OGRe: Optimal grid refinement protocol for accurate free energy surfaces and its application to proton hopping in zeolites and 2D COF stacking

Sander Borgmans, Sven M.J. Rogge,* Louis Vanduyfhuys, and Veronique Van Speybroeck*

Center for Molecular Modeling (CMM), Ghent University, Technologiepark-Zwijnaarde 46, 9052 Zwijnaarde, Belgium

E-mail: Sven.Rogge@UGent.be; Veronique.VanSpeybroeck@UGent.be

Abstract

While free energy surfaces are the crux of our understanding in many chemical and biological processes, their accuracy is generally unknown. Moreover, many developments to improve their accuracy are often complicated, impeding their general use. Luckily, several tools and guidelines are already in place to identify these shortcomings, but they are typically lacking in flexibility or fail to systematically determine how to improve the accuracy of the free energy calculation. To overcome these limitations, this work introduces OGRe–a python package for optimal grid refinement in an arbitrary number of dimensions. OGRe is based on three metrics which gauge the confinement, consistency, and overlap of each simulation in a series of umbrella sampling (US) simulations, an enhanced sampling technique ubiquitously adopted to construct free energy

surfaces for hindered processes. As these three metrics are fundamentally linked to the accuracy of the weighted histogram analysis method, adopted to generate free energy surfaces from US simulations, they facilitate a systematic construction of accurate free energy profiles, where each metric is driven by a specific umbrella parameter. This allows for the derivation of a consistent and optimal collection of umbrellas for each simulation, largely independent of the initial values, thereby dramatically increasing the ease-of-use towards accurate free energy surfaces. As such, OGRe is particularly suited to determined complex free energy surfaces, with large activation barriers and shallow minima, which underpin many physical and chemical transformations, and hence to further our fundamental understanding of these processes.

1 Introduction

Free energy surfaces (FESs) form the foundation of our understanding of many chemical and biological processes, and are an essential element to complement experimental studies. They are omnipresent to characterize complex physical and chemical transformation, such as reactions and phase transformations. Expressed as a function of judiciously chosen collective variables (CVs), a FES depicts the different (stable) system states and the transition paths between them. In this way, it sheds light on the underlying conformational thermodynamics and kinematics, and the effect thereon of specific system changes.[1–3] For instance, FESs are crucial to understand protein folding, and their accurate representation is of paramount importance to determine the stability, dynamics, and functional behavior of the designed compound.[4] As such, it is not surprising that, with the ever-increasing computational resources, FESs have found widespread use in (bio)chemistry, materials science, pharmacology, etc.[5,6] As the underlying theory has already been established for many years, recent endeavors focus on increasing the efficiency and accuracy by which these free energy profiles can be calculated.[7–10] Unfortunately, many recent methodological developments in the derivation

of FESs are typically overlooked due to their complexity or lack of reproducibility, warranting easy-to-use and intuitive practices to systematically improve the accuracy of free energy calculations. In this respect, some excellent guidelines and tools have already been introduced,[10,11] but they are limited to 1D, and only inform on inadequacies instead of automatically resolving them. In this work, we propose a systematic protocol to determine the free energy surfaces–in an arbitrary number of dimensions–based on umbrella sampling simulations, and introduce three simple and transparent metrics which allow to validate the accuracy of the FES prediction. It is applicable to activated processes, but is also particularly suited to describe transformations occupying a flat potential energy surface. This work was inspired by the layer dynamics, dominating the atomic geometry in 2D covalent organic frameworks, which is characterized by shallow minima. However, as we will demonstrate in this work through various model potentials and proton hopping taking place in zeolites, the protocol is much more generally applicable.

Choosing an optimal free energy calculation method is growing increasingly challenging. Aside from the sheer number of different techniques, and their varying complexity,[13–19] their efficacy generally depends, among others, on the system and the thermodynamic conditions. Yet, all these methods rely on an adequate sampling of the relevant portion of the phase space in order to capture the underlying partition function that gives rise to the overall probability density. In many cases, molecular dynamics (MD) simulations can explore a significant region of this phase space. However, many processes are highly activated, leading to non-ergodic sampling within a finite simulation time due to insurmountable barriers.[20] One of the more popular techniques to overcome this limitation is umbrella sampling (US), as illustrated in Figure 1.[9] It enables an enhanced sampling of the phase space by performing a series of short simulations, constraining each simulation to neighboring, and slightly overlapping, regions. These constraints are realized by introducing a bias potential or so-called umbrella. Because of this umbrella, each individual US simulation samples a limited region of the phase space, collectively encompassing the full region of interest. In this way, the full free

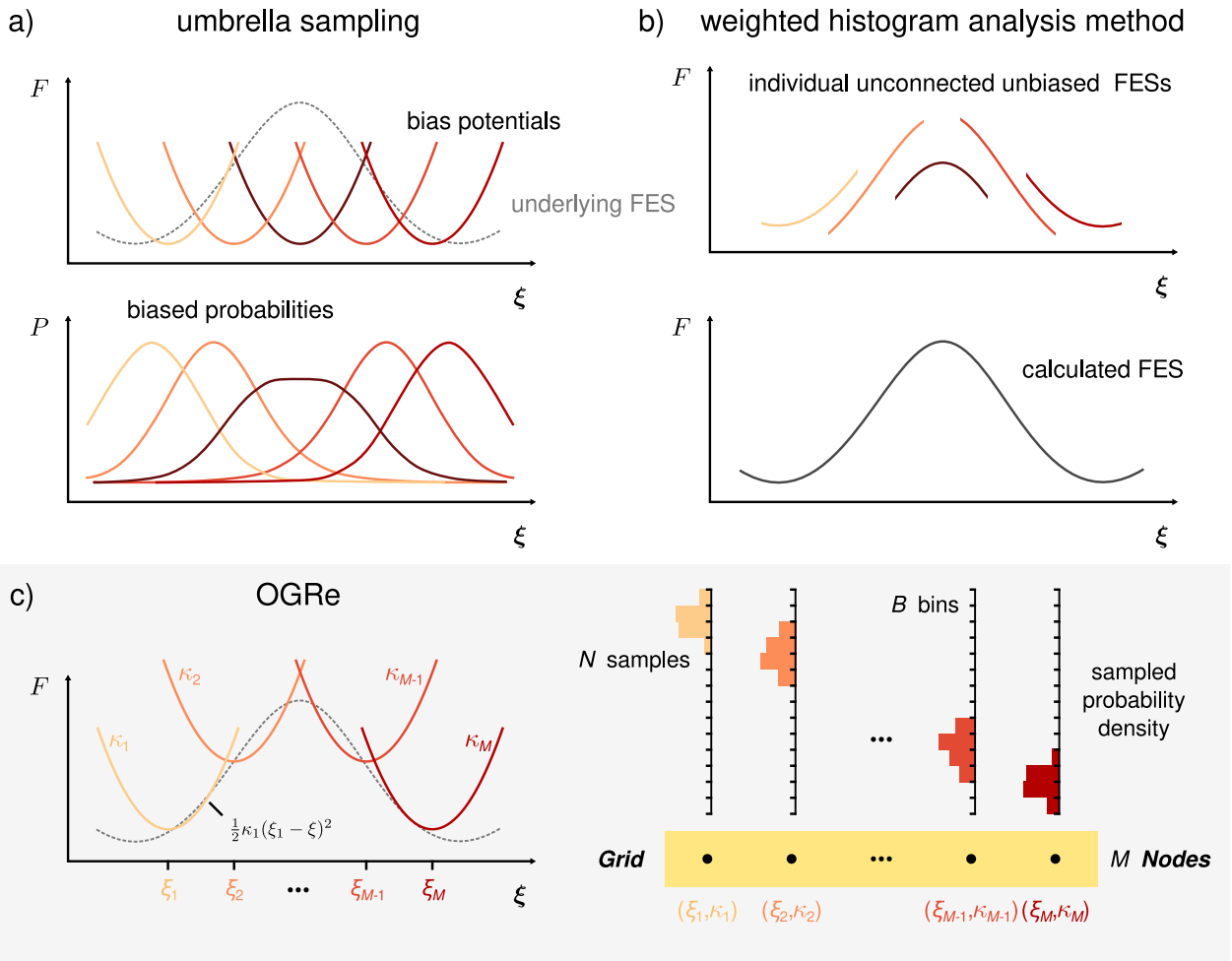

Figure 1: Overview of the umbrella sampling (US) and weighted histogram analysis method (WHAM) techniques, and their implementation in OGRe. (a) A series of US simulations samples the full region of interest by imposing bias potentials that force the sampling toward a specific region, resulting in biased probability densities. (b) WHAM converts these into an unbiased collective free energy profile for the full region of interest. (c) OGRe defines the collection of all US simulation parameters as the `Grid`, wherein each couple of parameters for a particular simulation is defined as a `Node`. Panels (a) and (b) are adapted from ref. 12 with permission of Elsevier, copyright 2017.

energy profile can be obtained after appropriately correcting for the applied biases, thereby relating the biased probability densities of neighboring regions, through, *e.g.*, the weighted histogram analysis method (WHAM).[21–23] Ideally, these biases are chosen to obtain a quasi-uniform sampling of the whole phase space of interest. However, this would require an exact knowledge of the sought-after FES, precluding an *a priori* optimal choice. Instead,

approximate approaches are adopted to identify (i) the locations around which the umbrellas are centered and (ii) the bias strength of these umbrellas, which we will collectively refer to as the umbrella grid parameters. These grid parameters can be estimated analytically,[24] or refined iteratively, such as in adaptive US.[13] Most grid parameter estimation methods rely on optimizing the bias strengths for a fixed collection of umbrella locations. These are often complemented by (manual) procedures that add additional umbrellas in regions where the sampling is inadequate, reposition the umbrellas for increased overlap in the region of interest,[25] or procedurally propagate the collection of umbrella locations (with a fixed step size) towards interesting regions.[18] While these approaches are valuable, they often require manual input and may lead to optimized grids that strongly depend on the initial conditions. This, in turn, may lead to suboptimal and inaccurate free energy estimates.

To overcome these limitations, we introduce in this work OGRe, an Optimal Grid Refinement protocol to construct accurate free energy surfaces through US. To this end, this standalone package iteratively and automatically refines an initial, uniform grid of umbrellas by adapting the bias strengths and locally adding umbrellas to create denser grids where necessary to improve the sampling. As we will demonstrate, this leads to a consistent and optimal grid refinement as the final grid definition–the umbrella centers and strengths–are largely independent of the initial values. This refinement is driven by several metrics, defining the reliability of individual simulations and the overlap between every pair of simulations with adjacent umbrellas, whose application could benefit any WHAM calculation and go beyond the OGRe protocol. By implementing our algorithms to accommodate an arbitrary number of dimensions, OGRe is intended to minimize the computational effort for free energy evaluation methods in N dimensions that require an overlap of simulated probability densities. The OGRe package is implemented in a user-friendly Python code, that is available from `https://github.com/SanderBorgmans/OGRe` and is easily adaptable and extendable for specific use cases.

The rest of the article is organized in the following way: in Section 2 a brief theoretical background is provided. We show that three metrics can be distilled from the WHAM equations–the confinement, overlap, and consistency–that quantify to which extent the US simulations are suitable to construct the free energy surface. We then discuss the methodology of the OGRe package, which naturally follows from these metrics. In Section 3, the OGRe approach is benchmarked on several examples to showcase its efficacy, by reproducing 1D and 2D analytic potentials. Finally, in Section 4, OGRe is applied to investigate the free energy landscape of two physical systems: proton hopping in zeolites, described by a 1D free energy profile, and layer stacking in a two-dimensional covalent organic framework, described by a 2D free energy surface. These examples are chosen to emphasize the generality of the approach, namely that the procedure is suited for processes both separated by non-negligible activation barriers and those characterized by shallow potential energy surfaces, thus being useful for the majority of realistic physical and chemical processes.

2 Methods

2.1 Theoretical background of umbrella sampling

Free energy calculation methods boil down to a partitioning of the phase space into macrostates–identified by a unique value of the set of collective variables (CVs), here denoted by the N-dimensional vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_N)$–and calculating the relative probabilities of these states being visited. This probability distribution function as a function of the CVs can, for instance, be calculated through US. Umbrella sampling facilitates an enhanced sampling by performing a series of simulations, where a bias potential or umbrella forces the sampling in each simulation to a specific region. In this work, each individual umbrella simulation i occurs on a so-called `Node`, which defines the center of the umbrella $\boldsymbol{\xi}_i$ in the N-dimensional

CV space, and the strength of the bias potential, defined by the N-dimensional vector $\boldsymbol{\kappa}_i$. In this notation, we allow for the bias strength to differ along each of the N CV directions. The complete collection of `Nodes` then defines the `Grid`, which represents the total sampling region of interest, illustrated in Figure 1c.

