results

We compare Vendi sampling against an uncoupled replica method with overdamped Langevin dynamics to showcase the added benefit of enforcing a repulsive force through the Vendi Score. We refer to this baseline as *Replica sampling* in the rest of the paper. We report free energy profiles on two model systems, the Prinz potential and the double well, and two molecular systems, Alanine dipeptide in vacuum and Chignolin in implicit solvent. In all cases we found Vendi sampling converges faster and explores energy surfaces better than Replica sampling.

3.1 Model systems

We first test Vendi sampling on two low-dimensional model systems that have exact expressions for their free energy functions. For these systems we use the simple kernel function:

$$k(x,x') = \frac{|x-x'|}{|x|+|x'|}. \tag{17}$$

3.1.1 Prinz Potential

The '*Prinz Potential*'Prinz et al. (2011) is a four-well 1D potential that features an energy barrier at $x = 0$. We test Vendi sampling's ability to both discover these wells and provide unbiased equilibrium sampling statistics. We initialized the replicas uniformly from one side of the central energy barrier ($\sim U[0,1]$). We performed this experiment 10 times for each choice of replica size, comparing against Replica sampling with an integration step of $\eta = 10^{-4}$ for both samplers. Hyperparameters were selected via a grid search over learning rate $\nu = \{1,10,40,100\}$ and linear annealing rates of $\alpha = \{0.01, 0.025, 0.1\}$, as detailed in section 2.2. On all replica sizes tested here, the grid search yielded $\nu = 100$, $\alpha = 0.1$.

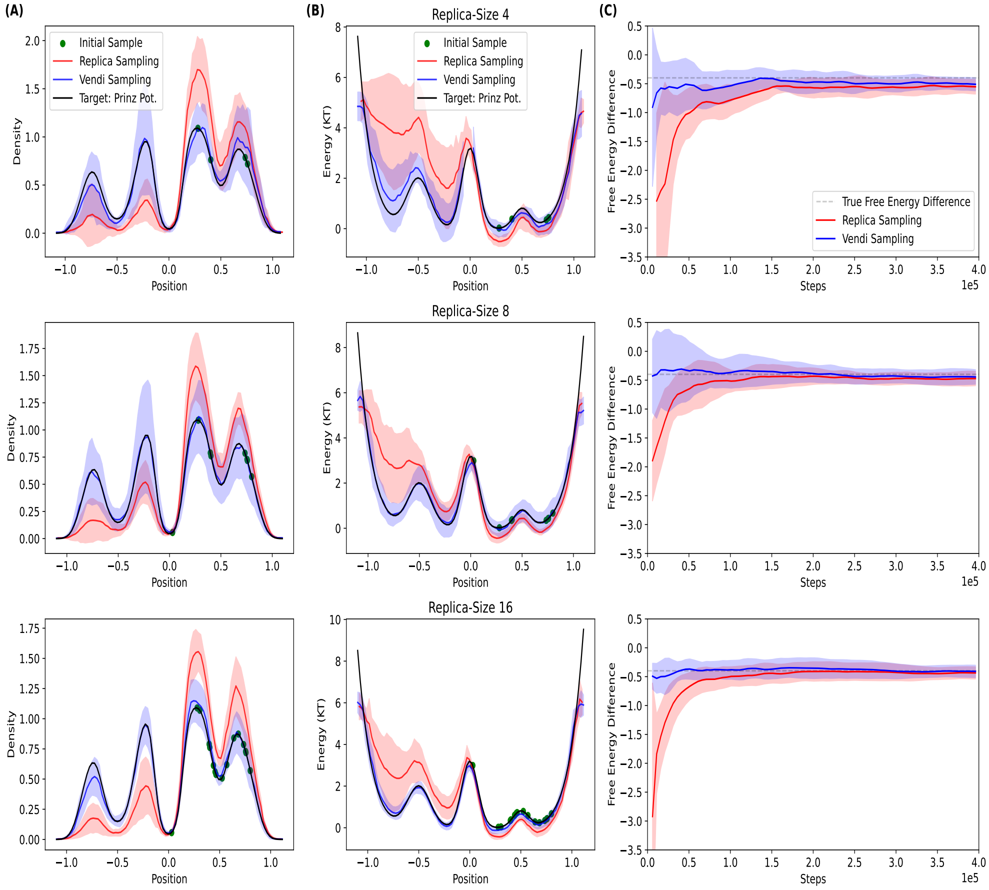

Figure 1: Vendi sampling rapidly identifies states and quickly recovers ground-truth equilibrium statistics. (A) Probability density functions for Vendi sampling and Replica sampling compared against the ground truth Prinz Potential density after 10^4 steps of each sampler for various replica sizes. (B) Energy functions for Vendi sampling and Replica sampling after 10^4 steps for various replica sizes. (C) Free energy difference across boundary over time for both samplers. In all cases, the shaded region reflects uncertainty as estimated using the standard deviation across $n = 10$ simulations after 10^3 burn-in steps.

