The Journal of Chemical Physics

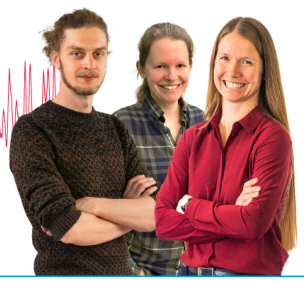
RAPID COMMUNICATION | OCTOBER 13 2025

On the reproducibility of free energy surfaces in machine-learned collective variable spaces

Special Collection: Michele Parrinello Festschrift

Florian M. Dietrich ; Matteo Salvalaglio

J. Chem. Phys. 163, 141102 (2025)
https://doi.org/10.1063/5.0287912



The Journal of Chemical Physics COMMUNICATION pubs.aip.org/aip/jcp

On the reproducibility of free energy surfaces in machine-learned collective variable spaces

Cite as: J. Chem. Phys. 163, 141102 (2025); doi: 10.1063/5.0287912
Submitted: 27 June 2025 • Accepted: 23 September 2025 •
Published Online: 13 October 2025

Florian M. Dietrich[a] and Matteo Salvalaglio[b]

AFFILIATIONS

Thomas Young Centre and Department of Chemical Engineering, University College London, London WC1E 7JE, United Kingdom

Note: This paper is part of the JCP Special Topic, Michele Parrinello Festschrift.
[a]florian.dietrich.20@ucl.ac.uk
[b]**Author to whom correspondence should be addressed:** m.salvalaglio@ucl.ac.uk

ABSTRACT

As Machine-Learned Collective Variables (MLCVs) are becoming increasingly relevant in the molecular simulation literature, we discuss the necessary conditions to enable reproducibility in calculating and representing free energy surfaces. We note that the variability of the training process and the roughness of the hyperparameter space impose inherent limits on the reproducibility of results even when the mathematical structure of the model defining a collective variable is consistent. To this end, we propose the adoption of a geometric (gauge invariant) free energy representation to obtain consistent free energy differences across training instances and architectures. Furthermore, we introduce a normalization factor to model gradients for biased enhanced sampling. This factor effectively unifies free energy definitions and addresses practical issues preventing the widespread use and deployment of MLCVs.

Machine learning (ML) techniques have revolutionized the development of approaches to process, classify, and rationalize high-dimensional data.[1] This includes the development of Collective Variables (CVs) for analyzing Molecular Dynamics (MD) trajectories.[2–7] Generally speaking, CVs are low-dimensional descriptors that capture the slow modes of a high-dimensional atomistic system, effectively compressing high-dimensional information into low-dimensional models. More rigorously, given a set of atomic coordinates $\mathbf{R}$, a CV $\mathbf{s}(\mathbf{R})$ is a mapping $\mathbf{s}: \mathbb{R}^n \to \mathbb{R}^m$, where $m \ll n$. In practice, $\mathbf{s}(\mathbf{R})$ provides a useful low-dimensional description that captures all the features of the process of interest. CVs play a key role in the quantitative analysis of atomistic simulations by providing a readable and often physically meaningful space to project configurational Boltzmann distributions, compute free energy differences between macrostates, and estimate macrostate-dependent structural observables.[5,6] A multitude of data-driven methods exists to aid in identifying such dimensionality reductions, be it in a linear unsupervised manner from the variance in the training data with methods such as principal component analysis (PCA),[8] or unsupervised non-linear mappings such as kernel-PCA, diffusion maps,[9–11] sketch-map,[2] or autoencoders.[12–14] Alternatively, one can formulate specific learning objectives in a supervised setting to efficiently approximate many-body CVs based on physical intuition[15] or to, in a semi-supervised manner, identify the slowest modes of the system.[16–20] More generally, the use of ML in CV development offers several advantages, including increased efficiency, more robust and automated data compression compared to traditional methods, and the potential for transferability of models across different systems. Furthermore, ML can be leveraged to learn and deploy CVs that achieve a more "ideal" approximation of rigorously defined reaction coordinates, using the committor.[7,21] Graph-based architectures, in particular, are easily adaptable to capture the complexities and inherent invariances of atomistic systems.[15,22–25] Unlike traditional CVs, which are often informed by physical intuition and experience, the definition of the analytical form of MLCVs—besides the adopted ML architecture—depends on the choice of hyperparameters and, crucially, on the training process.[3] In this Communication, we reflect on the implications of MLCVs' inherent variability on the low-dimensional representation of configurational probability distributions, the associated calculation of free energy surfaces (FESs),

and the implications for enhanced sampling simulations based on introducing a biasing potential along MLCVs.