Sampling this `Grid` results in a series of biased free energy profiles–one for each `Node`–which can be converted to the unbiased free energy profile for each `Node` by subtracting the constant bias potential. Subsequently, these local unbiased free energy profiles can be combined to obtain the total unbiased free energy profile through, *e.g.*, the weighted histogram analysis method (WHAM).[21] WHAM constructs the total unbiased probability density distribution as a linear combination of the individual unbiased probability densities of every simulation, by choosing the weight coefficients such that the variance of the total unbiased probability density is minimized. Suppose that our `Grid` consists of M `Nodes`, and that the free energy surface is evaluated along B points in the CV space, referred to as histogram bins (typically, $B \gg M$). The resulting nonlinear system of equations can be formulated in several ways,[3,20,21,23,26] such as:[23]

$$\begin{cases} P_k &= \frac{\sum_{i=1}^{M} H_{ik}}{\sum_{i=1}^{M} N_i f_i b_{ik}} \quad k \in 1 \dots B \\ f_i^{-1} &= \sum_{k=1}^{B} b_{ik} P_k \quad i \in 1 \dots M \\ 1 &= \sum_{k=1}^{B} P_k \end{cases} \tag{1}$$

In this expression, P_k represents the unbiased probability density at the center of histogram bin k, H_{ik} is the histogram count–representing the biased probability density–of `Node` i for bin k, N_i is the number of samples of `Node` i, f_i is the normalization factor for the biased probability density of `Node` i (obtained from the estimated unbiased probability density), and b_{ik} is the (integrated) Boltzmann factor of the biased system for bin k:

$$b_{ik} = \frac{1}{\Delta_k} \int_{\text{bin}_k} e^{-\beta U_i^b(\boldsymbol{\xi})} \mathrm{d}\boldsymbol{\xi} \tag{2}$$

with Δ_k the volume of bin k, $\boldsymbol{\xi}$ the CV, U_i^b the bias potential for `Node` i, and the integral is calculated over the bin volume of bin k. This system of nonlinear equations is then solved iteratively in WHAM, until converged estimates of P_k and f_i are obtained starting from some initial values.

The downside of this technique is that, to obtain a sound unbiased probability density, the biased probability density for each `Node` must have a "sufficient" overlap with the biased probability densities of its nearest neighbors.[11] While this requirement seems intuitive and trivial to relate all the probability densities to each other and reduce statistical errors, it is not evident from the WHAM equations. Furthermore, determining whether the overlap is "sufficient" is typically only examined by comparing the combined trajectory data visually, or by considering an overall error measure for the final free energy differences, which may originate from other sources.[10]

To illustrate this, we consider an extreme yet illustrative example for the WHAM equations that makes the requirement of overlap and its influence on the probability density evident. Assume our `Grid` contains M `Nodes`, where the corresponding simulations all have equal length ($N_i \equiv N$), and that the histogram counts for every simulation i are limited to a single, different, bin k_i. Then, $H_{ik} = N\delta_{kk_i}$, with k_i representing the relevant bin for `Node` i, and two biased probability densities never overlap. This can, for instance, be imposed through subjecting every simulation to a strong enough bias potential, centered at the corresponding bin location. For simplicity, we assume that the bias potentials have equal strength, such that $b_{ik} = b\delta_{kk_i}$. Then, the WHAM equations (1) reduce to:

$$\begin{cases} P_k & = \frac{\sum_{i=1}^{M} N_s \delta_{kk_i}}{\sum_{i=1}^{M} N_s f_i b \delta_{kk_i}} \Rightarrow P_{k_i} = \frac{1}{f_i b} \\ f_i^{-1} & = b P_{k_i} \\ 1 & = \sum_k P_k \end{cases} \tag{3}$$

This is an underdetermined system of equations, as the first two series of equations are identical. This can no longer be solved iteratively, and an infinite number of solutions exist, resulting in an infinitely high condition number. In other words, small variations in, *e.g.*, the sampled probabilities lead to infinitely high variations in the free energy. As such, the lack of overlap between simulations on adjacent `Nodes` can be related to the condition number of this system of equations, and thus the accuracy of the final free energy surface.

2.2 Metrics to gauge the quality of US simulations

In accordance with the theoretical background, three robust metrics are employed in the OGRe refinement procedure to ensure an accurate solution of the WHAM equations: (i) the confinement of each `Node`, (ii) the overlap between each pair of neighboring `Nodes`, and (iii) the consistency between the WHAM FES and the sampled trajectories. Each of these three metrics is elaborated upon below.

i. Confinement To maximize the statistical accuracy, convergence, and sampling of the orthogonal degrees of freedom, each `Node`'s umbrella simulation should obtain a high biased Boltzmann probability around its umbrella center $\boldsymbol{\xi}_i$.[3] To this end, the deviation of a `Node`'s sampling from its umbrella center is quantified, and defined as the confinement c. This confinement is expressed as a number between 0 and 1, determined by the fraction of the `Node`'s samples that fall inside the `Node`'s direct environment V_{cell}: an N-dimensional hypercube bounded by the location of adjacent `Nodes`. This definition implies that direct environments

of adjacent `Nodes` partially overlap. If all samples fall inside this direct environment, the confinement takes on the value of one, which reveals that the trajectory is fully confined. This is illustrated in Figure 2, and can be formally expressed as:

$$c_i = \int_{V_{\mathrm{cell},i}} P_i^b(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi}, \qquad \text{with } 0 \leq c_i \leq 1 \tag{4}$$

with $\boldsymbol{\xi}$ the CV, $P_i^b(\boldsymbol{\xi})$ the biased probability distribution of the simulation on Node i, and $V_{\mathrm{cell},i}$ the corresponding `Node` cell volume.

ii. Overlap Only when two neighboring `Nodes` are confined, it is sensible to test for the overlap between their biased probability densities. Here, similar to the confinement c, the overlap o is expressed as a number between 0 and 1. It is a property assigned to a set of two `Nodes` and is determined by the integral over the whole CV space ($\boldsymbol{\Omega}$) of the point-wise minimum of the two probability densities of these two `Nodes`, where 1 corresponds to an identical overlap. This is illustrated in Figure 3, and can be formally expressed as:

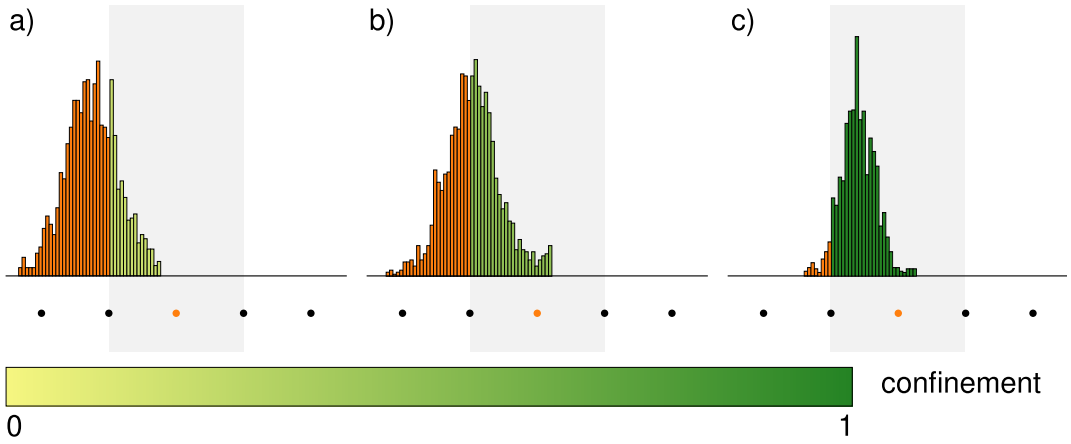

Figure 2: Illustration of the confinement metric in 1D. The panels represent the simulated trajectory of the orange-colored `Node`, resulting in a (a) low, (b) medium, and (c) high confinement value, indicated in correspondence with the color bar at the bottom. The gray box indicates the `Node`'s direct environment over which the integration of eq. 4 is performed.

$$o_{ij} = \int_{\mathbf{\Omega}} \min\left(P_i^b(\boldsymbol{\xi}), P_j^b(\boldsymbol{\xi})\right) \mathrm{d}\boldsymbol{\xi}, \qquad \text{with } 0 \leq o_{ij} \leq 1 \tag{5}$$

with $\boldsymbol{\xi}$ the CV, and $P_i^b(\boldsymbol{\xi})$ and $P_j^b(\boldsymbol{\xi})$ the biased probability distributions of the simulations on `Nodes` i and j. The overlap metric o is essential to ensure a low condition number for the coupled WHAM equations, such that the convergence of the estimated total unbiased probability density is contingent on the overlap between two or more simulated biased probability densities at every relevant `Node` location. In essence, trajectory overlap will ensure that neighboring windows have a similar sampling of the orthogonal degrees of freedom, resulting in consistent probability distribution estimations between them.[23]

iii. Consistency Even when all `Nodes` are considered confined and overlapping, the final unbiased probability resulting from the WHAM equations (eq. 1) can still deviate significantly from reality. Although the WHAM free energy profile is consistent in terms of the simultaneous convergence of the calculated unbiased probabilities P_k and the normalization

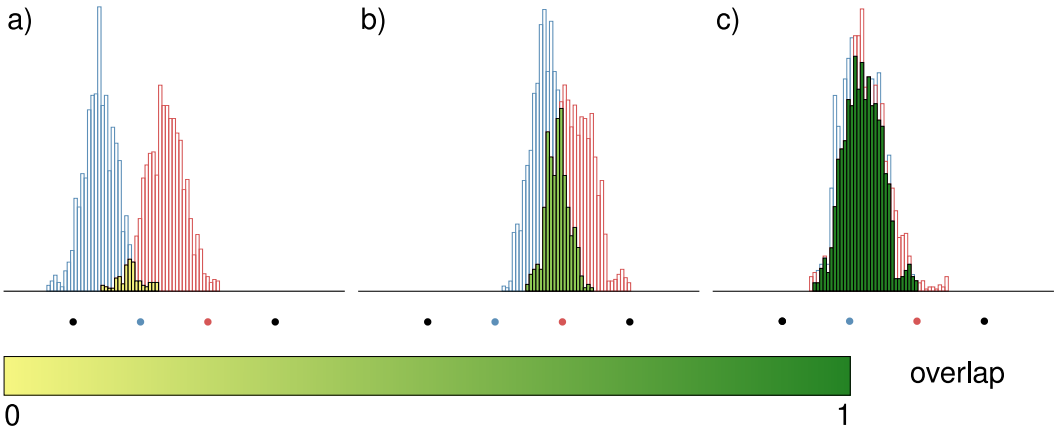

Figure 3: Illustration of the overlap metric in 1D. The panels represent the simulated trajectories of the blue- and red-colored `Nodes`, resulting in a (a) low, (b) medium, and (c) high overlap value. The density overlap is indicated in correspondence with the color bar at the bottom, whereas the individual density plots are indicated in blue and red outlines. The `Nodes` are displayed below each density plot.

factors f_i of the biased probabilities P_{ik}^b, this convergence does not guarantee an accurate free energy profile, as the individual US simulations do not guarantee an ergodic sampling. An example of this is shown in Figure 4. Starting from the predicted FES as calculated through WHAM in Figure 4a, a clear discrepancy can be observed with respect to the (unknown) underlying FES. As illustrated on the top of Figure 4c, this discrepancy can be traced back to a difference between the biased probability density predicted by WHAM and the actually sampled histogram for one or multiple `Nodes`. For instance, consider the US simulation around the orange-colored `Node` in Figure 4c. For this `Node`, the expected biased probability density can be calculated by incrementing the WHAM-predicted FES by the bias potential for that `Node`. In this case, this gives rise to a bimodal biased free energy profile and the corresponding histogram outlined in orange. However, given the free energy