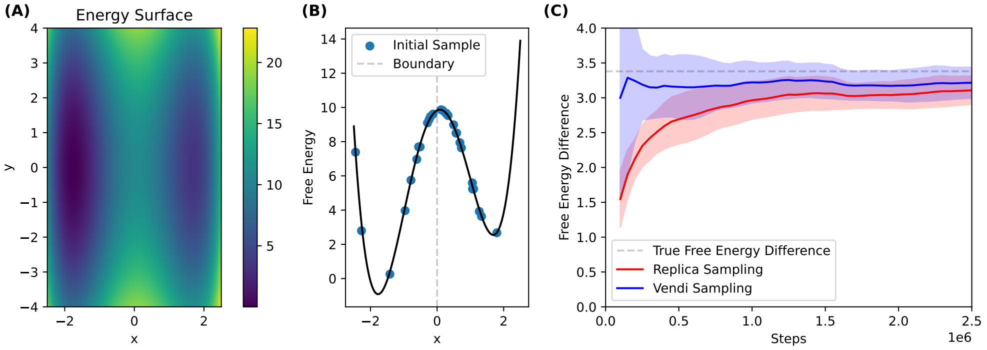

Figure 2: The **Vendi sampler rapidly converges in the double well potential**. (A) The underlying 2D free energy surface. (B) The marginal free energy along the x-axis. The initial conditions are shown as dots on the curve and the free energy boundary for free energy calculations is shown with a dashed line. (C) Comparison of convergence over 2.5 million steps for Vendi sampling and Replica sampling. The shaded area represents uncertainty as measured by the standard deviation over $n = 10$ experiments and after $50,000$ steps.

We observe that Vendi sampling rapidly identifies modes on the opposite side of the energy barrier (Fig. 1A & Fig. 1B). After $10,000$ time steps, Replica sampling poorly characterizes the left-most mode, while Vendi sampling is already able to approximate it relatively well.

We compare the convergence of Vendi sampling and Replica sampling, by observing the free energy difference (eq. 4) between the segments $A : \{x \in [-1,0]\}$ and $B : \{x \in [0,1]\}$ for both methods, as a function of simulation step. Vendi sampling consistently outperforms the Replica sampling baseline (Fig. 1C) by reaching the true free energy difference within 0.1 kT in fewer steps. Specifically, for 4 replicas Vendi sampling converged at 150’000 steps, whereas for 8 and 16 replicas Vendi Sampling reached similar values by the $50,000$th step. In contrast, Replica sampling required at least twice as many steps to achieve similar convergence.

Vendi sampling discovers all high probability (low energy) within $10,000$ time steps while Replica sampling requires an order of magnitude more steps to discover all states.

3.1.2 Double Well potential

To gauge the performance of Vendi sampling on systems with a higher free energy barrier ($> 10kT$), akin to those observed in protein folding-unfolding or in chemical reactions, we considered a previously studied two-dimensional double-well benchmark Noé et al. (2019).

We limited our experiments to 32 replicas for both Vendi sampling and Replica sampling, with initial starting positions drawn randomly in the interval $x \in [-2.5, 2.5]$ and $y \in [-2.5, 2.5]$. The free energy difference was calculate as before with Equation 4, using $A = \{x \in [-2.5, 0], y \in [-4, 4]\}$ and $B = \{x \in [0, 2.5], y \in [-4, 4]\}$. We used $\eta = 10^{-2}$ for both samplers and determined α, ν using procedures identical to those of the previous section. The grid search yields $\alpha = 2 \cdot 10^{-5}$, $\nu = 100$ as the optimal hyperparameters.

Vendi sampling converged within 0.2 kT of the true free energy difference by the $1,000,000$th step, whereas Replica sampling did not reach this over the course of the entire simulation (2C).

3.2 Molecular systems

Here we further investigate two molecular systems: capped alanine (alanine dipeptide) and the miniprotein chignolin, in vacuum and implicit solvent respectively. These systems are established benchmark systems to evaluate sampling methods in molecular dynamics (Chen et al., 2021; Invernizzi and Parrinello, 2020). We perform similar benchmarks as above, however, we use a kernel which is invariant to rotations and translation as described in section 2.3 using kernel bandwidth parameter $\sigma = \sqrt{\frac{1}{2}}$. This kernel ensures only internal degrees of freedom are encouraged to diversify, and avoids that the center of mass and orientation of the molecular system influences the Vendi score.

3.2.1 Alanine dipeptide in vacuum

Alanine dipeptide (Ala2) or capped alanine is a small, yet meaningful benchmark system to test enhanced sampling methods and the impact of solvation (Invernizzi and Parrinello, 2020; Brady and Karplus, 1985; Wu and Wang, 1998).

In vacuum, Ala2 exhibits a three-basin free-energy landscape in the Ramachandran plot of the back-bone dihedral angles ϕ (C-N-Cα-C) and ψ (N-Cα-C-N). Two of the basins occupy the β-strand like structures and are disconnected by only a small free energy barrier. The third, minor basin occupies the region of the Ramachandran plot associated with left-handed helices, and is disconnected from the β-strand states by a large free-free energy barrier $> 10 k_B T$ in the Amber96-SB forcefield at 300 K (Lindorff-Larsen et al., 2010) All simulations were carried out using the Langevin integrator implemented in OpenMM (v. 8.0) (Eastman et al., 2017) with a time step of 2.0 fs and a collision frequency of 1.0ps^{-1}. A free energy baseline was established by running 10 simulations of 100 ns each, as depicted in Supp. Figure S1. To analyze the conformational distribution, we recorded the ψ and ϕ backbone dihedral angles of each sample. The free energy difference was calculated via Equation 4 with $A = \{\phi \in [-\pi, 0], \psi \in [-\pi, \pi]\}$ and $B = \{\phi \in [0, \pi], \psi \in [-\pi, \pi]\}$.