Within computational physical chemistry, a free energy surface $F(\mathbf{s})$ is typically used to provide a useful low-dimensional representation of the configurational Boltzmann distribution of a multi-body, atomistic system. As such, $F(\mathbf{s})$ is defined as proportional to the negative logarithm of the marginal probability density $p(\mathbf{s})$,[26–31]

$$\begin{aligned} F(\mathbf{s}) &= -k_B T \ln p(\mathbf{s}), \\ p(\mathbf{s}) &= \frac{1}{Z}\int_{\mathbb{R}^n} e^{-\beta U(\mathbf{R})}\delta(\mathbf{s}(\mathbf{R})-\mathbf{s})\,\mathrm{d}\mathbf{R}, \end{aligned} \tag{1}$$

where k_B is the Boltzmann constant, T is the absolute temperature, $\beta = (k_B T)^{-1}$, $U(\mathbf{R})$ is the potential energy of the system, and $Z = \int e^{-\beta U(\mathbf{R})}\mathrm{d}\mathbf{R}$ is its configurational integral. The expression integrates over the entire configuration space $\mathbb{R}^n$ and selects configurations for which the value of $\mathbf{s}$ is constant in a process akin to projecting onto collective variable space $\mathbb{R}^m$. This "projection" can be made more explicit through application of the coarea formula,[32–35]

$$p(\mathbf{s}) = \frac{1}{Z}\int_{\Sigma_\mathbf{s}} e^{-\beta U(\mathbf{R})}\,\mathrm{vol}(J_\mathbf{s})^{-1}\,\mathrm{d}\sigma, \tag{2}$$

where we now integrate over the level set $\sum_\mathbf{s}$,

$$\Sigma_\mathbf{s} = \{\mathbf{R}\in\mathbb{R}^n : \mathbf{s}(\mathbf{R}) = \mathbf{s}\}, \tag{3}$$

where each level contains all configurations $\mathbf{R}\in\mathbb{R}^n$ that are degenerate at a given point in CV space, i.e., that map onto the same value of $\mathbf{s}(\mathbf{R})$. $\mathrm{vol}(J_\mathbf{s})$ is the volume of the Jacobian of $\mathbf{s}$ and accounts for the geometric distortion of the collective variable mapping.[36] The Jacobian of $\mathbf{s}$ is defined as

$$J_\mathbf{s} = \begin{bmatrix} \nabla^\mathsf{T} s_1 \\ \vdots \\ \nabla^\mathsf{T} s_m \end{bmatrix}, \tag{4}$$

$$\nabla^\mathsf{T} s = \begin{bmatrix} \frac{\partial s}{\partial x_1} \cdots \frac{\partial s}{\partial x_n} \end{bmatrix}, \tag{5}$$

where n is the number of coordinates and m is the number of CVs. By definition, $m \ll n$, and geometrically, for a rectangular matrix A, only the determinant of the smaller of the two square matrices $A^T A$ and AA^T can be interpreted as a volume.[37] This means, in this context, that the volume is defined as[38]

$$\mathrm{vol}(J_\mathbf{s}) = \sqrt{\det J_\mathbf{s} J_\mathbf{s}^T}. \tag{6}$$

Therefore, the nature of this projection depends on the exact choice of CV. In practice, this is often not a concern as the effects of this projection are grounded in physical intuition, e.g., projecting a volume onto a distance, and are consistently repeatable with every application of the same set of CVs. As a consequence, Eq. (1) is widely adopted in the enhanced sampling literature as it enables one to practically estimate the probability density $p(\mathbf{s})$ from a straightforward histogramming and reweighting procedure.[28–31] This implies that the expression defines $F(\mathbf{s})$ up to an immaterial constant shift and can be used to estimate free energy differences between ensembles of microstates that coexist in $\mathbf{s}$. The resulting FES can be straightforwardly interpreted as a marginal probability density encoding the statistics of the investigated system at thermal equilibrium.[35] However, Eq. (1) also implies that different CVs yield free energy surfaces with different shapes and extrema.