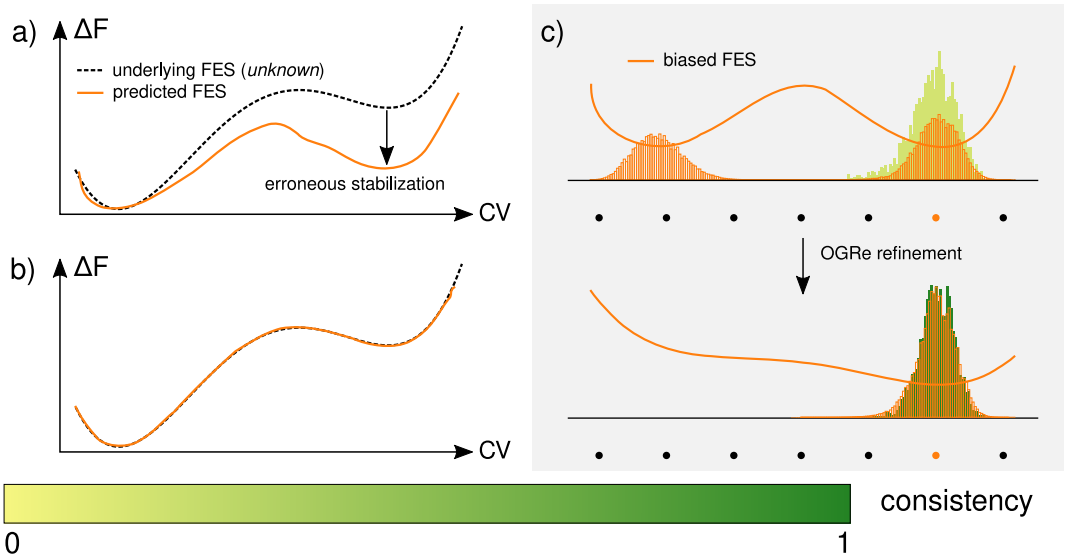

Figure 4: Illustration of the consistency metric in 1D. On the left, the panels represent the predicted and the underlying free energy profile in the case of (a) a single inconsistent orange-colored `Node` (low consistency), and (b) the refined counterpart (high consistency). On the right, the origin of the mismatch between the free energy profiles in (a) is found in the inconsistency between the sampled (in yellow-green) and the predicted biased probability density (in orange) of the responsible `Node`. The biased predicted density is calculated through eq. 9, using the predicted FES and the bias potential for that specific `Node`. The color of the sampled density is in correspondence to the degree of consistency using the color bar on the bottom.

barrier between the two minima, even in the biased free energy surface, the actual sampling histogram (in yellow) of the orange `Node` may be located in a single minimum instead, resulting in non-ergodicity. This would erroneously stabilize the sampled minimum compared to the unsampled one, which in turn distorts the whole probability distribution and the unbiased free energy surface predicted by WHAM. Hence, while these simulations may be both confined and overlapping, and while the WHAM equations converge, the actual sampled histograms will be inconsistent with the histogram predicted from the final WHAM equations. This ergodicity issue can be solved by simply increasing the umbrella bias strength $\boldsymbol{\kappa}$ for this `Node` (further confining the simulation, eliminating other free energy minima in the biased simulation), or further enhancing the sampling through, *e.g.*, replica exchange methods. However, identifying whether the issue has occurred, and for which simulation, is more difficult.

Starting from the WHAM equations (eq. 1), the biased probabilities can be defined as:

$$P_{ik}^{b} = \frac{1}{\sum_{l=1}^{B} b_{il}\ P_l} \cdot b_{ik} P_k = f_i b_{ik} P_k \tag{6}$$

Additionally, the biased probabilities can also be directly derived from the sampled histograms:

$$\widetilde{P}_{ik}^{b} = H_{ik}/N_i \tag{7}$$

with $N_i = \sum_l H_{il}$ the total amount of samples in the US simulation on `Node` i. As such, by comparing the sampled biased probabilities $\widetilde{P}_{ik}^{b}$ to the calculated biased probabilities P_{ik}^{b} for each `Node`, we can identify consistent `Nodes` through, for instance, the following metric:

$$s_i = 1 - \mathrm{JSD}(\widetilde{P}_i^b \parallel P_i^b), \quad \text{with } 0 \leq s_i \leq 1 \tag{8}$$

where JSD represent the Jensen-Shannon divergence.[27] The Jensen-Shannon divergence is a measure for the similarity between two probability distributions with finite bounds, where a value of 0 represents identical distributions and a value of 1 (when using a base 2 logarithm) represents the maximum divergence. As such, this consistency metric s has a value of 0 for maximally inconsistent `Nodes`, and a value of 1 for perfectly consistent `Nodes`. Consequently, s can be treated similarly to the previous metrics, where all `Nodes` that have a sufficiently high s value can be considered consistent. As both the confinement and consistency metrics relate to the umbrella strength of each `Node`, and is defined by the properties of each individual `Node` alone, a `Node` will only be considered 'reliable' when it is both confined and consistent. As discussed above, the overlap depends on two `Nodes` simultaneously instead.

Importantly, this consistency check can be added to any WHAM code, and can be calculated even when the WHAM equations did not converge to facilitate error correction. In case when only the resulting WHAM free energy is available, for instance during the OGRe post-processing steps, the biased probabilities of eq. 6 can also be derived through:

$$P_{ik}^b = \frac{1}{C_i} \exp\left[-\beta\left(F_k + U_{ik}^b\right)\right] \tag{9}$$

with C_i $(= \sum_k P_{ik}^b)$ a normalization constant, F_k the free energy predicted by WHAM evaluated in bin location k, and U_{ik}^b the bias potential for `Node` i at bin location k.

2.3 OGRe package

The OGRe package is an open-source Python package intended for optimal grid refinement in an arbitrary number of dimensions. It consists of two independent modules (`ogre.input` and `ogre.post`), corresponding to the input generation and the refinement module, respectively. This allows the user to choose their own simulation engine, which then interfaces with the `ogre.input` and `ogre.post` modules as elaborated in Section S1. A complementary package (OGRe-Yaff) is also available, which includes a simulation module (`ogre.sim`) that directly couples to our in-house simulation engine, Yaff,[28] making the package completely self-contained. This is schematically illustrated in Figure 5, and its practical usage is documented in Section S1.4.

The central object in OGRe is the `Grid` object, defining the collection of umbrella bias strengths and centers that collectively allow for an optimal derivation of the FES, individually referred to as `Nodes`. A schematic representation of this object is provided in Figure 5. This

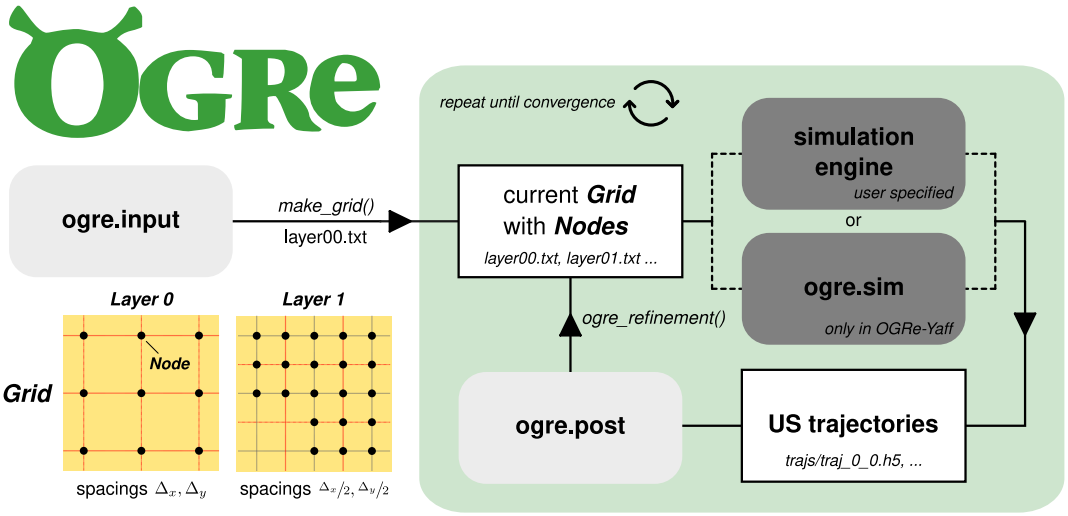

Figure 5: Schematic overview of the OGRe protocol, and the `Grid`, `Layer`, and `Node` objects. The OGRe protocol consists of two independent module (`ogre.input` and `ogre.post`), complemented by a simulation code or the complementary `ogre.sim` module from the OGRe-Yaff package. The `Grid` is the central object of the OGRe protocol, divided into `Layers`, which, in turn, are populated by equally spaced `Nodes`.

figure shows that within each `Grid`, sets of equidistant `Nodes` are grouped into `Layers`. These `Layers` relate to the increased spatial refinement of the `Grid` throughout the OGRe iterations, where each `Layer` is characterized by a fixed spacing $\boldsymbol{\Delta}_l$ (in each dimension) between the `Nodes`. In this way, the concept of neighboring `Nodes`, required for the overlap metric, can be consistently defined as those `Nodes` within the same `Layer` at the fixed spacing distance $\boldsymbol{\Delta}_l$ from each other.

2.3.1 Input generation

The OGRe procedure schematically illustrated in Figure 5 starts at `ogre.input`, where the first step towards the optimal grid refinement is to define the initial `Layer` and the (hyper)parameters for the refinement procedure. This definition is facilitated by an instance of the `OGRe_Input` class, as outlined in Figures S1-S2, and the execution of its `make_grid()` method. This writes the initial `Grid` information to a `layer00.txt` file, with the identity, umbrella center, umbrella strength, and `Node` type (*vide infra*) for each `Node`. This identity consists of the layer number (starting at zero) and the `Node` number (starting at zero), corresponding to an enumeration of the `Nodes` in that `Layer`. Additionally, all relevant information of the input object is saved to a human-readable YAML file, `data.yml`, which is later parsed for the output generation, and allows for easy access and adaptability.

The full `layer00.txt` information is then copied to the `run.txt` file, representing all remaining simulations to be performed, following the latest refinement. After every post-processing step, this will be updated along with all relevant grid files (`layer00.txt`, `layer01.txt` ...), until the protocol converges and no new simulations are added to the `run.txt` file. An extended discussion of the `OGRe_Input` class, the complementary `OGRe_Simulation` module, and post-processing module can be found in Section S1.

2.3.2 Post-processing

After the relevant simulations have been performed, the post-processing module is used to refine the existing `Layers` through the `ogre_refinement()` function, using the confinement, overlap, and consistency metrics defined above. By iterating over all `Layers`, every `Node` is first subjected individually to the confinement and consistency tests, which consider the deviation of its trajectory from the umbrella center and the deviation from the expected probability density, respectively. This identifies the reliable `Nodes`. Second, all reliable `Nodes` within the same `Layer` are pairwise compared with all neighboring `Nodes` through the overlap metric. While the former validates the individual umbrella strength of each `Node`, the latter verifies whether or not the simulated probability densities are sufficiently overlapping between each pair of two `Nodes` so to obtain accurate free energy differences from the WHAM equations. In turn, through identification of reliable `Nodes`, and overlapping pairs of `Nodes`, the `Nodes` within each `Layer` can be refined by increasing the umbrella strength for unreliable trajectories, and introducing new `Nodes` between reliable but non-overlapping pairs of `Nodes` in a subsequent `Layer`. As such, an iterative procedure emerges, as illustrated in the ogre-colored pane of Figure 5, where trajectories are collected based on the current `Grid`–through either a user-specified simulation engine or `ogre.sim`–which are then used to update the current `Grid`. This process is subsequently repeated, until the metric tests consider all `Nodes` to be reliable and overlapping, such that the grid refinement has converged.

Metric tests The reliability tests (confinement and consistency) are controlled by separate user-specified hyperparameters called `CONFINEMENT_THR` and `CONSISTENCY_THR` (see Section S1 for a full list), and are executed sequentially, as the consistency for a non-confined `Node` is of no interest. If either the confinement or consistency of a `Node` falls below its respective threshold, the `Node` is no longer considered reliable, and will be refined by increasing the umbrella strength ($\boldsymbol{\kappa}_i$) at that position. Each iteration will result in an increase of the original

$\boldsymbol{\kappa}_i$ value by a user-specified factor `KAPPA_GROWTH_FACTOR` to allow for maximal control. In this way, the confinement test first increases the $\boldsymbol{\kappa}_i$ value until the bias potential optimally compensates for the free energy gradient at that `Node` location, increasing the accuracy of the free energy estimate at the cost of a reduced width for the biased Boltzmann probability. Complementary, the consistency test enforces that only a single minimum is considered for each US simulation, preventing erroneous stabilization of free energy minima.

When $\boldsymbol{\kappa}_i$ grows too large, it becomes increasingly difficult for neighboring `Nodes` to overlap. Moreover, high umbrella strengths may lead to an over-representation of high-energy configurations, increased correlation between samples, and unstable simulations due to the finite time step in the MD algorithm, as the simulation can no longer emulate the frequency of the bias potential.[26,29] To accommodate for this, an additional hyperparameter `MAX_KAPPA` can be specified, which is the largest allowed value for the umbrella strength. Should the protocol attempt to increase the $\boldsymbol{\kappa}_i$ value above this limit, the `Node` is no longer refined (precluding a reliable state), and, as such, no longer takes part in overlap tests. This limits denser sampling in regions with prohibitively large unbiased free energy gradients, and decreases the numerical errors in the WHAM equations.[3] When this results in disjoint collections of overlapping regions, it is expected that the resulting free energy profile will not succeed in adequately reproducing the free energy difference between the regions, and either the `MAX_KAPPA` should be increased, or the disjoint regions should be considered separately as any transitions between them are deemed improbable.