For Vendi sampling we computed the Vendi Force that was applied to the Langevin dynamics at every timestep for the first 15 ps, with a learning rate of $\beta = 100$. After 15 ps, the Vendi Sampler reverted to using standard Langevin dynamics.

We experimented with 32-replica systems for both Vendi sampling and Replica sampling, initializing all replicas in the β-stranded state. We find that the Vendi sampler effectively used the Vendi force to break the hydrogen bonds between the amide nitrogen and carbonyl oxygen atoms stabilizing the β-stranded states, resulting in several replicas transitioning to the left-handed state within the first 1.0 ns (Fig. 3A & B). In contrast, the Replica sampler relies on thermal fluctuations to cross this free energy barrier. Replica sampling, further requires an approximate three-fold

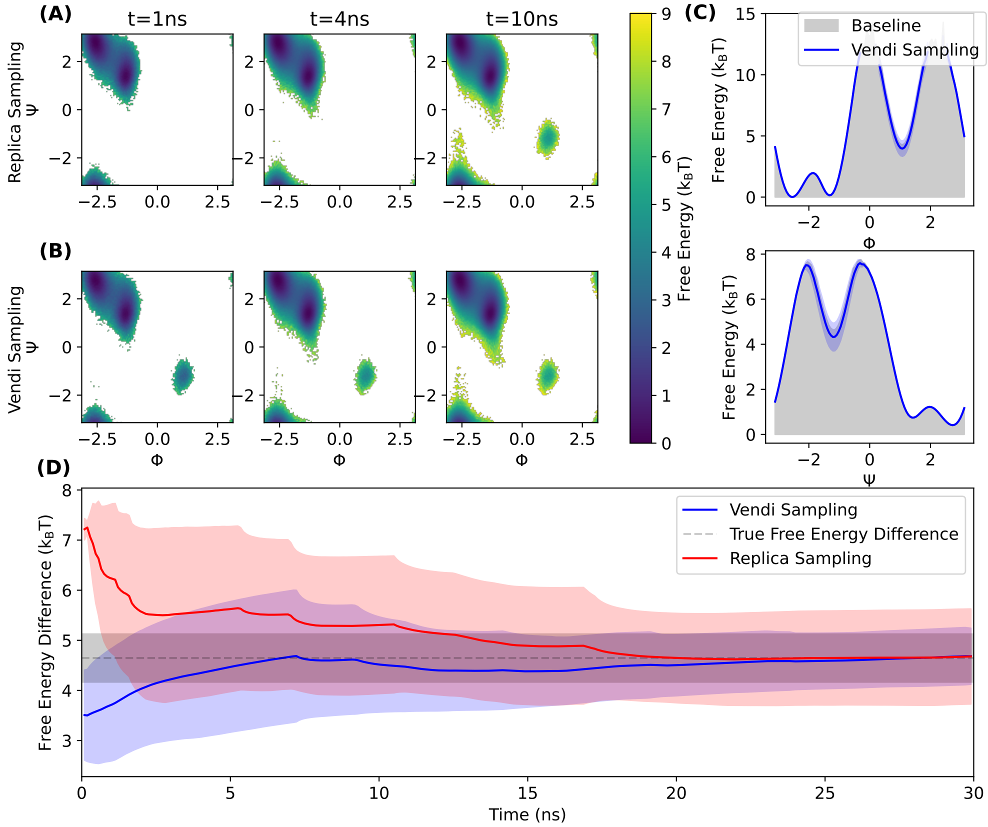

Figure 3: Vendi Sampler rapidly identifies modes, converges fast, and recovers accurate free energy difference in Ala2 in vacuum. (A) Free energy maps of dihedral angles of replica sampler at various time points. (B) Density maps of dihedral angles of the Vendi sampler at various time points show noticeable improvement in the time taken to discover all modes. (C) Comparison of the average learned 1D Free Energy functions along each angle compared against one baseline run. Bars show standard deviation across trials. (D) Free Energy Difference over time for each sampler across trials. Throughout, the shaded region for samplers represents uncertainty estimated as the standard deviation with $n = 10$ after $7,500$ burn-in steps. The black bar also represents uncertainty estimated as the standard deviation in the final free energy difference across 10 long runs.

larger simulation effort to equilibrate the average relative state populations within the accuracy of the baseline estimate. This demonstrates Vendi sampling's enhanced capacity for climbing the large energy barrier and discovering the left-handed minima quickly. The increased mode coverage during the early stages of the simulation facilitated faster convergence to the true free energy difference, as illustrated in Figure 3D. It also important that our sampler encourages mixing across modes in the MD trajectories. We observe that while the Vendi force is applied over the first 15 ps of the simulation, the Vendi Sampler provides much better mixing in the MD simulation (Fig. 4A,B&C). The Replica sampler would mostly transition to the left-handed state later in the simulation (Fig. 4D,E,&F), by which time the Vendi sampler had already converged.