While the shape of $J_\mathbf{s}$ is inherently well-defined for traditional CVs, the same cannot be said for MLCVs for two main reasons: The first is that MLCVs are typically defined as *model architectures* with problem-specific, tunable hyperparameters.[3,15,24] The second is that, even when hyperparameters are fixed, constraining the functional form of the MLCV, their parameterization depends on an inherently stochastic training process.[39] Consequently, every time one constructs an FES in the space of MLCVs, one constructs a representation of the free energy unique to the specific set of model weights that does not correspond to a shared understanding of the physics of the investigated problem.

A prominent application of MLCVs is within the context of *biased* enhanced sampling methods.[3,15,23,25] The inherent stochasticity associated with training here has an effect, too. During such simulations, a bias potential $V(\mathbf{s})$ is applied to drive the system to explore the given phase space. Doing so requires applying a force $\vec{f}$ to the particles of the system, which is proportional to the gradient of V in $\mathbf{s}$,

$$\vec{f} = \frac{\partial V}{\partial \mathbf{s}}\frac{\partial \mathbf{s}}{\partial \mathbf{x}}. \tag{7}$$

In this case, stochastic variations in $J_\mathbf{s}$ can lead to unexpected force spikes, vanishing gradients and inconsistent behavior between models with otherwise identical simulation parameters.

These practical and theoretical concerns can be alleviated by normalizing the gradients of MLCVs in production. In the following, we will show that normalized MLCVs sample a valid representation of the free energy and that this representation of the free energy is inherently more suitable for the domain of MLCVs.

An alternative definition of the free energy surface can be obtained by expressing the probability density $p(\mathbf{s})$ directly as the surface integral over level set $\Sigma_\mathbf{s}$. This formulation of the free energy is referred to as the *geometric* free energy and is typically denoted with G[32–35] or F_G,[40]

$$\begin{aligned} F_G(\mathbf{s}) &= -k_B T \ln q(\mathbf{s}), \\ q(\mathbf{s}) &= \int_{\Sigma_\mathbf{s}} e^{-\beta U}\mathrm{d}\sigma. \end{aligned} \tag{8}$$

Since now we integrate directly over the level set, this expression is invariant to any "degeneracy-preserving" transformation f, with f being any monotonic function with a non-zero gradient applied to any s in $\mathbf{s}$.[32]

The advantage of F_G, therefore, is that it is gauge invariant,

$$F_G(s) = F_G(f(s)). \tag{9}$$

In practical terms, gauge invariance implies that features extracted from F_G, i.e., local free energy minima, maxima, and saddle points, are not dependent on, for example, the units or exponent of the CV. This property makes it a useful representation to compare features of FESs between different sets of MLCVs.[41]

It is worth noting that both Eqs. (1) and (8) are valid definitions of the concept of free energy. The former provides a physically

intuitive way of encoding the relative stability of states, whereas the latter preserves kinetic information at the cost of a straightforward interpretation.[35]

Previous work by Hartmann and Schütte has shown that $F_G(\mathbf{s})$ can be recovered from $F(\mathbf{s})$ by correcting $F(\mathbf{s})$ to account for the ensemble average $\langle\cdot\rangle_{\mathbf{s}}$ of the volume of the Jacobian $J_{\mathbf{s}}$,[35]

$$F_G(\mathbf{s}) = F(\mathbf{s}) - k_B T \ln \langle \lambda^m \mathrm{vol}(J_{\mathbf{s}}) \rangle_{\mathbf{s}}, \tag{10}$$

where λ is a characteristic length scale that ensures that the logarithm is unitless and independent of the chosen unit system.[34]

Bal *et al.*[40] developed a scheme that allows for a straightforward construction of $F_G(\mathbf{s})$ from sampling. They introduce a correction that modifies the weight w of a given configuration to yield a geometric weight w_G. For a monodimensional CV space, Bal's expression reads

$$w_G = w \cdot \lambda^m \|\nabla s\|. \tag{11}$$

As shown in the Appendix, this expression can be derived by substituting $m = 1$ in the ensemble average of Eq. (10). Generalizing w_G for an arbitrary m-dimensional CV space $\mathbf{s}$ leads to[42]

$$w_G = w \cdot \lambda^m \sqrt{\det J_{\mathbf{s}} J_{\mathbf{s}}^T}. \tag{12}$$