The overlap test is similarly controlled by a single user-specified hyperparameter called `OVERLAP_THR`. If the overlap falls below this threshold, the neighboring `Nodes` are considered to be non-overlapping, and the `Grid` is locally densified, by introducing new `Nodes`, to amend this. These new `Nodes` are part of the subsequent `Layer`, as introduced in Figure 5, with a smaller fixed `Node` spacing $\boldsymbol{\Delta}_{l+1}$. The details of the `Node` generation are given below in the subsection "Grid refinement". In this way, new `Nodes` will account for undersampled regions,

while ensuring overlap with the surrounding `Nodes` for optimal convergence conditions of the unbiased probability density and the corresponding free energy estimate.

Similar to `MAX_KAPPA` for trajectory confinement, `MAX_LAYERS` can be introduced to limit the minimal $\boldsymbol{\Delta}_l$ value by limiting the number of `Layers` that may be generated. As an absolute maximum number of `Layers`, the smallest $\boldsymbol{\Delta}_l$ value should be larger than half the user-specified parameter `HISTOGRAM_BIN_WIDTHS`, which defines the bin widths in each dimension for calculating the trajectory overlap, as the overlap metric o will be identically 1 beyond that point. Moreover, ever-decreasing $\boldsymbol{\Delta}_l$ values will shift the balance between the equilibration time (of the orthogonal degrees of freedom) and the characteristic sampling time of the grid cell volume. As the $\boldsymbol{\Delta}_l$ value grows smaller, the time necessary to sample the ever smaller grid cell volume decreases, whereas the equilibration time remains unchanged, significantly lowering the efficiency of the US simulations.[19]

Notably, the predicted free energy profile changes during `Grid` refinement, possibly altering the value of the consistency metric for each `Node`. Ultimately, this may result in reliable `Nodes` reverting back to unreliable `Nodes` for a given `Layer` l–with reliable `Nodes` in all previous layers by construction. At that point, OGRe halts any refinement of the derived `Layers` (`Layers` with a higher layer number), and attempts to correct the unreliable `Nodes` in `Layer` l by increasing the $\boldsymbol{\kappa}_i$ value as outlined above. Once all `Nodes` in `Layer` l are reliable again–or the `MAX_KAPPA` is reached–OGRe resumes the refinement for the derived `Layers` $l+1, l+2, \ldots$

Grid refinement When two reliable neighboring trajectories fail to overlap, as defined above, a series of new `Nodes` is generated in the next `Layer` $l+1$ in a way to optimally explore the undersampled gap between them. This process of `Node` generation is illustrated in Figure 6 for one- and two-dimensional `Grids`, and can be generically extended to an arbitrary number of dimensions. In essence, a new `Node` is added in the middle of the two

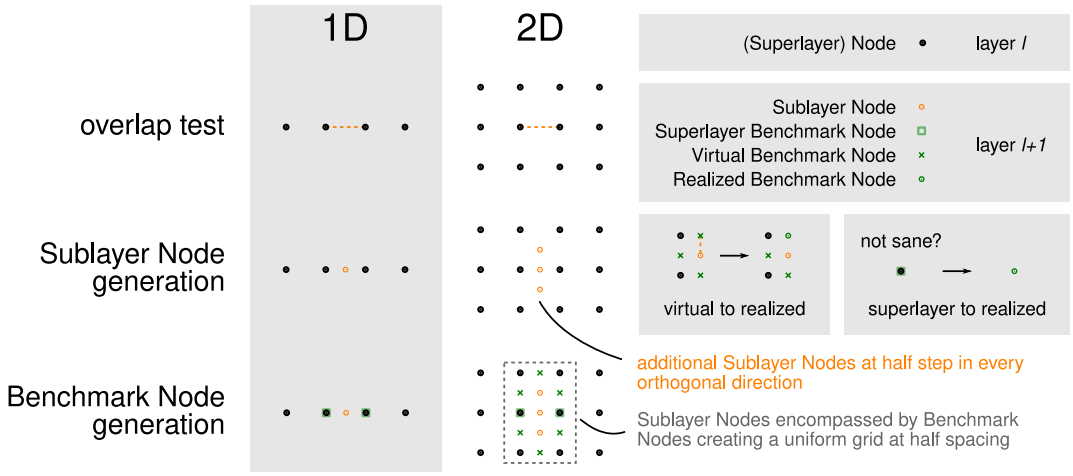

Figure 6: Illustration of the grid refinement in 1D and 2D. The procedure is illustrated by considering two Nodes with an overlap below the specified `OVERLAP_THR` in `Layer` l, indicated by the dashed orange line. When an overlap test fails, `Sublayer Nodes` (indicated in orange) and the `Superlayer` and `Virtual Benchmark Nodes` necessary to check overlap (indicated in green) are generated, so to create a uniform grid in `Layer` $l+1$ where $\boldsymbol{\Delta}_{l+1} = \boldsymbol{\Delta}_l/2$. If these `Superlayer` or `Virtual Benchmark Nodes` fail the confinement or consistency metrics, it is not marked as reliable and replaced by a Realized Benchmark Node.

non-overlapping nodes, with the addition of `Nodes` at every half spacing from that middle `Node` in every other dimension ($\boldsymbol{\Delta}_{l+1} = \boldsymbol{\Delta}_l/2$). In this way, a uniform subsampling is performed at half the original $\boldsymbol{\Delta}_l$ values around the newly introduced `Node`. This process finally results in a collection of `Layers`, each characterized by a unique $\boldsymbol{\Delta}_l$, which decreases exponentially for each dimension as the $\boldsymbol{\Delta}_l$ values halve per `Layer`.

In Figure 6, aside from the generation of new `Sublayer Nodes` in `Layer` $l+1$ in between the two non-overlapping `Nodes` of `Layer` l, and in every orthogonal direction (from 2D onward), `Benchmark Nodes` are introduced in `Layer` $l+1$. These `Benchmark Nodes` ensure that every new `Node` in `Layer` $l+1$ is fully encompassed by neighboring `Nodes` for the overlap test. In the one-dimensional case, these `Benchmark Nodes` identically correspond to `Nodes` from `Layer` l, effectively coupling subsequent `Layers`. As such, they are referred to as `Superlayer Benchmark Nodes` of `Layer` $l+1$, and are simply references to `Nodes` from `Layer` l. However, in more than one dimension, the perimeter of each connected region of

new `Nodes` lacks `Superlayer Benchmark Nodes`, as evident from the 2D case in Figure 6. To accommodate for this, `Virtual Benchmark Nodes` are introduced at the relevant locations in `Layer` $l+1$, indicated with green crosses. They use the combined trajectory data of their two neighboring (`Benchmark`) `Nodes` from `Layer` l. Importantly, `Virtual Benchmark Nodes` are only added to `Layer` $l+1$ when both its neighbors in `Layer` l are reliable. Otherwise, the `Benchmark Node` is omitted from the `Grid`, and no overlap test can take place, precluding spatial refinement in that local environment.

Evidently, the `Benchmark Nodes` are also subject to refinement. First, as with `Nodes`, `Superlayer Benchmark Nodes` are subjected to a confinement test in `Layer` $l+1$, using the same trajectory as the original `Node` from `Layer` l, but with the smaller grid cell volume of `Layer` $l+1$ (the consistency is unchanged). If this test fails, it is replaced by a `Realized Benchmark Node`, which is then refined as any `Node` would be. At this point, the trajectory of this `Node` in `Layer` $l+1$ is different from the one in `Layer` l. Contrastingly, a `Virtual Benchmark Nodes` is considered reliable by default, as it originates from two reliable and overlapping `Nodes` in `Layer` l. Second, since `Benchmark Nodes` are not necessarily fully encompassed by other `Nodes`, and are only intended as a benchmark, they are not subjected to an overlap test in a direct way. Instead, when iterating over all non-`Benchmark Nodes`, and performing the overlap test with respect to all their neighbors, every `Benchmark Node` will be taken into account by construction. If the overlap test fails for a `Virtual Benchmark Node`, it is replaced by a `Realized Benchmark Node` using the largest $\boldsymbol{\kappa}_i$ values of its neighboring `Nodes`, for which new simulations are started. In contrast, a failed overlap test for any other `Benchmark Node` is dealt with as before.

2.4 Limitations to OGRe

The maximum accuracy attainable through OGRe is bound by the choice of CVs and the convergence of the individual trajectories. To this end, there exist a multitude of techniques to select appropriate CVs, such as the time-structure based independent component analysis method,[30] as well as techniques to improve the convergence of the individual simulations, such as replica-exchange methods.[31,32] Essentially, a converged trajectory cannot be qualitatively distinguished from an infinitely long simulation. While the reliability metrics (confinement and consistency) do enforce convergence to some degree, by limiting the phase space with increased umbrella strengths, they do not provide a quantitative measure for convergence. An optional metric that is implemented in the OGRe code to check for convergence issues considers the similarity of the probability density between the first and second half of the trajectory. Similar to the consistency metric, the similarity is calculated though the Jensen-Shannon divergence, and a trajectory is considered converged when $1 - \mathrm{JSD}(P_{\mathrm{first}}|P_{\mathrm{second}})$ is larger than the provided `CONVERGENCE_THR`. Low convergence values typically occur when dealing with multiple minima, such that similar to the consistency metric, the simulation is repeated with a higher $\boldsymbol{\kappa}_i$ value to aid with convergence. While these convergence issues can be partially lifted in higher dimensions, as the phase space can be connected through infinitely more paths, the complexity of the phase space and the number of minima can increase as well (*vide infra*), such that the consistency and convergence thresholds can still be relevant.

Additionally, the OGRe protocol requires the umbrella potential at `Node` i to be defined by an umbrella strength parameter $\boldsymbol{\kappa}_i$, and an umbrella center parameter $\boldsymbol{\xi}_i$, for it to be refined by the post-processing protocol in its current version. In this work, the default harmonic potential $\kappa_{ij}(\xi_{ij} - \xi_j)^2/2$ is used for the j^{th} dimension in simulation i. Moreover, while OGRe aims to predict the same final free energy surface irrespective of the initial US parameters, the efficiency of the algorithm can be significantly impacted by this initial parameter choice.

Although the described refinement algorithms succeed in converging to consistent values for the `Grid`, a bad choice will evidently lead to a much larger number of iterations, and thus simulations. As such, its efficiency is, to some extent, reliant on the user. This will be further illustrated in Section 3 by performing a Pareto analysis on the error as a function of the number of simulations, and determining the Pareto optimal set, that balances the need for a low free energy error with the need to limit the amount of simulations.

Finally, the `HISTOGRAM_BIN_WIDTHS` parameter determines the spatial resolution of all histograms (and free energy profiles) during the OGRe run. As such, varying the bin width affects the overlap, consistency, and convergence checks as they are histogram-based, and can significantly affect the free energy profile. A large bin width will limit the maximal gradient in free energy, which may artificially lower barrier heights and free energy differences, whereas a small bin width results in noise, as the number of samples per bin decreases (see also Section S3). Ideally, the bin width should be chosen large enough such that each bin is still significantly sampled, but low enough to avoid cases where the data of a single trajectory is fully contained in a single bin. In the benchmarking simulations discussed below, the bin width is chosen as small as possible, to increase the `MAX_KAPPA` value (see eq. S3.3) at the cost of more noise on the FES data. Overall, this results in a lower error, especially for large `CONFINEMENT_THR`, which naturally require larger $\boldsymbol{\kappa}_i$ values.

