Integrating Path Sampling with Enhanced Sampling for Rare-event Kinetics

Dhiman Ray*

Department of Chemistry and Biochemistry, University of Oregon, Eugene, Oregon 97403, USA

E-mail: dray@uoregon.edu

Abstract

Studying the kinetics of long-timescale rare events is a fundamental challenge in molecular simulation. To address this problem, we propose an integration of two different rare-event sampling philosophies: biased enhanced sampling and unbiased path sampling. Enhanced sampling methods e.g. metadynamics can facilitate enthalpic barrier crossing by applying an external bias potential. On the contrary, path sampling methods like weighted ensemble (WE) lack explicit mechanisms to overcome energetic barriers. However, they can accelerate the exploration of rugged free energy surfaces through trajectory resampling. We show that a judicious combination of the weighted ensemble with a metadynamics-like algorithm, can synergize the strengths and mitigate the deficiencies of path sampling and enhanced sampling approaches. The resulting integrated sampling (IS) algorithm improves the computational efficiency of calculating the kinetics of peptide conformational transitions, protein unfolding, and the dissociation of a ligand-receptor complex. Furthermore, the IS approach can direct sampling along the minimum free energy pathway even when the collective variable used for biasing is suboptimal. These advantages make the integrated sampling algorithm suitable for studying the kinetics of complex molecular systems of biological and pharmaceutical relevance.

1 Introduction

Molecular dynamics (MD) simulations have found widespread applications in Chemistry, Biology, and Material Sciences due to their ability to study the mechanisms of molecular processes in atomistic resolution. There are two longstanding challenges in the field of MD simulations pertaining to the limited timescales and lengthscales accessible to conventional all-atom MD techniques.[1] The timescale problem stems from the presence of high energy barriers preventing transitions between interesting metastable confirmations in computationally accessible simulation timescales. In addition, the slow diffusion across rugged free energy

landscapes in systems such as protein folding and ligand-receptor binding also contributes to the increase of the transition times.

Several importance sampling algorithms have been developed over the past few decades to address this timescale problem in MD simulation. These methods can be categorized into two major groups: biased enhanced sampling and unbiased path sampling. In enhanced sampling approaches, an external biasing potential accelerates the dynamics and the escape from deep free energy minima.[2] Methods in this category include umbrella sampling (US),[3] metadynamics (MetaD),[4,5] adaptive biasing force (ABF),[6] Gaussian Accelerated Molecular Dynamics (GaMD),[7] etc. Contrarily, path-sampling algorithms avoid the application of external bias potential. They use the statistical properties of the unbiased trajectory ensembles to increase sampling in low-probability regions of the conformational space. Methods in this category include transition path sampling (TPS),[8,9] transition interface sampling (TIS),[10] forward flux sampling (FFS),[11,12] weighted ensemble (WE),[13,14] milestoning,[15,16] etc.

Although free energy landscapes provide a thermodynamic perspective, kinetics is often important to get a complete mechanistic picture of molecular processes e.g. ligand-binding, enzymatic reactions, heterogeneous catalysis, membrane permeation, and nucleation. The consideration kinetics is particularly relevant in computational drug design as the efficacy of small molecule drugs is better correlated with their residence time i.e. unbinding kinetics[17–19] than the binding free energy which is usually used to screen lead molecules.[20] While biased enhanced sampling methods can successfully compute free energy landscapes, the external bias distorts the natural dynamics of the system making it difficult to recover the correct kinetics. Specialized variants of existing enhanced sampling methods (e.g. infrequent metadynamics[21] and GaMD[7]) have been designed in recent days for calculating kinetics. Such methods have correctly predicted the kinetics of several molecular systems including drug-target residence times (see e.g. Ref. 22, 23). However, these methods require one to apply the bias conservatively to keep the transition states bias-free, a necessary condition for recovering the unbiased timescales.[21] It therefore requires one to compromise significantly on

the simulation efficiency.

Path sampling methods, contrarily, are unbiased by design and therefore the kinetics can be directly recovered. Except for TPS and TIS which initiate trajectories from transition regions, path sampling methods, in general, do not provide any explicit mechanism to overcome high enthalpic barriers to escape from deep free energy minima. Consequently, the overall computational costs of these methods are quite high. Therefore, one requires several microseconds of MD simulation to compute unbinding rate constants of realistic drugs molecules using either path sampling[24–29] or enhanced sampling[22,30–36] methods.

The fields of biased enhanced sampling and unbiased path sampling have developed independently over the past few decades. In this work, we introduce an integrated sampling (IS) approach that combines weighted ensemble with the enhanced sampling method on-the-fly probability enhanced sampling (OPES).[37] Specifically, we use the flooding variant[38] of the OPES algorithm as it is designed for calculating kinetics. The objective of this merger is to combine the strengths and mitigate the deficiencies of individual algorithms and to increase the efficiency of kinetics calculation from MD simulations. In our integrated approach, bias potential and trajectory resampling are performed simultaneously. The external bias potential facilitates the crossing of high energy barriers while the path sampling component accelerates the diffusion across the rugged free energy landscape often encountered in biophysical systems. This allows one to apply the bias potential conservatively to avoid biasing the transition state without having to compromise the efficiency of the simulation. It also provides a mechanism to accelerate the dynamics in the transition region which can be quite broad in many biophysical processes like protein folding and ligand binding. In the following sections, we describe the theoretical underpinnings of OPES flooding, weighted ensemble, and our integrated sampling algorithms, followed by the demonstration of our approach to peptide conformational transitions, protein unfolding, and ligand-receptor unbinding.