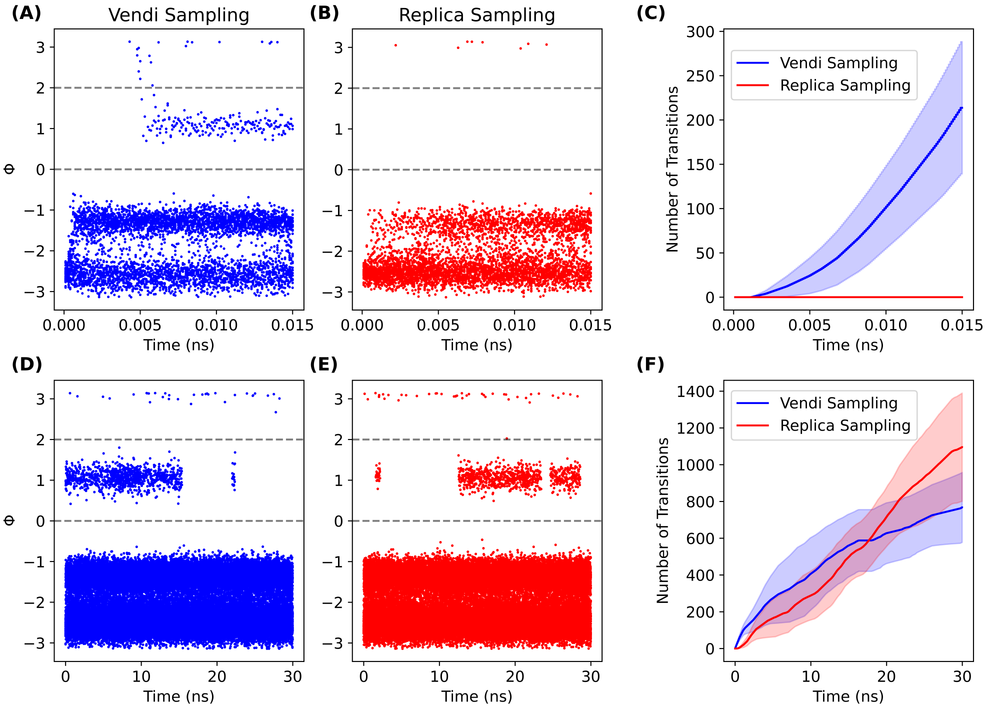

Figure 4: Vendi Sampling provides fast mixing in Ala2 system by rapidly crossing the energy barrier, performing multiple transitions. (A) & (B): Scatter plot of dihedral angle ϕ over time of samples from each method (first 15 ps are shown. The dashed line represents the region where specific Ala2 conformation is sampled. (C) Number of transitions in and out of the boxed region. (D) & (E): Scatter plot of the dihedral angle ϕ over time of samples from each method for an entire 30 ns-length simulation. (F) Number of transitions in and out of the boxed region. In all plots, the shaded region depicts uncertainty represented as an 80% confidence interval.

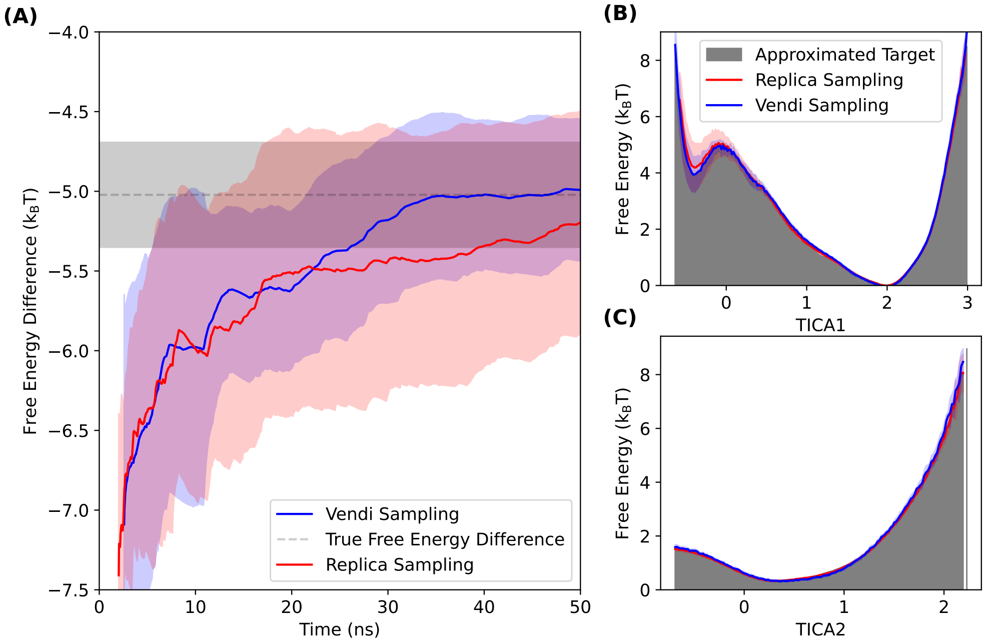

Figure 5: Vendi Sampling rapidly converges in Chignolin system. (A) Free energy Difference across TICA1=0. over time for the MD simulation of each sampler. Shaded region represents standard deviation across 10 trials after 80,000 burn-in steps. (B) & (C) Marginal free energy of each sampler along TICA axis compared against the against the average of 10 long unbiased runs.

3.3 Protein folding Chignolin in implicit solvent (CLN025)

Chignolin is a small artificial mini-protein with fast-folding kinetics, is a widely used model system for computational and experimental study of protein folding (Sobieraj and Setny, 2022; Chen et al., 2021; Okumura, 2012) Its folded state conformation is characterized by a strong hydrogen bonding network between the backbone amide and carbonyl groups forming a stable β-hairpin conformation.

We used the CHARMM22 protein force field from MacKerell Jr et al. (1998) with the OBC2 (Onufriev-Bashford-Case) implicit solvent model Onufriev et al. (2004) implemented by OpenMM to simulate Chignolin at a temperature of 350 K. Vendi sampling and Replica sampling both used the standard Langevin integrator from OpenMM with constraints on bonds to hydrogens and hydrogen mass repartitioning, time step of 4.0 fs and a collision frequency of 0.1 ps^{-1}. The Vendi sampler additionally incorporated a Vendi Force that was applied to the Langevin dynamics at every timestep for the first 320 ps, with a learning rate of $\nu = 250$. After 320 ps, the Vendi sampler reverted to using standard Langevin dynamics.