3 Benchmarking

The optimal values of the different hyperparameters are inherently correlated and depend on the initial `Grid` and simulation parameters. However, a set of reliable hyperparameters for accurate free energies is a requirement for ease-of-use and broad applicability of the proposed package. To this end, a benchmarking study is performed on several 1D and 2D analytic potentials, illustrated in Figure 7. These potentials facilitate a fast parameter screening

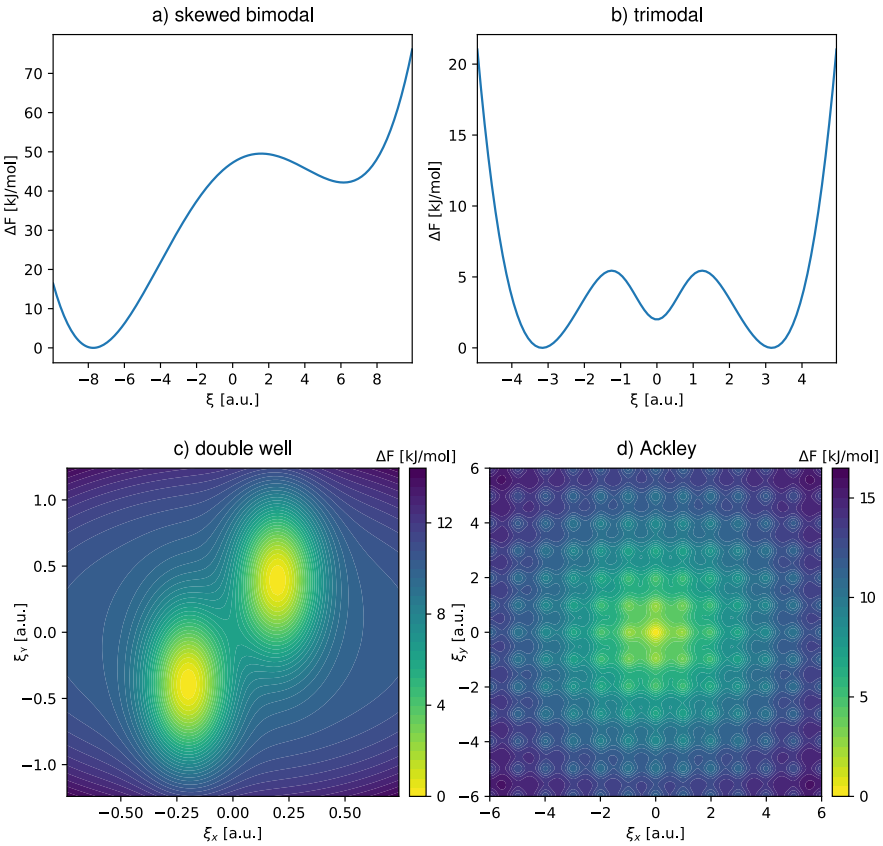

Figure 7: Illustration of the four 1D and 2D analytical benchmark potentials. The benchmark potentials constitute two 1D potentials: (a) the skewed bimodal, and (b) the trimodal potentials; and two 2D potentials: (c) the double well, and (d) the Ackley potentials.

and direct error measure through comparison of the predicted FES with the analytical FES used as input, as discussed in Section S2. The potentials were selected to showcase difficult to capture FES features, such as flat energy profiles, equivalent minima, and high barriers, typically used when testing optimization algorithms,[33] and which are prone to consistency errors.

For these analytic potentials, the error can easily be defined as the root mean squared error (RMSE) between the predicted and the analytical free energy differences, summed over al bins, relative to their respective free energy minimum ($\Delta F_k = F_k - \min(F)$):

$$\text{RMSE} = \sqrt{\frac{1}{B}\sum_{k=1}^{B}\left(\Delta F_k^{\text{calc}} - \Delta F_k^{\text{analytic}}\right)^2} \tag{10}$$

1D: skewed bimodal As an initial test, the OGRe protocol is applied to reproduce the following analytical potential:

$$U^b(\xi) = 3\xi - \xi^2 + 0.01\xi^4 \quad \text{with} \quad -10 \leq \xi \leq 10 \tag{11}$$

This potential, illustrated in Figure 7a, represents a bimodal free energy profile–typically found for activated processes–with a significant barrier height and difference in the free energy of the minima ($k_BT \approx 2.5$ kJ/mol at 300 K). In what follows, we will investigate how different OGRe input parameters (initial $\boldsymbol{\kappa}_i$ value and initial $\boldsymbol{\Delta}_l$ value) and hyperparameters (`CONFINEMENT_THR`, `OVERLAP_THR`, `MAX_LAYERS`, and `KAPPA_GROWTH_FACTOR`) affect the final error on the free energy profile, and the number of US simulations. This investigation is performed through consideration of a large set of parameter combinations, as presented in Figure S4. As such, we highlight which parameter combinations should be used, and which parameter values should be avoided entirely.

Below, in Figures 8-10, specific combinations are compared to explore the effect of different initial grid parameters and hyperparameters on the FES error. Then, the convergence of the OGRe protocol is discussed, which relates to the different hatches that overlay the error data. Finally, the efficiency is discussed based on the Pareto plot in Figure 11. In all simulations, the `MAX_KAPPA` hyperparameter is fixed in correspondence with the Nyquist-Shannon sampling theorem,[34] to avoid unstable simulations, while also accounting for the finite `HISTOGRAM_BIN_WIDTHS` as elaborated in Section S3. Additionally, fixed `CONSISTENCY_THR` and `CONVERGENCE_THR` are chosen to enforce convergence of each `Node` in the refined `Grid`. Their values are chosen high enough to avoid multiple minima in the biased free energy profile,

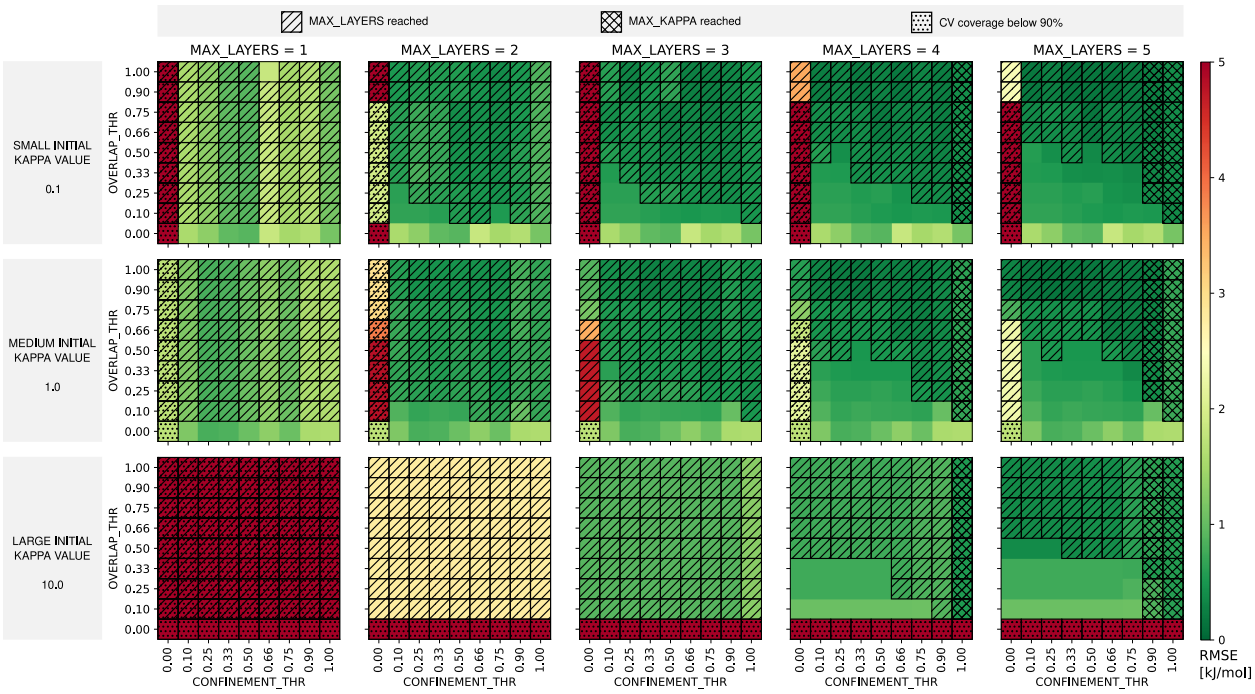

Figure 8: Effect of the initial $\boldsymbol{\kappa}_i$ value on the efficacy of the OGRe protocol using the skewed bimodal potential. The error (in kJ/mol) with respect to the analytical FES for varying initial $\boldsymbol{\kappa}_i$ values as a function of the `CONFINEMENT_THR` and `OVERLAP_THR`, for an increasing `MAX_LAYERS` limit. A fixed initial $\boldsymbol{\Delta}_l$ of 2.0 and `KAPPA_GROWTH_FACTOR` of 2 is employed for all OGRe runs in this figure. The definition of the possible hatches is given at the top.

mainly impacting the $\boldsymbol{\kappa}_i$ values of `Nodes` around the secondary minimum of the skewed bimodal potential. The numerical values of all fixed hyperparameters and simulation parameters are reported in Tables S4-S5.

Generally, the error decreases with increasing `CONFINEMENT_THR` and `OVERLAP_THR`, resulting from increased sampling at the steepest free energy gradients and increased overlap between the biased probability densities. This is for instance clear from Figure 8, with an increasingly green hue towards the upper right corner. Additionally, this figure illustrates the effect of increasing the initial $\boldsymbol{\kappa}_i$ value for a fixed initial $\boldsymbol{\Delta}_l$ of 2.0 and `KAPPA_GROWTH_FACTOR` of 2. A clear difference can be observed between small and large initial $\boldsymbol{\kappa}_i$ values. In the latter case, several spatial refinements are necessary (*i.e.*, several layers) before overlap occurs and the

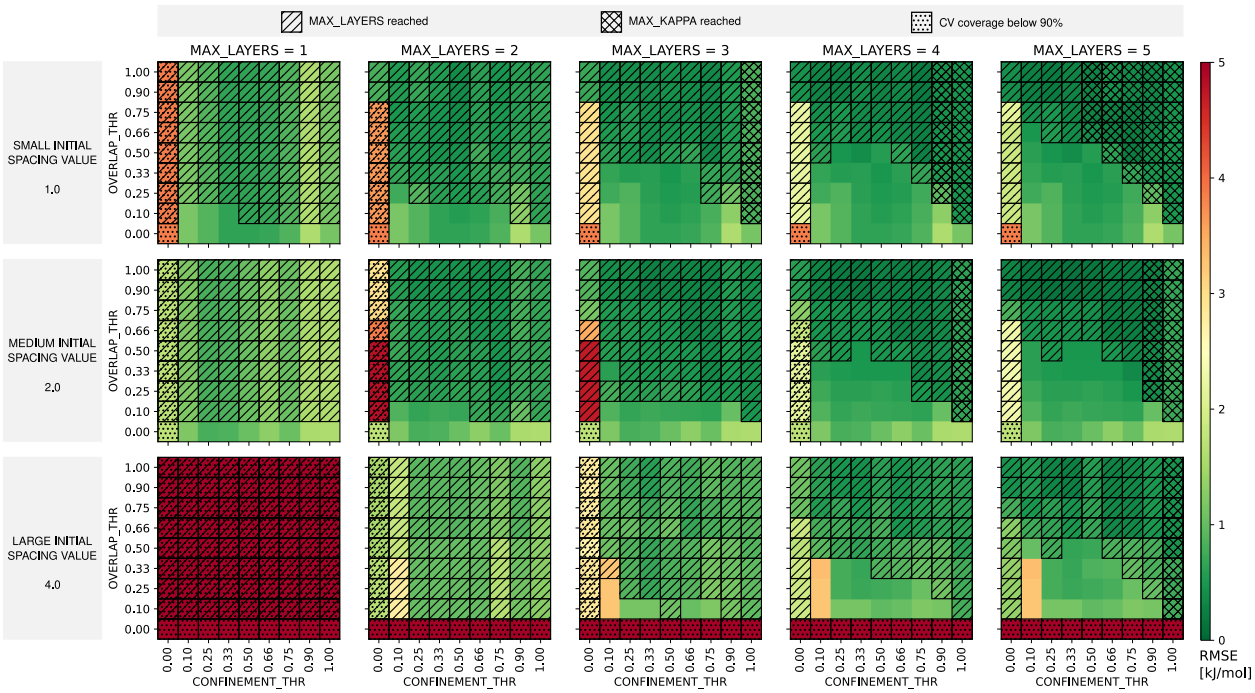

Figure 9: Effect of the initial $\boldsymbol{\Delta}_l$ value on the efficacy of the OGRe protocol using the skewed bimodal potential. The error (in kJ/mol) with respect to the analytical FES for varying initial $\boldsymbol{\Delta}_l$ values as a function of the `CONFINEMENT_THR` and `OVERLAP_THR`, for an increasing `MAX_LAYERS` limit. A fixed initial $\boldsymbol{\kappa}_i$ of 1.0 and `KAPPA_GROWTH_FACTOR` of 2 is employed for all OGRe runs in this figure. The definition of the possible hatches is given at the top.

free energy profile can be accurately reproduced, which is apparent from the completely red and yellow hues in the first and second layer for large initial $\boldsymbol{\kappa}_i$ values. In contrast, for very small initial $\boldsymbol{\kappa}_i$, the protocol can refine the $\boldsymbol{\kappa}_i$ values at each `Node` towards its optimal value, resulting in a low RMSE, at the cost of additional simulations. A very similar picture can be observed in Figure 9, which illustrates the effect of different initial $\boldsymbol{\Delta}_l$ values. With a fixed initial $\boldsymbol{\kappa}_i$ of 1.0 and `KAPPA_GROWTH_FACTOR` of 2, a large initial $\boldsymbol{\Delta}_l$ requires several spatial refinements before overlap can occur, in contrast to the smaller initial $\boldsymbol{\Delta}_l$ value, where some overlap is immediately possible.