2 Theory

2.1 OPES-Flooding

OPES-flooding ($\mathrm{OPES_f}$) is an enhanced sampling algorithm for calculating kinetics of molecular rare-events.[38] It is based on the OPES algorithm[37] which leads to a quicker convergence of the bias deposited in the initial state compared to its predecessor, metadynamics.[4,5] In OPES, the bias potential $V(\mathbf{s})$ is built along a low-dimensional collective variable space $\mathbf{s}$ which is designed to encode the slow modes of the system. $V(\mathbf{s})$ is constructed from an on-the-fly estimate of the unbiased probability distribution $P(\mathbf{s})$:

$$V(\mathbf{s}) = -\frac{1}{\beta} \ln \frac{p^{\mathrm{tg}}(\mathbf{s})}{P(\mathbf{s})} \,. \tag{1}$$

where $p^{\mathrm{tg}}(\mathbf{s})$ is the target distribution sampled in biased simulation. As the target distribution we use the well-tempered distribution $p^{\mathrm{tg}}(\mathbf{s}) \propto P(\mathbf{s})^{1/\gamma}$, where $\gamma = \beta \Delta E$. The ΔE parameter dictates the maximum amount of bias that can be deposited during an OPES simulation. The bias potential in the n-th iteration is given by

$$V_n(\mathbf{s}) = (1 - 1/\gamma)\frac{1}{\beta} \ln \left(\frac{P_n(\mathbf{s})}{Z_n} + \epsilon \right) , \tag{2}$$

where $P_n(\mathbf{s})$ is the estimated unbiased marginal probability distribution along $\mathbf{s}$ at step n: $P_n(s) = \sum_k^n w_k G_k(s, s_k) / \sum_k^n w_k$. Here $G_k(s, s_k)$ are Gaussian Kernels and w_k is the weight of k-th kernel computed as $w_k = \exp(\beta V_{k-1}(s_k))$. The Z and ϵ are a normalization factor and a regularization term, respectively. They are introduced to ensure numerical stability. In $\mathrm{OPES_f}$ simulation, the ΔE is set to be lower than the barrier height one wishes to overcome. In addition, an excluded region $\chi_{\mathrm{exc}}(s, s_{\mathrm{exc}})$ parameter is used to avoid biasing beyond the threshold of $s = s_{\mathrm{exc}}$ A careful choice of the ΔE and s_{exc} parameter ensures that no bias is deposited in the transition state. Under these conditions, unbiased transition time τ can be

obtained by rescaling the biased simulation timescales:

$$\tau = \sum_{i}^{N_{tot}} \Delta t \exp(\beta V(\mathbf{s}_i)) \tag{3}$$

where Δt is the integration timestep, $\mathbf{s}_i$ is the location of the system in the CV space at step i, and N_{tot} is the total number of steps propagated to observe the transition.

2.2 Weighted Ensemble

Weighted ensemble (WE) is a path-sampling algorithm that accelerates the simulation of rare events through statistical resampling of the trajectory ensemble. In this approach, one stratifies the configuration space into multiple bins. Several trajectories (say N) are initiated at the starting configuration. An initial weight (w_1) of $1/N$ is assigned to each trajectory segment. The progress of the simulation is monitored at a fixed time interval δt by projecting the trajectory along a "progress coordinate" which is equivalent to the CV (s) in enhanced sampling. If a trajectory segment reaches a new bin it is stopped and some new trajectories are initiated from its endpoint. The weight of the old (parent) trajectory is distributed equally among the new (daughter) ones. This resampling is continued such that every occupied bin contains exactly N trajectories. If more than N replicas enter a bin, the excess ones are terminated and their weights are redistributed among the surviving trajectories. This ensures that the total probability remains conserved while increasing sampling in less probable regions of the conformational space. As no external bias is applied, the natural dynamics of the system is preserved, making it possible to recover kinetic properties directly. Unlike enhanced sampling where weights are assigned to configurations, WE method assigns weights to trajectory segments making it possible to calculate the properties of the trajectory ensemble through appropriate reweighting. However, the kinetics are computed, in general, by establishing a non-equilibrium steady state by recycling trajectories from the target state. The rate constant k and mean first passage time (MFPT) can be estimated by the "Hill

relation":[39]

$$k = \frac{1}{\mathrm{MFPT}(A \to B)} = \mathrm{flux}(SS, A \to B) = \frac{\sum_k w_k}{t} \tag{4}$$

where SS refers to steady state, A and B are the initial and target states respectively, w_k is the weight of the k-th reactive trajectory and t is the simulation time. More sophisticated algorithms such as the history-augmented Markov state modeling (haMSM)[40] and the rate from event durations (RED) scheme[41] can provide a better estimation of the kinetics.

2.3 Integrated Sampling

Here we re-iterate our motivation behind integrating path sampling and enhanced sampling algorithms. Enhanced sampling methods are efficient in crossing enthalpic barriers due to their bias deposition protocol which elevates the potential energy of the metastable states and helps escape free energy minima. However, they are not ideal for accelerating diffusive processes on a free energy plateau. Methods like weighted ensemble perform well in accelerating diffusive processes as demonstrated in its successful application to many protein folding processes. However, they do not have any explicit mechanism to overcome enthalpic barriers, making them suboptimal for escaping deep free energy minima, where the potential energy increases sharply as a function of the progress coordinate (Fig. 1).

The OPES-flooding algorithm, in particular, has another limitation. As no bias should be deposited in the transition state region one needs to have pre-existing knowledge about the free energy landscape which can be very expensive to compute. This can be particularly challenging in the case of a multi-state system or a process that involves escaping a deep energy minimum followed by diffusion across a rough landscape. The latter situation is pervasive in biomolecular processes. In such situations, the only option is to deposit bias in the initial state minimum and wait for the system to cross the barrier and diffuse toward the final state. This can be inefficient with large biomolecules where the transition time, despite being small compared to the total first passage time, can be in the beyond microsecond

regime and computationally unaffordable. Furthermore, when the transition state region is broad and diffusive, more time is spent traversing through the TS region than sampling in the initial state basin. Such dynamics, however, cannot be accelerated through metadynamics-like methods when one needs to ensure that no bias is deposited in the TS.