As for Ala2, we used a 32-replica setup to compare Vendi sampling and Replica sampling, initializing all replicas in the misfolded state. A free energy baseline was established by running 10 simulations of 100 ns each. Using previously published simulation data on Chignolin in explicit water Lindorff-Larsen et al. (2011) to identify collective variables via TICA (time-lagged independent component analysis) Pérez-Hernández et al. (2013); Schwantes and Pande (2015). We used all pair-wise

Cα distances as features and a lag-time of 5 ns. The first two TICs separate folded and unfolded states (S2), and are used to compute free energies. The two primary TICs are shown in Supp. Figure S2A. The free energy difference was calculated via Equation 4 with $A = \{\text{TICA1} < 0\}$ and $B = \{\text{TICA1} > 0\}$.

We observe that the Vendi sampling average across trials converges within one standard deviation of the estimated true free energy difference in 15 ns less than the replica sampler (Fig 5). We see from the marginal free energy show in Fig 5B and the reference free energy surface in Fig Supp. S2A that the Vendi sampler is fitting the Chignolin folded state probability better.

Vendi sampling does not seem to provide a noticeable increase in performance in the early steps of the simulation. Perhaps this is due to the broad entropic basin of Chignolin's unfolded state Bicout and Szabo (2000). Such a basin would allow the Vendi Force to achieve structural diversity without necessarily passing the free energy barrier of interest.

Supp. Table S1 shows that the Vendi Force calculation is quite expensive, but the force is only applied over a small fraction (0.6%) of steps. It's application over even a small number of steps is enough to achieve a noticeable improvement in performance.

4 Discussion and Conclusion

We introduced Vendi sampling, a replica-based sampling method for molecular simulations. Vendi sampling overcomes large free energy barriers and enhances sampling. While there is a wide variety of enhanced sampling techniques available in the literature Henin et al. (2022), Vendi sampling, to our best knowledge, takes a unique approach to the problem. It is driven by a diversity metric, the Vendi score, which is computed according to a kernel matrix that performs pairwise comparisons of the states of all replicas. Consequently, the system of coupled replicas defined by Vendi sampling forms an extended ensemble that is neither driven by the choice or modulation of a macroscopic thermodynamic parameter (e.g. temperature or Hamiltonian) nor by the definition of a collective variable. Instead, the extended ensemble is defined by a kernel function. The kernel function computes inner products in feature spaces, e.g. high-dimensional collective variables. However, since we do not need to explicitly compute the features in the kernel formalism, the underlying features can, in principle, be infinite-dimensional. We can thereby avoid costly collective variable identification, without suffering from 'the curse of dimensionality.'

It is important to note that the use of kernels is not new in enhanced sampling, however, previous use has focused on estimating an adaptive biasing potential Invernizzi and Parrinello (2020); Mones et al. (2016); Debnath and Parrinello (2020); Maragakis et al. (2009). While these approaches can yield low-variance free energy estimates, they intrinsically have to balance exploration and accuracy, Invernizzi and Parrinello (2022), e.g. by slowly tampering off the estimation of the adaptive biasing potential Barducci et al. (2008). On the contrary, Vendi sampling yields high-variance sample estimates, yet avoids problems associated with adaptive biasing potential estimation.

Through our analyses of tractable model systems and benchmark molecular systems, we find that Vendi sampling indeed excels in cases that are defined by high free energy barriers in two ways: it rapidly covers the local minima of the free energy landscape and enables fast mixing across free energy barriers to allow for fast convergence of free energy estimates.

Future work will explore whether Vendi sampling can benefit from more advanced kernel functions, e.g. in the context of sampling interfacial water molecules, which may be important for molecular processes such as ligand binding. Exploring alternative annealing strategies of the Vendi force during sampling may lead to further efficiency gains.

Acknowledgements

Adji Bousso Dieng is supported by the National Science Foundation, Office of Advanced Cyberinfrastructure (OAC): #2118201, and Schmidt Futures' AI2050 Early Career Fellowship. Simon Olsson is supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Simon Olsson thanks NAISS for providing part of the computational infrastructure used to generate preliminary results for this work (Project: SNIC 2022/22-616). Amey Pasarkar is supported by an NSF-GRFP fellowship. We thank Michele Invernizzi (FU Berlin) for insightful discussions.

References

Abrams, C. and Bussi, G. (2013). Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199.

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100(2).

Best, R. B. and Vendruscolo, M. (2004). Determination of protein structures consistent with NMR order parameters. *Journal of the American Chemical Society*, 126(26):8090–8091.

Bicout, D. J. and Szabo, A. (2000). Entropic barriers, transition states, funnels, and exponential protein folding kinetics: A simple model. *Protein Science*, 9(3):452–465.

Bonati, L., Piccini, G., and Parrinello, M. (2021). Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences*, 118(44):e2113533118.

Brady, J. and Karplus, M. (1985). Configuration entropy of the alanine dipeptide in vacuum and in solution: a molecular dynamics study. *Journal of the American Chemical Society*, 107(21):6103–6105.

Buch, I., Giorgino, T., and De Fabritiis, G. (2011). Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 108(25):10184–10189.