The weight w depends on the ensemble where sampling is carried out. For instance, $w = 1$ when samples are drawn from an unperturbed ensemble, while when sampling is performed in a biased ensemble, $w \propto \exp \beta V_B(\mathbf{s})$, where $V_B(\mathbf{s})$ is the bias perturbing the system's Hamiltonian.[27]

This brief introduction of $F(\mathbf{s})$ and $F_G(\mathbf{s})$ underscores a key difference between the two estimators of thermodynamic stability. While $F_G(\mathbf{s})$ explicitly compensates for the volume change associated with the mapping $\mathbf{R} :\to \mathbf{s}$, $F(\mathbf{s})$ is dependent on the exact shape and magnitude of $\mathbf{J}_s$.[34,40]

From Eq. (11), it is evident that biasing in the space of normalized MLCVs directly samples the geometric FES in the 1D case and the geometric FES up to a cross-correction term in the general case. This lends a theoretical justification to the practically motivated decision to normalize the gradients, i.e., the vectors $\nabla s_1, \ldots, \nabla s_m$.

As previously stated, while sampling alone affects the accuracy of free energy surfaces as a function of traditional CVs when MLCVs are employed, both hyperparameter optimization and training crucially affect the topology of low-dimensional representations and the accuracy of free energy estimates. Beyond their practical benefits, geometric free energy surfaces also theoretically solve this issue by removing the explicit dependence on d$\mathbf{s}$. However, to have the same F_G, two CVs must share the same level sets, i.e., be related via a gauge transformation.

Any machine-learning application is an inherently stochastic process, and any supervised machine-learning application assumes that the true relationship between an input $\mathbf{x}$ and a label y can be "learned" by finding an optimal set of parameters θ_{opt} for a function $f(\mathbf{x};\ \theta_{opt})$ up to a normally distributed irreducible noise term ϵ,[43]

$$y = f(\mathbf{x}; \theta_{opt}) + \epsilon, \text{with} \langle \epsilon \rangle = 0. \tag{13}$$

The error of a non-ideal model can be decomposed into a contribution from insufficient model flexibility to capture the true relationship between label and data, a contribution from the training set composition, and the previously mentioned irreducible error ϵ. Controlling for the first two components, by fully converging two models of the same architecture on the same training set, the difference in the two model outputs on the same input $\mathbf{x}$ is 2ϵ.

For two CVs to share the same F_G, they have to share the same level sets, i.e., if a set of configurations maps to a single value in $\mathbf{s}_1$, they have to map to a single value in $\mathbf{s}_2$, although the two values do not have to match. In the case of two machine-learned CVs s_{m1} and s_{m2}, trained identically, the surfaces at any given point are assumed to have a corresponding distance of 2ϵ but an expected average distance of $\langle 2\epsilon \rangle = 0$ and therefore,

$$F_{G,s_{m1}} \approx F_{G,s_{m2}}. \tag{14}$$

This approximation weakens when the two model CVs have different architectures, training parameters, and training sets.

Here, we devise a minimal experiment to motivate the use of F_G for comparing free energy differences between metastable states and show that this concept, which has thus far been mostly confined to the study of kinetics, has additional utility in the field of machine-learned CVs.

We train a simple autoencoder (AE) with a varying number of parameters on a toy system consisting of a 100-dimensional (100D) space with four embedded macrostates of equal probability (i.e., where $\Delta F_{i,j} = 0, \forall i,j$ macrostates). Figure 1 shows the training of three such models with an increasing number of parameters and their corresponding $F(\mathbf{s})$. All three models successfully separate the four macrostates in two dimensions, but the relative shape of the resulting basins is largely randomized. As a consequence, when estimating the relative stability of these basins as $\Delta F_{i,j} = \min F_j - \min F_i$, as is often the case in the literature, instead of integrating over the whole basin, one obtains free energy differences that deviate strongly from the expected $\Delta F = 0$ until the geometric correction is applied.

To further quantify this effect between different realizations of the *same* model architecture, 100 independent models are trained for 100 epochs for each of the three model sizes to approximately the same training loss.

Figure 2 reports the distribution of $\Delta F_{i,j}$ computed in the space of the MLCV $\mathbf{s}$ with (ΔF_G) and without (ΔF) the Gauge-invariance correction to the configurational weights. We note that, as expected, the average $\langle \Delta F_G \rangle$ is close to zero and less affected by the additional variance from increasing model complexity. By contrast, $\langle \Delta F \rangle$ (Fig. 2) scales with increasing model complexity, with individual realizations accounting for deviations from the expected ΔF by up to $\approx 2k_BT$. In realistic applications where both the dimensionality of $\mathbf{R}$ and the number of parameters of the MLCVs are higher, we envisage these deviations to be larger, less predictable, and more complex to disentangle from inherent sampling uncertainties.