To overcome these issues we propose an integrated sampling scheme where OPES-flooding and WE are performed simultaneously. The OPES-flooding simulation is performed to converge the bias potential in the initial state free energy minimum. The weighted ensemble resampling is performed thereafter to accelerate the diffusive dynamics while keeping the OPES$_f$ bias ($V(\mathbf{s})$) constant (Fig. 1). The time rescaling procedure from Eq. 3 is then performed on the trajectory traces of successful transitions (Fig. 2). A weighted average of these rescaled transition times is computed from the trajectory ensemble as:

$$\langle \tau \rangle = \frac{\sum_k^M w_k \tau_k}{\sum_k^M w_k}, \tag{5}$$

where M is the total number of successful transitions sampled, τ_k is the rescaled time for the k-th transition and w_k is the corresponding weight. The mean first passage time (MFPT) is estimated to be equal to this weighted average. The Eq. 5 is different from the Hill relation (Eq. 4), commonly used to compute rate constants from WE simulations. As the transition times are rescaled in our integrated sampling approach, two trajectory segments reaching the final state in the same iteration can have vastly different τ based on their splitting-merging history. Therefore, it is difficult to estimate the flux from the WE iterations alone. Furthermore, the rescaled transition times are dominated by the biased portion of the trajectory as the rescaling factor scales exponentially with the bias. Our integrated scheme, therefore, can be understood as a modified form of OPES-flooding where the weighted ensemble algorithm is used only to accelerate the diffusive part of the dynamics. So, the kinetics can be estimated with reasonable accuracy by performing a weighted average over the rescaled transition times of all successful events.

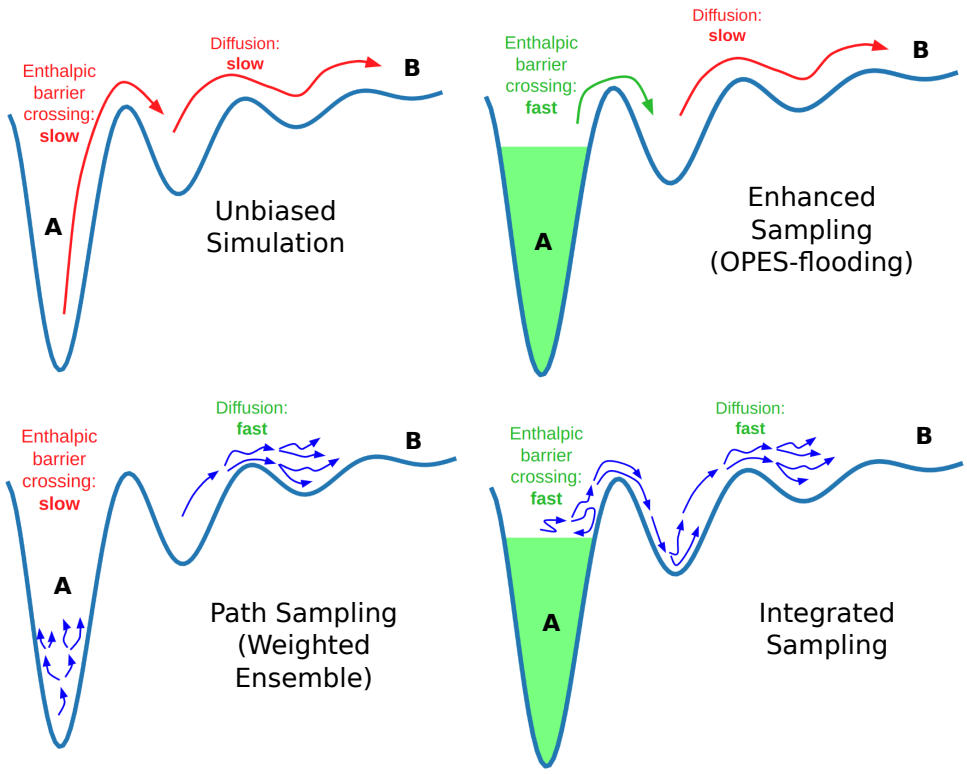

Figure 1: Schematic of different approaches of rare event sampling in molecules. The objective is to calculate the kinetics of going from state A to state B. In enhanced sampling approaches such as OPES flooding, external bias (solid green shade) is deposited in the initial state basin. In path sampling methods like weighted ensemble multiple trajectory segments (blue arrows) are generated through a resampling procedure. In the integrated sampling method, both are performed simultaneously.

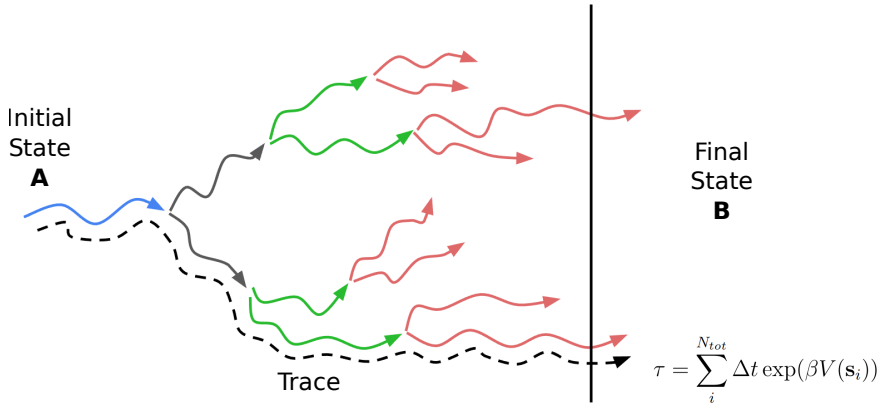

Figure 2: In weighted ensemble, many trajectory segments sample the conformational space. In integrated sampling, the kinetics is only computed by tracing a successful transition back to its origin in the initial state. One of the two such traces is depicted in a gray dashed line. The unbiased transition time is then obtained by reweighting the timescale of the trace based on the external bias deposited along its path.