Finally, the effect of the `KAPPA_GROWTH_FACTOR` is examined in Figure 10, with a fixed initial $\boldsymbol{\kappa}_i$ of 1.0 and $\boldsymbol{\Delta}_l$ of 2.0. The `KAPPA_GROWTH_FACTOR` mainly determines the trade-off

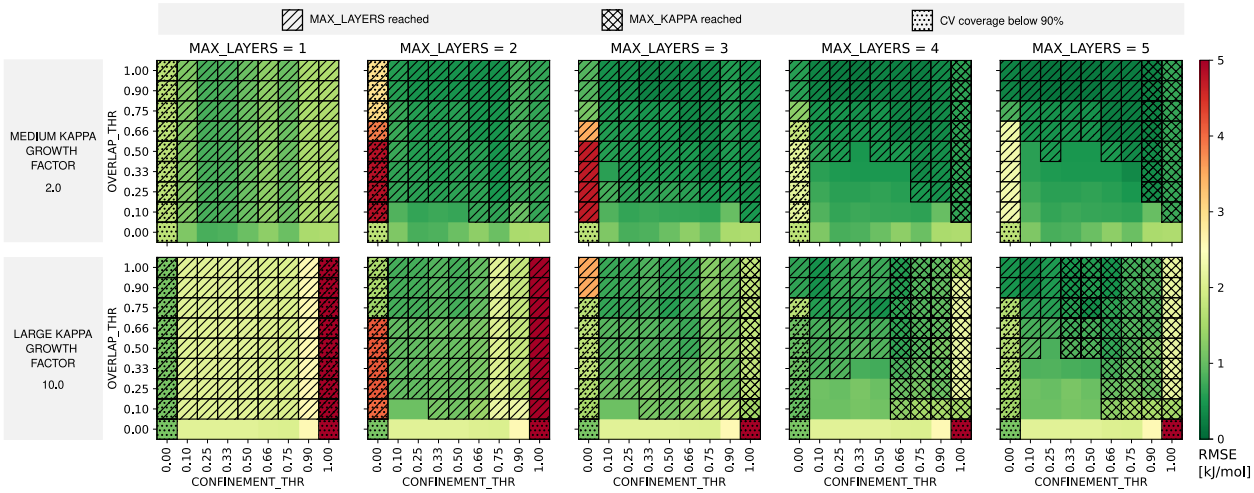

Figure 10: Effect of the `KAPPA_GROWTH_FACTOR` value on the efficacy of the OGRe protocol using the skewed bimodal potential. The error (in kJ/mol) with respect to the analytical FES for varying initial $\boldsymbol{\Delta}_l$ values as a function of the `CONFINEMENT_THR` and `OVERLAP_THR`, for an increasing `MAX_LAYERS` limit. A fixed initial $\boldsymbol{\kappa}_i$ of 1.0 and $\boldsymbol{\Delta}_l$ of 2.0 is employed for all OGRe runs in this figure. The definition of the possible hatches is given at the top.

between convergence speed and accuracy of the refinement procedure, affecting the FES error in a more indirect way. This follows from the fact that low multiplication factors for the refinement procedure allow for extremely well refined $\boldsymbol{\kappa}_i$ values to accommodate the free energy gradient, at the cost of many refinements, whereas a large `KAPPA_GROWTH_FACTOR` will quickly increase the $\boldsymbol{\kappa}_i$ value to be high enough. However, when increasing the $\boldsymbol{\kappa}_i$ value too fast, the change in confinement might be too drastic, requiring many spatial refinements before overlap can occur.

The convergence behavior in Figures 8-10 is determined by the finite `MAX_LAYERS` and `MAX_KAPPA`. When an OGRe run attempts to create a new `Layer` with an increased density beyond `MAX_LAYERS`, the protocol is not converged, indicated by a diagonal hatch. This is especially encountered when choosing large initial $\boldsymbol{\kappa}$ values, large initial Δ_l values, and/or large `OVERLAP_THR` values. Similarly, when the protocol attempts to increase the $\boldsymbol{\kappa}_i$ value for a single `Node` beyond `MAX_KAPPA`, the protocol is not converged, indicated by crossed hatches when at least one `Node` reaches this limit. This is especially encountered for large `CONFINEMENT_THR`

values, if the MAX_LAYERS hyperparameter allows it. Additionally, as the error can only be determined for those points of the calculated FES with a finite energy (non-zero probability), a dotted hatch indicates those OGRe runs where the FES could not be determined for at least 90% of the desired CV range. As the error could be artificially high or low, depending on which (unconnected) regions are covered, these parameter combinations should also not be taken into account. Low coverage is typically caused by a severe lack of overlap or confinement in certain CV regions, and typically occurs at the lowest CONFINEMENT_THR or OVERLAP_THR in Figures 8-10.

Several general trends in the convergence behavior can be observed. First, the number of converged OGRe runs increases as a function of the number of layers, creating a lower triangular section of converged simulations in the error matrix. This follows from the fact that an increased CONFINEMENT_THR results in a reduced potential for overlap, as already discussed in Section 2. However, deviations from the lower triangular shape can be expected due to the imposed CONSISTENCY_THR, which can supersede the required CONFINEMENT_-THR. Second, a zero OVERLAP_THR will always lead to convergence (not taking MAX_KAPPA into account), as no new Nodes are ever required and the existing Nodes will be refined towards the required CONFINEMENT_THR. Third, when only a single layer is allowed, the OVERLAP_THR has no effect, and all parameter combinations in any given column have the same value. Fourth, a zero CONFINEMENT_THR precludes any $\boldsymbol{\kappa}_i$ refinement, such that some overlap thresholds can never be attained, as the umbrella potential never dominates certain free energy gradients of the underlying profile. Finally, as mentioned above, for those runs with a limited CONFINEMENT_THR, a limited OVERLAP_THR, or both, the free energy profile may not be adequately defined within the required CV range.

Looking at the variation as a function of the initial $\boldsymbol{\kappa}_i$ and $\boldsymbol{\Delta}_l$ value, it is clear more OGRe runs converge for a given MAX_LAYERS for the lowest initial values, as there is a higher potential for overlap through either more closely spaced Nodes (initial $\boldsymbol{\Delta}_l$) or broader umbrellas

(initial $\boldsymbol{\kappa}_i$). However, as the $\boldsymbol{\Delta}_l$ values grow smaller, it becomes more difficult to attain confinement within the `Node` cell volume, resulting in an increase in κ value until the `MAX_KAPPA` value is achieved (indicated by the crossed hatches). Finally, larger `KAPPA_GROWTH_FACTOR` values decrease the number of OGRe runs that are converged in terms of both the `MAX_KAPPA` (diagonal hatches) and the lack of overlap for a given `MAX_LAYERS` (crossed hatches).

As evidenced by this proof of concept, our OGRe protocol can accommodate for a large range of initial parameters, determining the 1D free energy profile well beyond chemical accuracy (4 kJ/mol). However, up to this point, we have not taken efficiency into account. To increase the efficiency, the user could adapt its guesses for the initial parameters based on prior knowledge of the profile, or after the first grid run. When all simulations fail either on the confinement, consistency, or, optionally, convergence metric, it is recommended to increase the initial $\boldsymbol{\kappa}_i$ value. In contrast, when all simulations fail their overlap test, it is recommended to decrease the initial $\boldsymbol{\kappa}_i$ values or the initial $\boldsymbol{\Delta}_l$ value. In general, a lower initial $\boldsymbol{\kappa}_i$ value will always result in a better reproduction of the profile at the expense of more simulations required.

Based on all the converged OGRe runs, their error, and the corresponding number of simulations (Figure S4), a Pareto analysis can be performed to find the optimal hyperparameter combinations that simultaneously minimize the error and the required number of simulations. This is visualized in Figure 11, whereas the corresponding Pareto optimal set is reported in Table S1. Complementary, the insets in Figure 11 depict the average values of the corresponding initial parameters and hyperparameters. This shows that increasing the `OVERLAP_THR` generally decreases the RMSE, although the data is biased against high `OVERLAP_THR` values as these OGRe runs typically did not converge. Moreover, the Pareto front is accompanied by median `CONFINEMENT_THR` values, with the error and number of simulations typically increasing for large `CONFINEMENT_THR` values. OGRe runs with a large `KAPPA_GROWTH_FACTOR` are clustered at low numbers of simulations, although those com-

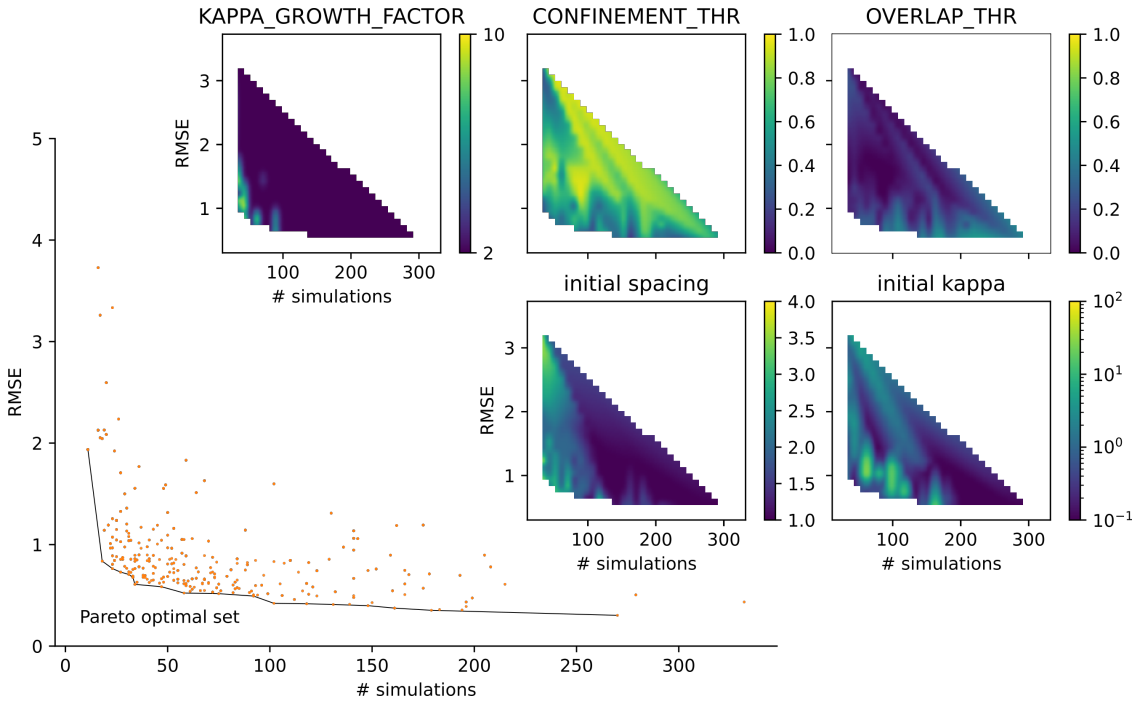

Figure 11: Pareto analysis as a function of the RMSE and number of simulations using the skewed bimodal potential. The Pareto curve connects those points that are optimal in this multi-object optimization of error and computational cost. They represent those combinations that cannot be improved in one of the aspects without worsening the other. The insets represent interpolated average values of the initial parameters ($\boldsymbol{\Delta}_l$ and $\boldsymbol{\kappa}_i$) and the hyperparameters (`CONFINEMENT_THR`, `OVERLAP_THR`, and `KAPPA_GROWTH_FACTOR`).

binations with large `KAPPA_GROWTH_FACTOR` in the Pareto optimal set generally show large RMSE values (see Table S1), in line with the conclusions from Figure 10. Finally, lower initial $\boldsymbol{\Delta}_l$ values and $\boldsymbol{\kappa}_i$ values mainly result in a larger number of simulation, but are generally accompanied by lower RMSE values, in line with Figures 8-9.

1D: trimodal Similarly, the OGRe protocol is applied to reproduce the following analytical potential:

$$U^b(\xi) = 0.1\xi^4 - 2\xi^2 - 8e^{-\xi^2} + 10 \quad \text{with} \quad -6 \leq \xi \leq 6 \tag{12}$$

as illustrated in Figure 7b, similar to a process with a stable intermediate. The effect of the aforementioned hyperparameters and initial parameters on the error and the number of simulations is illustrated in Figure S5. As the region of interest for the CV was halved with respect to the skewed bimodal potential, an even lower initial $\mathbf{\Delta}_l$ value was considered here. Despite the change in CV range, a lower initial $\mathbf{\Delta}_l$, and the different 1D potential energy profile (with more minima and different energy barriers), the protocol performs robustly and produces free energy profiles beyond chemical accuracy when avoiding the lowest `CONFINEMENT_THR` and `OVERLAP_THR`, as seen in the Pareto and parameter density plots illustrated in Figures S6-S7.