Buchete, N.-V. and Hummer, G. (2008). Coarse master equations for peptide folding dynamics. *The Journal of Physical Chemistry B*, 112(19):6057–6069.

Camilloni, C., Cavalli, A., and Vendruscolo, M. (2013). Replica-averaged metadynamics. *Journal of Chemical Theory and Computation*, 9(12):5610–5617.

Cavalli, A., Camilloni, C., and Vendruscolo, M. (2013). Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *The Journal of Chemical Physics*, 138(9):094112.

Cavanagh, J., Fairbrother, W. J., Palmer III, A. G., and Skelton, N. J. (1996). *Protein NMR spectroscopy: principles and practice*. Academic press.

Chakrabarti, K. S., Olsson, S., Pratihar, S., Giller, K., Overkamp, K., Lee, K. O., Gapsys, V., Ryu, K.-S., de Groot, B. L., Noé, F., et al. (2022). A litmus test for classifying recognition mechanisms of transiently binding proteins. *Nature Communications*, 13(1):3792.

Chen, W. and Ferguson, A. L. (2018). Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of Computational Chemistry*, 39(25):2079–2102.

Chen, Y., Krämer, A., Charron, N. E., Husic, B. E., Clementi, C., and Noé, F. (2021).

Machine learning implicit solvation for molecular dynamics. *The Journal of Chemical Physics*, 155(8):084101.

Debnath, J. and Parrinello, M. (2020). Gaussian mixture-based enhanced sampling for statics and dynamics. *The Journal of Physical Chemistry Letters*, 11(13):5076–5080.

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., et al. (2017). Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659.

Ferrarotti, M. J., Bottaro, S., Pérez-Villa, A., and Bussi, G. (2014). Accurate multiple time step in biased molecular simulations. *Journal of Chemical Theory and Computation*, 11(1):139–146.

Ferrenberg, A. M. and Swendsen, R. H. (1989). Optimized monte carlo data analysis. *Computers in Physics*, 3(5):101–104.

Friedman, D. and Dieng, A. B. (2022). The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*.

Galama, M. M., Wu, H., Krämer, A., Sadeghi, M., and Noé, F. (2023). Stochastic approximation to MBAR and TRAM: Batchwise free energy estimation. *Journal of Chemical Theory and Computation*, 19(3):758–766.

Grubmüller, H. (1995). Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, 52(3):2893–2906.

Hempel, T., Olsson, S., and Noé, F. (2022). Markov field models: Scaling molecular kinetics approaches to large molecular machines. *Current Opinion in Structural Biology*, 77:102458.

Henin, J., Lelievre, T., Shirts, M. R., Valsson, O., and Delemotte, L. (2022). Enhanced sampling methods for molecular dynamics simulations [article v1.0]. *Living Journal of Computational Molecular Science*, 4(1).

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450(7172):964–972.

Hummer, G. and Köfinger, J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics*, 143(24):243150.

Invernizzi, M. and Parrinello, M. (2020). Rethinking metadynamics: From bias potentials to probability distributions. *The journal of physical chemistry letters*, 11(7):2731–2736.

Invernizzi, M. and Parrinello, M. (2022). Exploration vs convergence speed in adaptive-bias enhanced sampling. *Journal of Chemical Theory and Computation*, 18(6):3988–3996.

Jaini, P., Holdijk, L., and Welling, M. (2021). Learning equivariant energy based models with equivariant stein variational gradient descent. *Advances in Neural Information Processing Systems*, 34:16727–16737.

Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. (2022). Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*.

Köhler, J., Klein, L., and Noé, F. (2020). Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pages 5361–5370. PMLR.

Köhler, J., Krämer, A., and Noé, F. (2021). Smooth normalizing flows. *Advances in Neural Information Processing Systems*, 34:2796–2809.

Kříž, P., Šućur, Z., and Spiwok, V. (2017). Free-energy surface prediction by flying gaussian method: Multisystem representation. *The Journal of Physical Chemistry B*, 121(46):10479–10483.

Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566.

Limongelli, V., Bonomi, M., and Parrinello, M. (2013). Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences*, 110(16):6358–6363.

Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132.

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How fast-folding proteins fold. *Science*, 334(6055):517–520.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010). Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958.

MacKerell Jr, A. D., Bashford, D., Bellott, M., Dunbrack Jr, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616.

Maragakis, P., van der Vaart, A., and Karplus, M. (2009). Gaussian-mixture umbrella sampling. *The Journal of Physical Chemistry B*, 113(14):4664–4673.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new monte carlo scheme. *Europhysics Letters (EPL)*, 19(6):451–458.

Mones, L., Bernstein, N., and Csányi, G. (2016). Exploration, sampling, and reconstruction of free energy surfaces with gaussian process regression. *Journal of Chemical Theory and Computation*, 12(10):5100–5110.

Noé, F., Doose, S., Daidone, I., Löllmann, M., Sauer, M., Chodera, J. D., and Smith, J. C. (2011). Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proceedings of the National Academy of Sciences*, 108(12):4822–4827.

Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147.

Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L., and Weikl, T. R. (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences*, 106(45):19011–19016.

Okumura, H. (2012). Temperature and pressure denaturation of chignolin: Folding and unfolding simulation by multibaric-multithermal molecular dynamics method. *Proteins: Structure, Function, and Bioinformatics*, 80(10):2397–2416.

Olsson, S. and Cavalli, A. (2015). Quantification of entropy-loss in replica-averaged modeling. *Journal of Chemical Theory and Computation*, 11(9):3973–3977.