3 Results

In this section, we demonstrate the application of the integrated sampling algorithm on the conformational transition of alanine dipeptide, unfolding of the chignolin mini-protein, and the ligand unbinding from the calixerene host.

3.1 Alanine Dipeptide

First, we tested our integrated sampling algorithm on the $C7_{eq}$ to $C7_{ax}$ conformational transition in Alanine dipeptide with a simulation setup identical to Ref. 38. A periodic function of the ϕ torsion angle is used as the CV for OPES-flooding and the progress coordinates for WE simulation. We applied the bias conservatively and avoided biasing the TS through an appropriate choice of the barrier parameter (ΔE) and excluded region s_{exc}. OPES$_f$ simulations are conducted until we observe the system to escape the initial state ($C7_{eq}$) minimum. Then the OPES$_f$ bias is used as a static bias to perform weighted ensemble simulations. Trajectories are terminated and recycled when they reach the $C7_{ax}$ state.

The estimated Mean First Passage Time (MFPT) is $\langle\tau\rangle = 0.84 \pm 0.29$ μs after 500 iterations, and $\langle\tau\rangle = 1.02 \pm 0.35$ μs after the complete run (1000 iterations) (SI Table S1). These results are in agreement with the unbiased estimate of $\langle\tau\rangle_{unbiased} = 1.28\mu$s.[38] The timescales converged within 500 iterations of the integrated sampling (Fig. 3). The first 500 iterations required ∼21.6 ns of total simulation time (including all WE segments) for each independent run resulting in 165, 119, and 112 transitions respectively. In comparison, OPES$_f$ simulations with identical conditions required ∼10.9 ns of total simulation time to observe 30 transitions. However, the transitions observed in the WE method are correlated as the trajectory traces share some portions of their propagation history. The free energy surface of alanine dipeptide is rather simple with a single barrier-crossing event with no requirement for traversing a rugged free energy landscape. Therefore, simple OPES flooding is sufficient for this problem and the computational gain for the integrated sampling method

is not apparent. Nevertheless, this proof of concept example demonstrates that the IS algorithm can predict the correct unbiased kinetics.

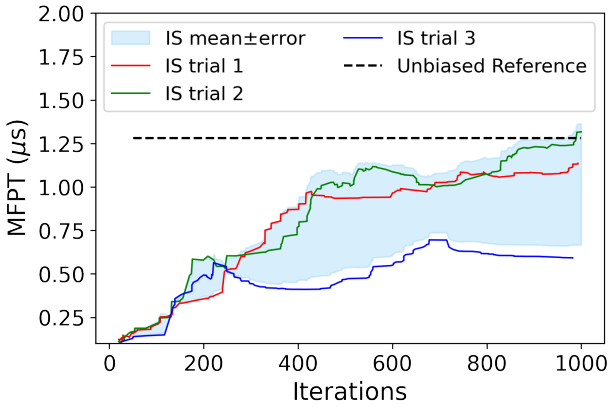

Figure 3: Convergence of the mean first passage time of $C7_{eq}$ to $C7_{ax}$ transition in alanine dipeptide. The uncertainty (light blue shade) is computed as the 95% confidence interval from three independent sets of integrated sampling (IS) simulations.

3.2 Chignolin

Next, we tested our method on the unfolding of the chignolin mini-protein. Chignolin is a 10-residue polypeptide with a free energy landscape resembling typical protein folding landscapes. In the folded configuration it remains as a β-hairpin which can unfold into a disordered state in the μs timescale. In previous work, the folding-unfolding dynamics of Chignolin have been studied using the Anton supercomputer by performing a $\sim$107 μs unbiased MD simulation, which predicted the unfolding timescale to be 2.2$\pm$0.4 μs.[42] In our previous work,[38] we used the OPES-flooding algorithm to study the kinetics of Chignolin unfolding using simulation conditions identical to Ref. 42 to make a direct comparison with the unbiased data. We use the same setup in the present study. As the biasing CV and the WE progress coordinate, we use the Harmonic Linear Discriminant Analysis (HLDA) CV introduced by Mendels et al.[43] It is described by a linear combination of 6 pairwise contacts between the protein atoms.[44] We showed earlier[38,45] that the HLDA CV is sub-optimal as the efficiency and accuracy of the predicted unfolding kinetics was poorer with the HLDA

CV compared to the neural network based CVs such as Deep Linear Discriminant Analysis (Deep-LDA),[46] Deep Time-lagged Independent Component Analysis (Deep-TICA),[47] and Deep Targeted discriminant Analysis[48] . Using OPES-flooding we could obtain an acceleration factor close to 100,[38] which is acceptable for fast-folding proteins such as chignolin but is not adequate for the study of physiologically relevant processes with beyond millisecond timescales. Therefore, chignolin unfolding with HLDA CV also serves as a good test case on how our algorithm performs with a sub-optimal CV. Initially, we performed 3 independent OPES-flooding simulations to converge the bias in the folded state minimum. These trajectories did not reach the unfolded state, as that would require significant time to diffuse across the rough free energy landscape. Weighted ensemble simulations were initiated from the folded state in the presence of the static bias generated in initial OPES flooding simulations (Fig 4a).