2D potentials Finally, the OGRe protocol is applied to the two-dimensional analytic potentials, the double well potential:

$$\begin{aligned} U^b(\xi_x, \xi_y) = (\xi_x^2 + \xi_y^2)^2 - 10\exp\left[-30(\xi_x - 0.2)^2 - 3(\xi_y - 0.4)^2\right] \\ - 10\exp\left[-30(\xi_x + 0.2)^2 - 3(\xi_y + 0.4)^2\right] \end{aligned} \tag{13}$$

and the Ackley potential:

$$U^b(\xi_x, \xi_y) = -20\exp\left[0.2\sqrt{\frac{\xi_x^2 + \xi_y^2}{2}}\right] - \exp\left[\frac{\cos(2\pi\xi_x) + \cos(2\pi\xi_y)}{2}\right] + \exp(1) + 20 \tag{14}$$

The first potential is the two-dimensional equivalent of the previously investigated skewed bimodal potential, showcasing two equivalent minima separated by a barrier, shown in Figure 7c. In contrast, the second potential challenges the protocol with a single global minimum, embedded in a very flat energy surface with a periodic arrangement of different equivalent

minima, shown in Figure 7d. Similar to the 1D potentials, Figures S8 and S12 depict the RMSE and convergence behavior of the double well and Ackley potentials, in line with the 1D results, which again illustrate that the OGRe protocol can successfully reproduce the free energy surface beyond chemical accuracy. This is explicitly visualized in Figure 12 for the Ackley potential, by comparing the analytical FES and the calculated FES for a particular parameter combination. Notably, to accommodate the exponentially larger phase space for the 2D potentials, the number of MD steps per US simulation was increased significantly compared to the 1D simulations (see Table S5). If instead the original number of MD steps is used, as illustrated in Figure S16, the OGRe protocol compensates by requiring more simulations, as the required overlap thresholds are more difficult to meet, and generally showcases a higher RMSE. In Figures S10 and S14, the average parameter values as a function of the error and the number of simulations are illustrated for both 2D potentials, which allows for very similar conclusions to be drawn with respect to the 1D potentials. The effect of the `CONSISTENCY_THR` was also investigated, to assess its influence in higher dimensions, as discussed in Section 2.4. As illustrated in Figures S11 and S15, a significant decrease in the error can be obtained, at the cost of more simulations, which is more pronounced for the Ackley potential. This likely follows from the fact that the Ackley potential has many more local minima which can influence the consistency of each simulation, counteracting the increased number of paths connecting the phase space in higher dimensions. Consequently, it is recommended for the `CONSISTENCY_THR` to be used for any number of dimensions, when dealing with multiple minima.

4 Application on physical systems

To showcase the performance of the OGRe package on physical systems, it is first applied on a well-documented 1D case study, namely proton hopping in zeolites. Afterwards, it is

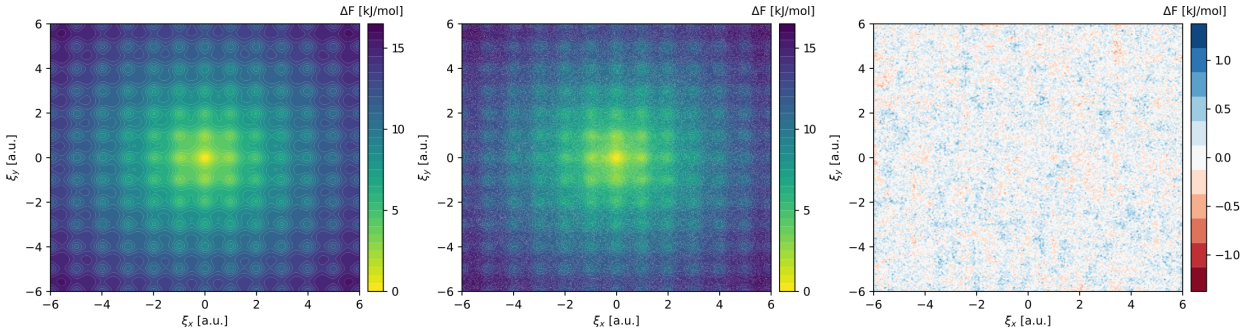

Figure 12: Efficacy of the OGRe protocol. Comparison between the (left) analytical, (middle) calculated, and (right) difference plot of the free energy surface of the Ackley potential, with an RMSE value of 0.29. Calculated with an initial $\boldsymbol{\Delta}_l$ of 2.0, initial $\boldsymbol{\kappa}_i$ of 1.0, `KAPPA_GROWTH_FACTOR` of 2, `CONFINEMENT_THR` of 0.33, `OVERLAP_THR` of 0.50, and `CONSISTENCY_THR` of 0.96.

applied on the difficult to capture two-dimensional free energy landscape of layer stacking in the prototypical two-dimensional covalent organic framework, COF-5.

Proton hopping in zeolites Although proton hopping is a fundamental activated event within zeolite chemistry, and despite its straightforward transition mechanism (see Figure 13), there is a large spread in both the experimentally and theoretically predicted transition barriers. A recent study by Bocus *et al.* highlighted a fundamental lack of nuclear quantum effects in the theoretical calculations of these barriers, which is prohibitively expensive to take into account at the density functional theory (DFT) level.[35] To this end, they derived a reactive machine learning potential (MLP), dramatically increasing the attainable space and time scales, with a similar accuracy to DFT. This allowed them to perform extended US simulations on a dense grid, with increased convergence of the individual simulations.

We validated the OGRe protocol on a specific proton hopping path at 873 K, denoted as 2-3 in ref. 35, where a difference in the coordination number between the proton and the two oxygen atoms involved in the hopping was used as the CV (see Figure 13 for the nomenclature). Evaluating the OGRe metrics on the available trajectory data of the corresponding

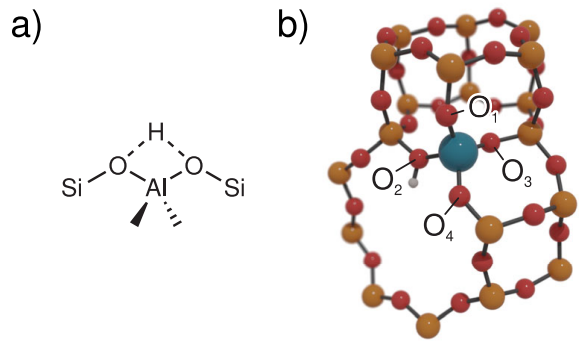

Figure 13: Proton hopping reaction within the H-CHA zeolite. (a) Schematic illustration of the hopping reaction. (b) Part of the H-CHA unit cell, showing the conventional nomenclature of the oxygen atoms in the first coordination sphere of the Al defect adopted herein (Si is in yellow, O in red, Al in blue and H in white). Reproduced from ref. 35 with permission of Springer Nature, copyright 2023.

US simulations revealed large overlaps between adjacent simulations, but highlighted a confinement of zero at the edges of the FES, and low confinements in the intermediate regions between the minima and the maximum, as shown in Figure 14b, due to the large free energy gradient. By limiting the range of the CVs to [-0.85,0.85] instead of [-0.95,0.95], and performing several umbrella strength refinements, the OGRe protocol did converge with only a handful of additional simulations, also performed with Yaff in correspondence to the original simulations.[28] This refinement was based on a `CONFINEMENT_THR` and `OVERLAP_THR` of 0.33, and a `CONSISTENCY_THR` of 0.95. With these thresholds, the OGRe refinement resulted in a slight decrease of the free energy difference between the two minima, as illustrated in Figure 14a. The bin width was also lowered with respect to the original bin width in accordance with Section S3, to allow for a sufficiently high `MAX_KAPPA` value. All OGRe parameters are reported in Tables S6-S7, whereas the computational details are discussed in Section S6.

While no consistency issues were raised by the OGRe protocol at a `CONSISTENCY_THR` value of 0.95, both the reference profile and the OGRe profile did showcase some deviation between the sampled and predicted biased probability distribution, as illustrated in Figure 14c. In an attempt to improve on this FES, the `CONSISTENCY_THR` was increased to 0.98. However, no

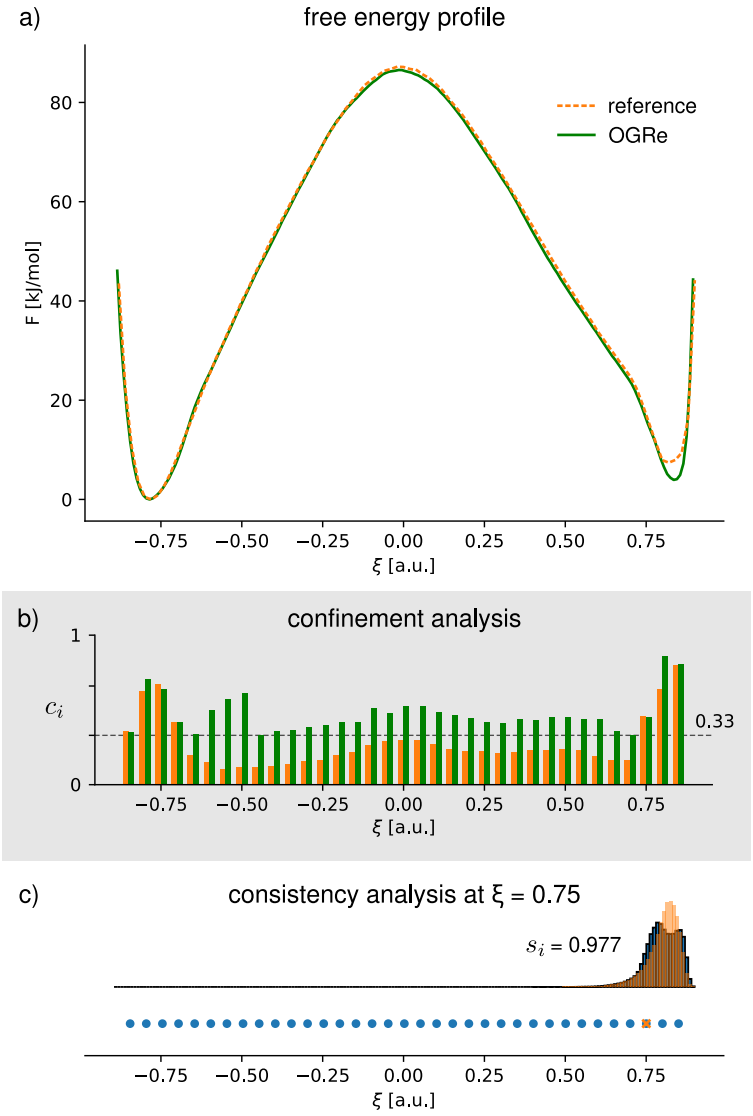

Figure 14: Application of the OGRe protocol on proton hopping in zeolites. (a) Comparison between the converged FES as calculated through the OGRe protocol, and the reference pattern from ref. 35. (b) Overview of the confinement metric c_i, for the reference trajectories (in orange) and the OGRe trajectories (in green). The imposed `CONFINEMENT_THR` of 0.33 is indicated with the dashed line. (c) Consistency analysis in the secondary minimum, showing a minor deviation between the sampled and predicted biased probability distribution, indicating that there is likely a finer structure to the FES.

amount of umbrella strength refinements could resolve the inconsistency. As such, it is likely that the consistency issue is not related to any ergodicity problems in terms of the CV, but rather the orthogonal degrees of freedom that are not biased. Evidence of this conjecture can also be found in Supplementary Figure 6 of ref. 35, where a two-dimensional expansion

of the one-dimensional FES along the H–O_2 and H–O_3 distances highlights a structure that the coordination number alone might not be able to describe.