This behavior is consistent with Eq. (14), underscoring the fact that introducing a gauge correction and representing distributions in low-dimensional CV spaces using *geometric* free energy surfaces should be inherently preferred when dealing with MLCVs, where stochastic model variations are to be expected. This is particularly important when evaluating the free energy associated with an

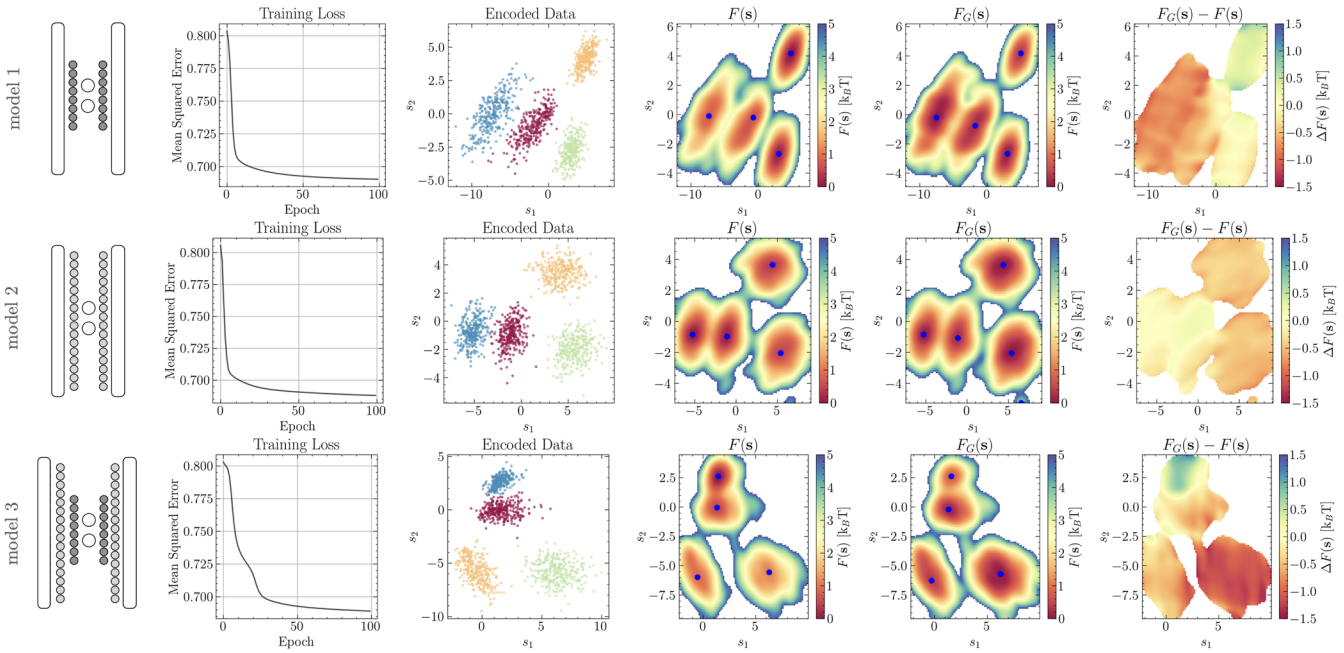


FIG. 1. Three autoencoders of increasing model complexity are trained to embed a 100-dimensional distribution with four equally probable macrostates into a two-dimensional space. The pictograms on the left illustrate each model architecture, followed by the respective training curves and learned low-dimensional representations. This demonstrates that all models converge to a comparable loss under an analogous training schedule. Despite encoding the data into distinct collective variable spaces $\mathbf{s} = [s_1, s_2]$, the resulting embeddings are of similar quality. The final three columns display the corresponding free energy surfaces $F(\mathbf{s})$, the geometric free energy surfaces $F_G(\mathbf{s})$, and their pointwise difference $F_G(\mathbf{s}) - F(\mathbf{s})$, illustrating how geometric corrections recover consistency in free energy estimates across model instances.