We observed multiple unfolding events during the 1000 iterations of the integrated sampling simulation of the chignolin mini protein system. For computing kinetics, we only considered transitions for which the final weight is $\geq 10^{-8}$. Such transitions only appeared after ∼300 iterations of IS for all three independent replicas. In Fig 4b, we show that an integrated sampling trajectory spends less time in the transition region compared to an OPES-flooding trajectory, before reaching the unfolded state. This acceleration is a result of the ability of the weighted ensemble algorithm to accelerate diffusive processes. After 800 iterations the predicted $\langle\tau\rangle_{unfolding}$ (1.81 $\pm$ 0.43 μs) is virtually identical to the unbiased timescale of 2.2$\pm$0.4 μs. However, the unfolding timescales are well within an order of magnitude of the unbiased timescales throughout the entire duration of simulations (Fig. 4c). The total computational cost of 1000 iterations of integrated sampling simulation is ∼460 ns which is lower compared to the ∼825 ns of OPES-flooding simulation required to observe 15 unfolding events in Ref. 38. But, if we consider the fact that kinetics of similar accuracy as OPES$_f$[38] could be obtained from the first 300 iterations (∼150 ns) of the IS simulations, we can appreciate the significant improvement of the computational efficiency of the com-

bined approach. Notably, similar to the alanine dipeptide, the unfolding of chignolin also involves a free energy landscape with two prominent minima. Therefore, despite the increase in complexity, it is still a single barrier-crossing event. Even in this case, we can observe an improvement in sampling efficiency for our combined algorithm compared to standard enhanced sampling.

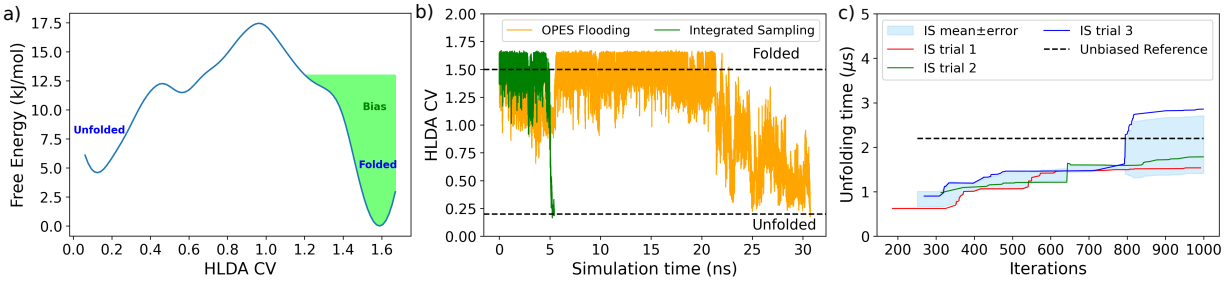

Figure 4: (a) Bias deposition scheme along the HLDA CV for sampling the unfolding transitions in chignolin mini-protein. (b) Representative unfolding trajectories, sampled using OPES-flooding and integrated sampling, projected along the HLDA CV. Representative trajectories were chosen as the ones with rescaled transition time closest to the estimated MFPT from all trajectories. In case of integrated sampling, the simulation time axis is equivalent to the molecular time in the weighted ensemble. (c) Convergence of the unfolding time obtained from integrated sampling. The uncertainty (light blue shade) is computed as the 95% confidence interval from three independent simulations.

3.3 Ligand-Receptor Complex

Next, we studied the unbinding of guest 4 (G4) from the OAMe calixerene host. This system and similar calixerine-based host-guest complexes have been used to benchmark importance sampling algorithms for calculating free energy[49,50] and ligand residence time.[51] The z-projection of the center of mass distance between the ligand and the binding pocket is used as CV. Previous studies have indicated that water coordination plays a prominent role in the ligand unbinding process for this system.[50,51] However, we excluded any description of water coordination to make the CV sup-optimal. Similar to the HLDA CV for chignolin, it allows us to evaluate this method for situations where the CV is not highly optimized. Out of the different guest molecules investigated for this specific host, the G4 shows a rather complex

free energy landscape due to the presence of a metastable intermediate state between the bound and unbound configuration.[50] Therefore, deciding on an optimal excluded region is difficult even with the knowledge of the complete free energy landscape. In this work, we took a conservative approach and restricted the bias deposition to only the bound state minimum (Fig. 5a). In addition to the integrated sampling, OPES-flooding simulations have been performed to compare the results.

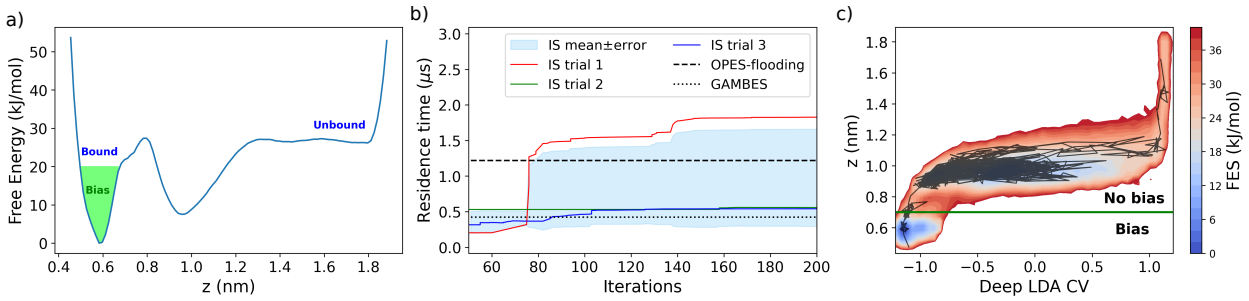

Figure 5: (a) Bias deposition scheme along the z-axis CV for sampling the dissociation of the OAMe-G4 host guest complex. The biasing scheme is designed blindly by only depositing bias in the bound state minimum and not the intermediate. (b) Convergence of the ligand residence time of G4 obtained from integrated sampling. The uncertainty (light blue shade) is computed as the 95% confidence interval from the three independent simulations. (c) Projection of the trajectory with the highest weight on the 2D free energy surface along the biasing CV (z) and the Deep-LDA CV trained to model the water hydration behavior.[50] Although we did not bias the water, the most probable trace of the integrated sampling algorithm sampled the minimum free energy pathway in the 2D space. Similar plots for the 10 most probable trajectory traces are provided in the supporting information.