Layer stacking in COF-5 It has already been well established that 2D COFs showcase dynamic layer behavior, with a random stacking orientation, yet a fixed offset.[36–43] This has been attributed to the interplay of electrostatic and van der Waals interactions, with in particular Pauli repulsion. Although the 2D potential energy surface for COF-5 has already been reported several times,[38,40,43] the free energy surface at operating condition has not been predicted before. For optimal accuracy, while avoiding a new hyperparameter screening, the hyperparameters for the OGRe protocol were chosen in correspondence with Section 3, and have been tabulated along with the simulation parameters in Tables S6-S7. The CVs to describe layer stacking were chosen in correspondence to previous work,[41,43] as described in Section S5.1, together with a derivation of the equations required to bias along these CVs in a US simulation. Finally, a $1 \times 1 \times 2$ supercell (with the last dimension along the stacking direction) was used to allow for sufficient freedom, as illustrated in Figure 15a, whereas the interactions were modeled with a previously derived force field,[41] and the simulations were performed with Yaff,[28] as discussed in Section S6. Figure 15b visualizes the resulting free energy surface, where the expected fixed layer offset at with a random orientation is apparent from the ring of minimal energy, centered around the origin indicating perfect alignment. Looking beyond, at higher layer offsets, the structure and symmetry of the layer geometry is apparent, originating from the variations in the interplay of interactions. Notably, the layer offset predicted by the force field (2.3 Å) is slightly larger than the DFT predicted value at 0 K (1.6 Å).[38,43] However, as discussed in Section S7, this shift in the layer offset is not driven by entropic effects, but rather enthalpically driven through the force field (see Figure S18).

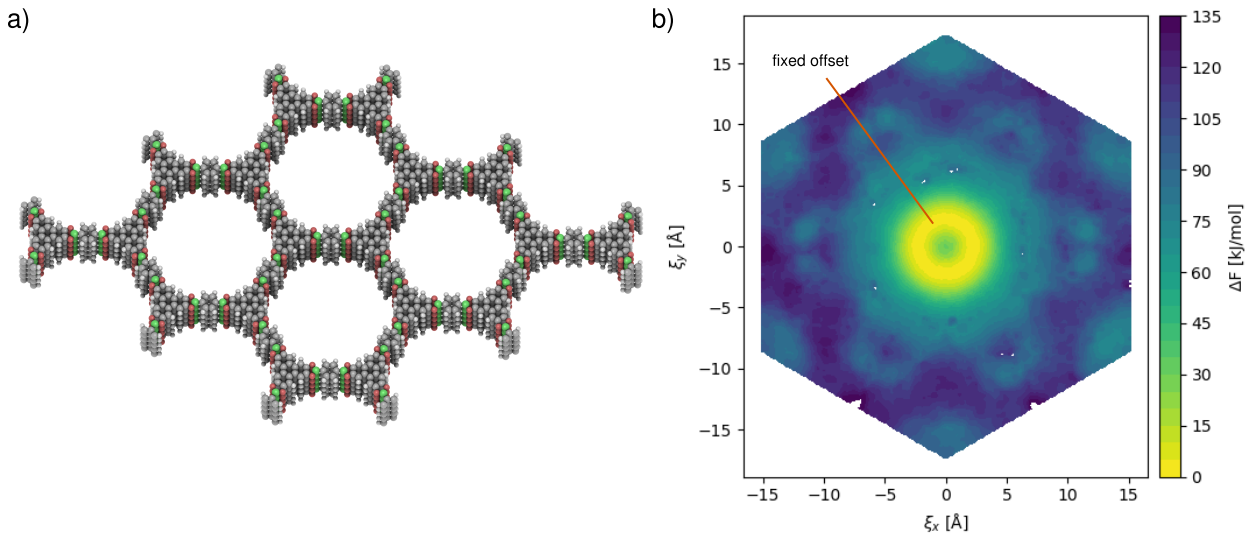

Figure 15: Application of the OGRe protocol on layer stacking in COF-5 (a) The atomic structure of COF-5. (b) Free energy surface of the layer stacking in COF-5, where the hexagonal boundaries delineate the periodicity of the FES.

5 Conclusions

We herein presented OGRe, an easy-to-use and flexible Python package aiming to create an optimal sampling grid for free energy evaluation methods that require an overlap of simulated probability densities, such as WHAM, based on underlying US simulations. Through an iterative procedure, the OGRe protocol refines an initial grid in both bias strength and local sampling density, to obtain a maximum accuracy with minimal computational effort. This process is facilitated by three metrics that gauge the confinement, consistency, and overlap of each simulation, based on the requirements for an accurate WHAM calculation of the FES. The efficacy of these metrics is clearly established, and rules were established on how OGRe leverages these metrics to improve on the initial grid of umbrella parameters. In particular, the consistency metric proves to be an excellent gauge for possible ergodicity errors, which should prove highly valuable for free energy methods in general.

Because the choice of hyperparameters for the refinement protocol is crucial, benchmarked

values have been calculated, with a robust working range. Generally, by increasing the `CONFINEMENT_THR` and `OVERLAP_THR` values, the error on the generated free energy surfaces decreases, following the increased sampling for the steepest free energy gradients and better conditioned WHAM equations through increased overlap. However, higher threshold values also increase the required number of simulations for convergence as evidenced by the Pareto analysis. Regardless, the difficult choice of grid density and bias strength for umbrella sampling simulations is partially lifted through automatic refinement, and the robust working range for each parameter in OGRe. As such OGRe proves to be successful in deriving FESs within chemical accuracy, largely independent of the initially provided parameters, while its efficiency remains, to some extent, user-driven. Moreover, through the benchmarking on analytic potentials, several issues with histogram-based methods are highlighted, along with suggested workarounds. In particular, due to the finiteness of the bin width, care should be taken with the maximal $\boldsymbol{\kappa}_i$ value. When the umbrella strength grows too large, all samples could be assigned to a single bin, resulting in a poor reproduction of the underlying probability distribution.

When applied on physical systems, the OGRe protocol is capable of improving upon previously reported free energy profiles, and can highlight the potential for further refinement, as indicated with our example on proton hopping in zeolites. Through application on the layer stacking of 2D COFs, we furthermore highlight its potential to characterize the underlying phase landscape for layer alignment, driven by the collective variables as defined in this work. Hopefully, the OGRe protocol, and its metrics, will prove to be very valuable towards a perfect chemical precision for free energy calculations, and this work provides the impetus to uncover previously challenging free energy landscapes through automated procedures.

Acknowledgement

This work is supported by the Research Board of Ghent University (BOF) through a Concerted Research Action (GOA010-17). S.M.J.R. acknowledges the Research Foundation Flanders (FWO) for a postdoctoral fellowship (grant no. 12T3522N). V.V.S. acknowledges the Research Board of Ghent University (BOF). The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government – department EWI.

References

(1) Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. Basic ingredients of free energy calculations: A review. *J. Comput. Chem.* **2010**, *31*, 1569–1582.

(2) Paul, S.; Nair, N. N.; Vashisth, H. Phase space and collective variable based simulation methods for studies of rare events. *Mol. Simul.* **2019**, *45*, 1273–1284.

(3) Pal, A.; Pal, S.; Verma, S.; Shiga, M.; Nair, N. N. Mean force based temperature accelerated sliced sampling: Efficient reconstruction of high dimensional free energy landscapes. *J. Comput. Chem.* **2021**, *42*, 1996–2003.

(4) Tavernelli, I.; Cotesta, S.; Iorio, E. E. D. Protein Dynamics, Thermal Stability, and Free-Energy Landscapes: A Molecular Dynamics Investigation. *Biophys. J.* **2003**, *85*, 2641–2649.

(5) Shirts, M. R. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Springer New York: New York, NY, 2012; pp 425–467.

(6) Borgmans, S.; Rogge, S. M. J.; De Vos, J. S.; Van Der Voort, P.; Van Speybroeck, V. Ex-

ploring the phase stability in interpenetrated diamondoid covalent organic frameworks. *Commun. Chem.* **2023**, *6*, 5.

(7) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.

(8) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245–268.

(9) Torrie, G.; Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(10) Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.

(11) Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the analysis of free energy calculations. *J. Comput. Aided Mol. Des.* **2015**, *29*, 397–411.

(12) Morita, A. *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*; Elsevier, 2017.

(13) Mezei, M. Adaptive umbrella sampling: Self-consistent determination of the non-Boltzmann bias. *J. Comput. Phys.* **1987**, *68*, 237–248.

(14) Berne, B. J.; Straub, J. E. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struct. Biol.* **1997**, *7*, 181–189.

(15) Marsili, S.; Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V. Self-healing Umbrella Sampling: A Non-equilibrium Approach for Quantitative Free Energy Calculations. *J. Phys. Chem. B* **2006**, *110*, 14011–14013.

(16) Maragakis, P.; van der Vaart, A.; Karplus, M. Gaussian-Mixture Umbrella Sampling. *J. Phys. Chem. B* **2009**, *113*, 4664–4673.

(17) Rousset, M.; Stoltz, G.; Lelievre, T. *Free Energy Computations*; Imperial College Press, 2010.

(18) Wojtas-Niziurski, W.; Meng, Y.; Roux, B.; Bernèche, S. Self-Learning Adaptive Umbrella Sampling Method for the Determination of Free Energy Landscapes in Multiple Dimensions. *J. Chem. Theory Comput.* **2013**, *9*, 1885–1895.

(19) Frenkel, D.; Smit, B. *Understanding molecular simulation: from algorithms to applications*; Elsevier, 2001; Vol. 1.

(20) Bartels, C.; Karplus, M. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comput. Chem.* **1997**, *18*, 1450–1462.

(21) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(22) Roux, B. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **1995**, *91*, 275–282.

(23) Zhu, F.; Hummer, G. Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* **2011**, *33*, 453–465.

(24) Kästner, J.; Thiel, W. Analysis of the statistical error in umbrella sampling simulations by umbrella integration. *J. Chem. Phys.* **2006**, *124*, 234106.

(25) Park, S.; Im, W. Theory of Adaptive Optimization for Umbrella Sampling. *J. Chem. Theory Comput.* **2014**, *10*, 2719–2728.

(26) Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 932–942.

(27) Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.

(28) Verstraelen, T.; Vanduyfhuys, L.; Vandenbrande, S.; Rogge, S. M. J. Yaff, yet another force field. `http://molmod.ugent.be/software/`.

(29) Straatsma, T. P.; McCammon, J. A. Multiconfiguration thermodynamic integration. *J. Chem. Phys.* **1991**, *95*, 1175–1188.

(30) Demuynck, R.; Wieme, J.; Rogge, S. M. J.; Dedecker, K. D.; Vanduyfhuys, L.; Waroquier, M.; Van Speybroeck, V. Protocol for Identifying Accurate Collective Variables in Enhanced Molecular Dynamics Simulations for the Description of Structural Transformations in Flexible Metal–Organic Frameworks. *J. Chem. Theory Comput.* **2018**, *14*, 5511–5526.

(31) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(32) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

(33) Oldenhuis, R. Test functions for global optimization algorithms. **2023**,

(34) Shannon, C. Communication in the Presence of Noise. *Proc. IRE* **1949**, *37*, 10–21.

(35) Bocus, M.; Goeminne, R.; Lamaire, A.; Cools-Ceuppens, M.; Verstraelen, T.; Van Speybroeck, V. Nuclear quantum effects on zeolite proton hopping kinetics explored with machine learning potentials and path integral molecular dynamics. *Nat. Commun.* **2023**, *14*, 1008.

(36) Lukose, B.; Kuc, A.; Frenzel, J.; Heine, T. On the reticular construction concept of covalent organic frameworks. *Beilstein J. Nanotechnol.* **2010**, *1*, 60–70.

(37) Lukose, B.; Kuc, A.; Heine, T. The structure of layered covalent-organic frameworks. *Chemistry* **2011**, *17*, 2388–2392.

(38) Koo, B. T.; Dichtel, W. R.; Clancy, P. A classification scheme for the stacking of two-dimensional boronate ester-linked covalent organic frameworks. *J. Mater. Chem.* **2012**, *22*, 17460–17469.

(39) Pütz, A. M.; Terban, M. W.; Bette, S.; Haase, F.; Dinnebier, R. E.; Lotsch, B. V. Total scattering reveals the hidden stacking disorder in a 2D covalent organic framework. *Chem. Sci.* **2020**, *11*, 12647–12654.

(40) Winkler, C.; Kamencek, T.; Zojer, E. Understanding the origin of serrated stacking motifs in planar two-dimensional covalent organic frameworks. *Nanoscale* **2021**, *13*, 9339–9353.

(41) Borgmans, S.; Rogge, S. M. J.; De Vos, J. S.; Stevens, C. V.; Van Der Voort, P.; Van Speybroeck, V. Quantifying the Likelihood of Structural Models through a Dynamically Enhanced Powder X-Ray Diffraction Protocol. *Angew. Chem. Int. Ed.* **2021**, *60*, 8913–8922.

(42) Zhang, Y.; Položij, M.; Heine, T. Statistical Representation of Stacking Disorder in Layered Covalent Organic Frameworks. *Chem. Mater.* **2022**, *34*, 2376–2381.

(43) Rawat, K. S.; Borgmans, S.; Braeckevelt, T.; Stevens, C. V.; Van Der Voort, P.; Van Speybroeck, V. How the Layer Alignment in Two-Dimensional Nanoporous Covalent Organic Frameworks Impacts Its Electronic Properties. *ACS Appl. Nano Mater.* **2022**, *5*, 14377–14387.