Olsson, S. and Noé, F. (2016). Mechanistic models of chemical exchange induced relaxation in protein NMR. *Journal of the American Chemical Society*, 139(1):200–210.

Olsson, S. and Noé, F. (2019). Dynamic graphical models of molecular kinetics. *Proceedings of the National Academy of Sciences*, 116(30):15001–15006.

Onufriev, A., Bashford, D., and Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics*, 55(2):383–394.

Opanasyuk, O., Barth, A., Peulen, T.-O., Felekyan, S., Kalinin, S., Sanabria, H., and Seidel, C. A. M. (2022). Unraveling multi-state molecular dynamics in single-molecule FRET experiments. II. quantitative analysis of multi-state kinetic networks. *The Journal of Chemical Physics*, 157(3):031501.

Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G., and Noé, F. (2013). Identification of slow molecular order parameters for Markov model construction. *Journal of Chemical Physics*, 139(1).

Piana, S., Lindorff-Larsen, K., and Shaw, D. E. (2013). Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences*, 110(15):5915–5920.

Pitera, J. W. and Chodera, J. D. (2012). On the use of experimental observations to bias simulated ensembles. *Journal of Chemical Theory and Computation*, 8(10):3445–3451.

Plattner, N., Doerr, S., De Fabritiis, G., and Noé, F. (2017). Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling. *Nature chemistry*, 9(10):1005–1011.

Plattner, N. and Noé, F. (2015). Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and markov models. *Nature communications*, 6(1):7653.

Prinz, J.-H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schütte, C., and Noé, F. (2011). Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17):174105.

Ribeiro, J. M. L., Bravo, P., Wang, Y., and Tiwary, P. (2018a). Reweighted autoencoded variational bayes for enhanced sampling (rave). *The Journal of chemical physics*, 149(7):072301.

Ribeiro, J. M. L., Bravo, P., Wang, Y., and Tiwary, P. (2018b). Reweighted autoencoded variational bayes for enhanced sampling (RAVE). *The Journal of Chemical Physics*, 149(7):072301.

Roux, B. and Weare, J. (2013). On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *The Journal of Chemical Physics*, 138(8):084107.

Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schwantes, C. R. and Pande, V. S. (2015). Modeling molecular kinetics with tica and the kernel trick. *Journal of chemical theory and computation*, 11(2):600–608.

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346.

Shirts, M. R. and Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12):124105.

Sobieraj, M. and Setny, P. (2022). Granger causality analysis of chignolin folding. *Journal of Chemical Theory and Computation*, 18(3):1936–1944.

Stelzl, L. S. and Hummer, G. (2017). Kinetics from replica exchange molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 13(8):3927–3935.

Šućur, Z. and Spiwok, V. (2016). Sampling enhancement and free energy prediction by the flying gaussian method. *Journal of Chemical Theory and Computation*, 12(9):4644–4650.

Sugita, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151.

Sultan, M. M. and Pande, V. S. (2017). tICA-metadynamics: Accelerating metadynamics by using kinetically selected collective variables. *Journal of Chemical Theory and Computation*, 13(6):2440–2447.

Swendsen, R. H. and Wang, J.-S. (1986). Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609.

Tiwary, P. and Berne, B. (2016). Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences*, 113(11):2839–2844.

Tuckerman, M. (2010). *Statistical mechanics: theory and molecular simulation*. Oxford University Press, USA.

Wu, H., Paul, F., Wehmeyer, C., and Noé, F. (2016). Multiensemble markov models of molecular thermodynamics and kinetics. *Proceedings of the National Academy of Sciences*, 113(23):E3221–E3230.

Wu, X. and Wang, S. (1998). Self-guided molecular dynamics simulation for efficient conformational search. *The Journal of Physical Chemistry B*, 102(37):7238–7250.