Within the first 200 iterations of the integrated sampling, the residence time estimate is converged in all three independent replicas. The computed values (Fig. 5b and Table S3) are well within one order of magnitude of the kinetics estimated from OPES-flooding and Gaussian Mixture Based Enhanced Sampling (GAMBES) simulations.[51,52] The cumulative simulation time of 200 iterations of integrated sampling is less than 100 ns while the OPES-flooding estimate was obtained from 30 transitions sampled from ∼500 ns of simulation. It should be noted that the kinetics obtained from IS can be considered converged at any point beyond the first 100 iterations for all practical purposes. However, we confirmed that the results truly converged by extending all three replicas to 1000 iterations each (Supporting

Information Table S7 and Fig. S2).

As we did not use any description of water in the CV space, it is important to verify whether the unbinding pathways sampled from our IS algorithm are consistent with previous works. Therefore, we projected the transition paths with the highest weights onto a two dimensional free energy landscape composed of a distance CV and a water CV. For the distance CV, we chose the z CV used for biasing while for the water CV we chose the Deep Linear Discriminant Analysis (Deep-LDA) CV[46] trained by Rizzi et al. using the water coordination of the host and guest atoms as descriptors.[50] We observe that the most probable unbinding transitions obtained from IS simulations follow the minimum free energy pathway in this 2D space (Fig. 5c). Notably, no bias is applied outside the bound state minimum. Therefore, the WE resampling is the only driver of the transitions in this region. The performance of WE sampling is not as sensitive to the choice of CV as the OPES simulations. So the role of water is correctly recovered even when the water coordination description is not explicitly included in the CV space.

4 Discussions and Conclusions

The two classes of importance sampling methods: unbiased path sampling and biased enhanced sampling, offer unique advantages. But, when combined effectively, they can significantly improve the sampling efficiency. In this work, we demonstrate one such scenario for calculating the kinetics of processes involving one barrier-crossing event followed by a diffusive process. Such a situation is commonly observed in biomolecular processes such protein-ligand unbinding and protein conformational transitions. Enhanced sampling methods i.e. metadynamics and OPES can accelerate the enthalpic barrier crossing events but are not as effective in accelerating diffusive processes. Weighted Ensemble, a path sampling algorithm, can accelerate diffusive dynamics while being less efficient in sampling barrier crossing transitions due to the absence of external biasing force. We show that a well-designed combination

of these two techniques can synergize the advantages and mitigate the deficiencies of each other. We conduct the flooding variant of OPES simulation at the initial state to stimulate the escape from a deep free energy minimum. The Weighted Ensemble resampling technique then accelerates the diffusion on this modified potential energy surface (due to the presence of the OPES bias) and helps the system reach the target state. The unbiased kinetics is recovered by appropriately reweighting the biased transition times taking into account both the external bias potential and the trajectory weights from the WE simulation. This approach does not need one to reach the nonequilibrium steady state usually required for standard WE simulations. In this sense, the integrated sampling method is inherently different from an earlier attempt to combine WE with GaMD by Ahn et al. They performed GaMD to sample protein conformational landscape for optimal selection of the starting coordinates of a weighted ensemble simulation.[53] Although this approach increased the convergence speed of the WE simulation, both methods did not contribute simultaneously to the exploration of the free energy landscape. Combinations of different path-sampling algorithms have also been shown to increase the computational efficiency in predicting rare event kinetics (see e.g. weighted ensemble milestoning[54] and Markovian weighted ensemble milestoning[55]). But, our integrated sampling approach comes with the benefit of sampling continuous pathways from the initial to the final state, which is inaccessible to milestoning-based methods.

We demonstrated the successful application of our integrated sampling algorithm on a model system of alanine dipeptide as well as the unfolding of chignolin mini-protein and the dissociation of a host-guest complex. Despite the apparent simplicity of these systems, we could observe faster convergence of kinetics in comparison to OPES flooding. We also observe that the efficiency increases with the increasing complexity of the system, particularly with the presence of additional kinetic and diffusive bottlenecks outside the initial state minimum.

We envision that the true potential of the integrated sampling approach can be realized in complex drug-receptor unbinding problems where traversing the free energy landscape involves both high enthalpic barriers as well as rugged plateaus where the system needs to

diffuse towards the final state (see e.g. Ref. 56). Through the improvement in computational efficiency, the integrated sampling algorithm will help address the challenge of large-scale screening of drug candidates based on their ligand residence time, a property that correlates better with their physiological activity compared to binding affinity. At a fundamental level, it should also be possible to design alternative schemes of combining weighted ensemble with metadynamics and related methods to study other challenging problems including conformational exploration and free energy surface reconstruction for highly complex systems with multiple barriers and a rugged free energy surface. We hope that this work will encourage future studies in these directions.

Acknowledgement

The author thanks Sudip Das and Jintu Zhang for critically reading and providing feedback on the manuscript.

5 Data Availability Statement

The input files for all simulations performed in this work are provided in the GitHub repository: `https://github.com/dhimanray/Enhanced_Sampling_Path_Sampling.git`. The input files will also be made available through the PLUMED NEST repository.[57] All simulations are performed with the GROMACS 2021[58] package patched with PLUMED 2.9[59] and WESTPA 2.0.[60] These are all open-source software.

Supporting Information Available

Computational details and additional results concerning the convergence and computational cost of integrated sampling simulations are provided in the supplementary information.

References

(1) Hollingsworth, S. A.; Dror, R. O. Molecular dynamics simulation for all. *Neuron* **2018**, *99*, 1129–1143.

(2) Hénin, J.; Lelièvre, T.; Shirts, M.; Valsson, O.; Delemotte, L. Enhanced sampling methods for molecular dynamics simulations. *Living Journal of Computational Molecular Science* **2022**, *4*.

(3) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of computational physics* **1977**, *23*, 187–199.

(4) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the national academy of sciences* **2002**, *99*, 12562–12566.

(5) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters* **2008**, *100*, 020603.

(6) Darve, E.; Pohorille, A. Calculating free energies using average force. *The Journal of chemical physics* **2001**, *115*, 9169–9183.

(7) Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. *Journal of chemical theory and computation* **2015**, *11*, 3584–3595.

(8) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants. *The Journal of chemical physics* **1998**, *108*, 1964–1977.

(9) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry* **2002**, *53*, 291–318.

(10) van Erp, T. S.; Moroni, D.; Bolhuis, P. G. A novel path sampling method for the calculation of rate constants. *The Journal of Chemical Physics* **2003**, *118*, 7762–7774.

(11) Allen, R. J.; Warren, P. B.; Ten Wolde, P. R. Sampling rare switching events in biochemical networks. *Physical review letters* **2005**, *94*, 018104.

(12) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. Simulating rare events in equilibrium or nonequilibrium stochastic systems. *The Journal of chemical physics* **2006**, *124*.

(13) Huber, G. A.; Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophysical journal* **1996**, *70*, 97–110.

(14) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *The Journal of chemical physics* **2010**, *132*.

(15) Faradjian, A. K.; Elber, R. Computing time scales from reaction coordinates by milestoning. *The Journal of chemical physics* **2004**, *120*, 10880–10889.

(16) Bello-Rivas, J. M.; Elber, R. Exact milestoning. *The Journal of Chemical Physics* **2015**, *142*.

(17) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nature Reviews Drug Discovery* **2006**, *5*, 730–739.

(18) Guo, D.; Mulder-Krieger, T.; IJzerman, A. P.; Heitman, L. H. Functional efficacy of adenosine A 2A receptor agonists is positively correlated to their receptor residence time. *British Journal of Pharmacology* **2012**, *166*, 1846–1859.

(19) Copeland, R. A. The drug–target residence time model: a 10-year retrospective. *Nature Reviews Drug Discovery 2015 15:2* **2015**, *15*, 87–95.

(20) York, D. M. Modern alchemical free energy methods for drug discovery explained. *ACS Physical Chemistry Au* **2023**, *3*, 478–491.

(21) Tiwary, P.; Parrinello, M. From metadynamics to dynamics. *Physical review letters* **2013**, *111*, 230602.

(22) Ray, D.; Parrinello, M. Kinetics from metadynamics: Principles, applications, and outlook. *Journal of Chemical Theory and Computation* **2023**, *19*, 5649–5670.

(23) Wang, J.; Arantes, P. R.; Bhattarai, A.; Hsu, R. V.; Pawnikar, S.; Huang, Y.-m. M.; Palermo, G.; Miao, Y. Gaussian accelerated molecular dynamics: Principles and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021**, *11*, e1521.

(24) Lotz, S. D.; Dickson, A. Unbiased molecular dynamics of 11 min timescale drug unbinding reveals transition state stabilizing interactions. *Journal of the American Chemical Society* **2018**, *140*, 618–628.

(25) Bose, S.; Lotz, S. D.; Deb, I.; Shuck, M.; Lee, K. S. S.; Dickson, A. How robust is the ligand binding transition state? *Journal of the American Chemical Society* **2023**, *145*, 25318–25331.

(26) Dixon, T.; Uyar, A.; Ferguson-Miller, S.; Dickson, A. Membrane-mediated ligand unbinding of the PK-11195 ligand from TSPO. *Biophysical journal* **2021**, *120*, 158–167.

(27) Narayan, B.; Buchete, N.-V.; Elber, R. Computer simulations of the dissociation mechanism of Gleevec from Abl Kinase with milestoning. *The Journal of Physical Chemistry B* **2021**, *125*, 5706–5715.

(28) Rathnayake, S.; Narayan, B.; Elber, R.; Wong, C. F. Milestoning simulation of ligand dissociation from the glycogen synthase kinase 3β. *Proteins: Structure, Function, and Bioinformatics* **2023**, *91*, 209–217.

(29) Ojha, A. A.; Srivastava, A.; Votapka, L. W.; Amaro, R. E. Selectivity and ranking of

tight-binding JAK-STAT inhibitors using Markovian milestoning with Voronoi tessellations. *Journal of Chemical Information and Modeling* **2023**, *63*, 2469–2482.

(30) Tiwary, P.; Mondal, J.; Berne, B. J. How and when does an anticancer drug leave its binding site? *Science advances* **2017**, *3*, e1700014.

(31) Casasnovas, R.; Limongelli, V.; Tiwary, P.; Carloni, P.; Parrinello, M. Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations. *Journal of the American Chemical Society* **2017**, *139*, 4780–4788.

(32) Ahmad, K.; Rizzi, A.; Capelli, R.; Mandelli, D.; Lyu, W.; Carloni, P. Enhanced-sampling simulations for the estimation of ligand binding kinetics: current status and perspective. *Frontiers in molecular biosciences* **2022**, *9*, 899805.

(33) Lamim Ribeiro, J. M.; Provasi, D.; Filizola, M. A combination of machine learning and infrequent metadynamics to efficiently predict kinetic rates, transition states, and molecular determinants of drug dissociation from G protein-coupled receptors. *The Journal of Chemical Physics* **2020**, *153*.

(34) Shekhar, M.; Smith, Z.; Seeliger, M. A.; Tiwary, P. Protein flexibility and dissociation pathway differentiation can explain onset of resistance mutations in kinases. *Angewandte Chemie International Edition* **2022**, *61*, e202200983.

(35) Wang, Y.-T.; Liao, J.-M.; Lin, W.-W.; Li, C.-C.; Huang, B.-C.; Cheng, T.-L.; Chen, T.-C. Structural insights into Nirmatrelvir (PF-07321332)-3C-like SARS-CoV-2 protease complexation: a ligand Gaussian accelerated molecular dynamics study. *Physical Chemistry Chemical Physics* **2022**, *24*, 22898–22904.