5 Supplemental Information

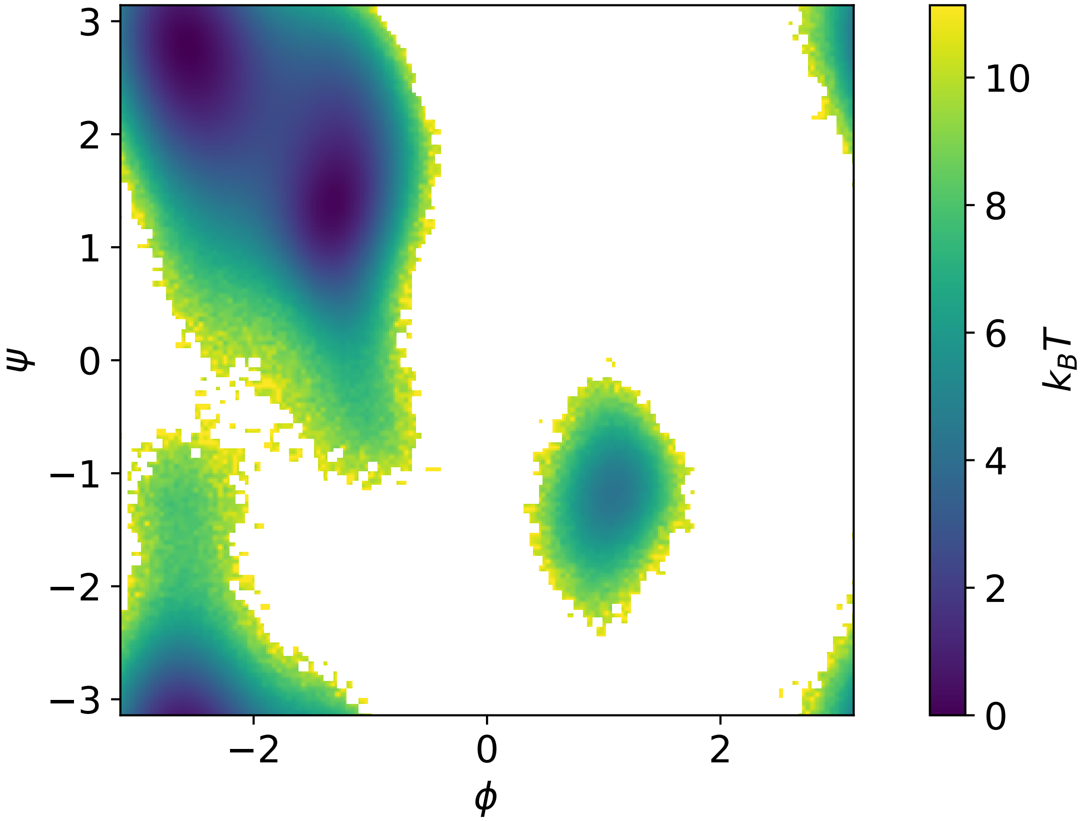

Figure S1: Free Energy Surface of unbiased long-run for Alanine Dipeptide in vacuum

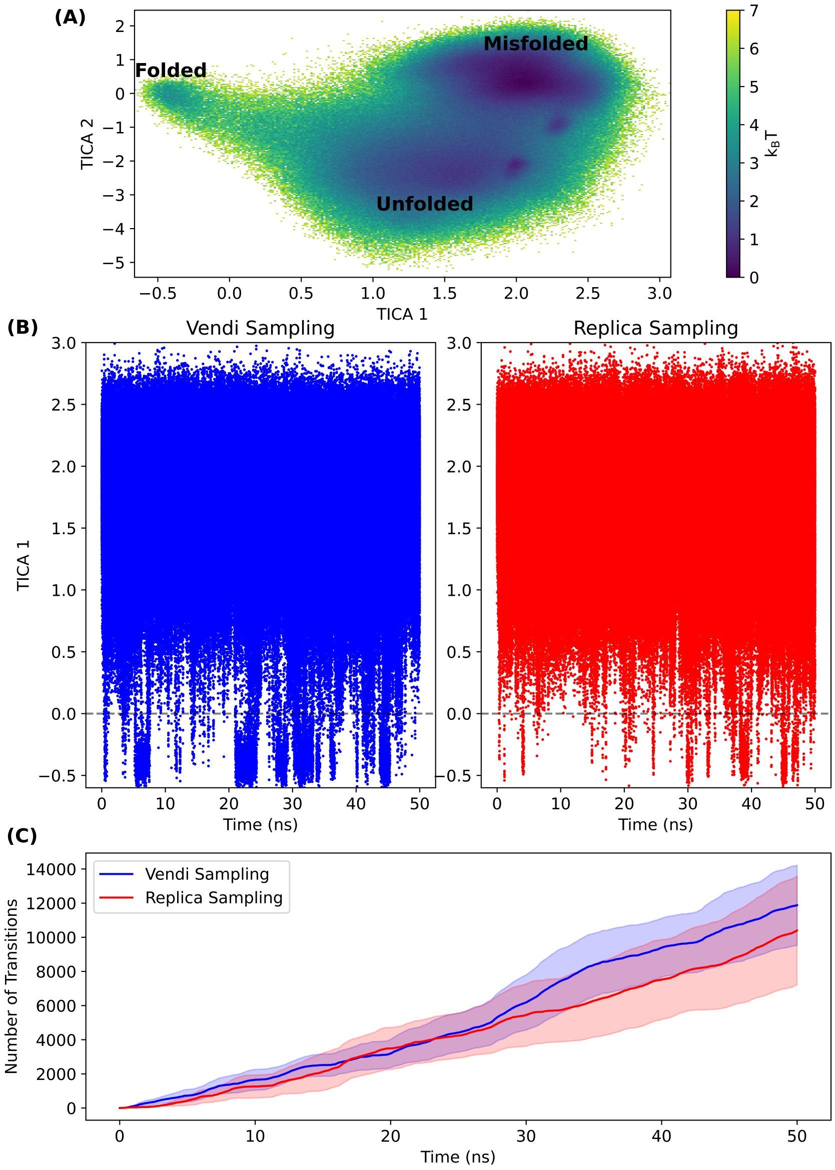

Figure S2: Vendi Sampling improves mixing in Chignolin system. (A) Reference free energy surface along first two TICA axes from a long ubiased run. Chignolin folding states are annotated. (B) Scatter plot of TICA 1 feature over time for Vendi & Replica sampling respectively. Dashed line separates region of misfolded Chignolin, showing state of samples over time for each sampler. (C) Number of state transitions in and out of misfolded Chignolin over time for each sampler. Shaded region depicts uncertainty represented as 80% confidence interval.

Sampling Method	% Time Computing				**MD (ns per day)**
	Langevin Step	Kernel	Vendi Score	Vendi Force	
Vendi Sampling	2	42	<1	55	6
Replica Sampling	100	NA	NA	NA	260

Table S1: Comparison of computational speed for Vendi sampling and Replica sampling. Breakdown of computation overhead to calculate the Vendi force when using 32 replicas on Chignolin in implicit solvent.