(36) Wang, J.; Do, H. N.; Koirala, K.; Miao, Y. Predicting biomolecular binding kinetics: A review. *Journal of Chemical Theory and Computation* **2023**, *19*, 2135–2148.

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

(38) Ray, D.; Ansari, N.; Rizzi, V.; Invernizzi, M.; Parrinello, M. Rare event kinetics from adaptive bias enhanced sampling. *Journal of Chemical Theory and Computation* **2022**, *18*, 6500–6509.

(39) Bhatt, D.; Zhang, B. W.; Zuckerman, D. M. Steady-state simulations using weighted ensemble path sampling. *The Journal of chemical physics* **2010**, *133*.

(40) Suárez, E.; Lettieri, S.; Zwier, M. C.; Subramanian, S. R.; Chong, L. T.; Zuckerman, D. M. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *Biophysical Journal* **2014**, *106*, 406a.

(41) DeGrave, A. J.; Bogetti, A. T.; Chong, L. T. The RED scheme: Rate-constant estimation from pre-steady state weighted ensemble simulations. *The Journal of Chemical Physics* **2021**, *154*.

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

(43) Mendels, D.; Piccini, G.; Parrinello, M. Collective variables from local fluctuations. *The journal of physical chemistry letters* **2018**, *9*, 2776–2781.

(44) Mendels, D.; Piccini, G.; Brotzakis, Z. F.; Yang, Y. I.; Parrinello, M. Folding a small protein using harmonic linear discriminant analysis. *The Journal of chemical physics* **2018**, *149*.

(45) Ray, D.; Trizio, E.; Parrinello, M. Deep learning collective variables from transition path ensemble. *The Journal of Chemical Physics* **2023**, *158*.

(46) Bonati, L.; Rizzi, V.; Parrinello, M. Data-driven collective variables for enhanced sampling. *The journal of physical chemistry letters* **2020**, *11*, 2998–3004.

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

(48) Trizio, E.; Parrinello, M. From enhanced sampling to reaction profiles. *The Journal of Physical Chemistry Letters* **2021**, *12*, 8621–8626.

(49) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K. Overview of the SAMPL5 host–guest challenge: Are we doing better? *Journal of computer-aided molecular design* **2017**, *31*, 1–19.

(50) Rizzi, V.; Bonati, L.; Ansari, N.; Parrinello, M. The role of water in host-guest interaction. *Nature Communications* **2021**, *12*, 93.

(51) Debnath, J.; Parrinello, M. Computing rates and understanding unbinding mechanisms in host–guest systems. *Journal of Chemical Theory and Computation* **2022**, *18*, 1314–1319.

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

(53) Ahn, S.-H.; Ojha, A. A.; Amaro, R. E.; McCammon, J. A. Gaussian-accelerated molecular dynamics with the weighted ensemble method: A hybrid method improves thermodynamic and kinetic sampling. *Journal of chemical theory and computation* **2021**, *17*, 7938–7951.

(54) Ray, D.; Andricioaei, I. Weighted ensemble milestoning (WEM): A combined approach for rare event simulations. *The Journal of chemical physics* **2020**, *152*.

(55) Ray, D.; Stone, S. E.; Andricioaei, I. Markovian weighted ensemble milestoning (M-WEM): Long-time kinetics from short trajectories. *Journal of Chemical Theory and Computation* **2021**, *18*, 79–95.

(56) Tang, Z.; Chen, S.-H.; Chang, C.-E. A. Transient states and barriers from molecular simulations and the milestoning Theory: kinetics in ligand–protein recognition and compound design. *Journal of chemical theory and computation* **2020**, *16*, 1882–1895.

(57) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; Capelli, R.; Carloni, P.; Ceriotti, M.; Cesari, A.; Chen, H.; Chen, W.; Colizzi, F.; De, S.; De La Pierre, M.; Donadio, D.; Drobot, V.; Ensing, B.; Ferguson, A. L.; Filizola, M.; Fraser, J. S.; Fu, H.; Gasparotto, P.; Gervasio, F. L.; Giberti, F.; Gil-Ley, A.; Giorgino, T.; Heller, G. T.; Hocky, G. M.; Iannuzzi, M.; Invernizzi, M.; Jelfs, K. E.; Jussupow, A.; Kirilin, E.; Laio, A.; Limongelli, V.; Lindorff-Larsen, K.; Löhr, T.; Marinelli, F.; Martin-Samos, L.; Masetti, M.; Meyer, R.; Michaelides, A.; Molteni, C.; Morishita, T.; Nava, M.; Paissoni, C.; Papaleo, E.; Parrinello, M.; Pfaendtner, J.; Piaggi, P.; Piccini, G.; Pietropaolo, A.; Pietrucci, F.; Pipolo, S.; Provasi, D.; Quigley, D.; Raiteri, P.; Raniolo, S.; Rydzewski, J.; Salvalaglio, M.; Sosso, G. C.; Spiwok, V.; Šponer, J.; Swenson, D. W. H.; Tiwary, P.; Valsson, O.; Vendruscolo, M.; Voth, G. A.; White, A.; consortium, T. P. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods* **2019**, *16*, 670–673.

(58) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(59) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer physics communications* **2014**, *185*, 604–613.

(60) Russo, J. D.; Zhang, S.; Leung, J. M. G.; Bogetti, A. T.; Thompson, J. P.; DeGrave, A. J.; Torrillo, P. A.; Pratt, A. J.; Wong, K. F.; Xia, J.; Copperman, J.; Adelman, J. L.; Zwier, M. C.; LeBard, D. N.; Zuckerman, D. M.; Chong, L. T. WESTPA 2.0: High-Performance Upgrades for Weighted Ensemble Simulations and Analysis of

Longer-Timescale Applications. *Journal of Chemical Theory and Computation* **2022**, *18*, 638–649, PMID: 35043623.