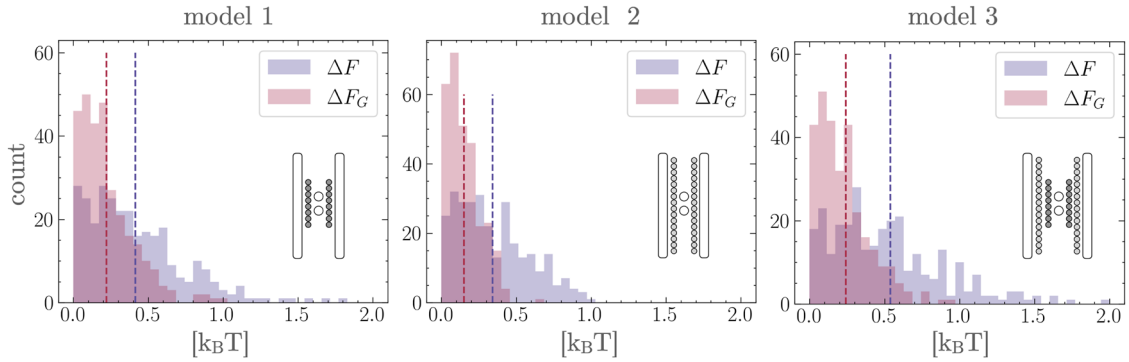


FIG. 2. Three histograms of the relative stabilities ΔF and ΔF_G extracted from the free and geometric free energy surfaces corresponding to 100 independently trained instances of the indicated model architectures. The (geometric) free energy difference between any two basins i and j is evaluated as $\Delta F_{ij} = \min F_j - \min F_i$. The dashed lines indicate the averages of the sampled distributions. While free-energy differences between basins are more rigorously defined by integrating over basin ensembles, here we report point-to-point ΔF values to highlight the maximum variability introduced by retraining MLCVs. This choice reflects that integration is only meaningful for well-defined (meta)stable basins, whereas point estimates remain a common practice even for traditional CVs.

ensemble of configurations that project on point features of the FES, such as transition states.

As an example to illustrate how the relative magnitude of individual gradient components and the gradient norms of machine-learning-based CVs can vary between model trainings even when using all the same parameters, we train four graph-neural networks (GNNs) to mimic the behavior of the analytical collective variable $n(Q6)$, which counts the number of particles in a simple colloidal system with a Steinhardt parameter higher than a specific cutoff. For details on the models, the collective variable, and the system, we refer to our previous publications.[15,44]

Figure 3 shows the result of training four such models with two different architectures (10 latent dimensions + 1 graph convolutional layer and 25 latent dimensions + 1 graph convolutional layer)

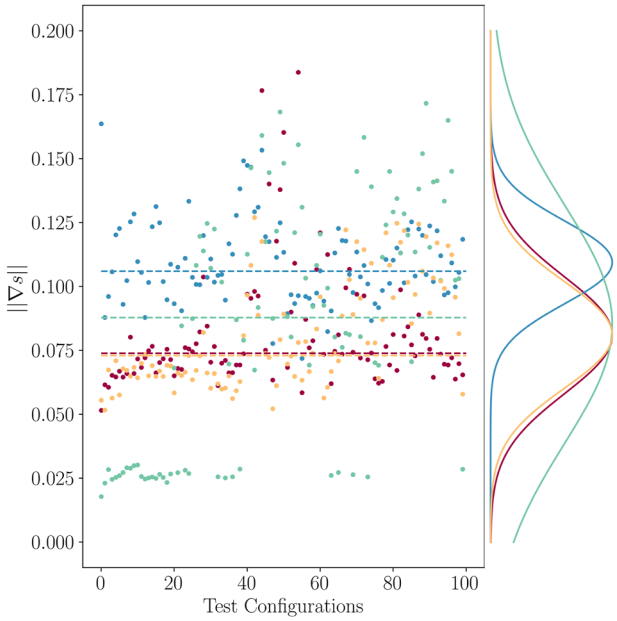


FIG. 3. Spread of gradient norms of four models with two different architectures (25 latent dimensions + 1 graph convolutional layer, red and blue, and 10 latent dimensions + 1 graph convolutional layer, gold and green) trained to the same accuracy over 100 random configurations. The dashed lines indicate the median norm of a given model.

to approximately the same training loss and plotting their gradient norms on 100 test configurations. Here, each column of points represents the different gradient norms on the *same* test configurations. It is easy to see that the models all behave differently, even the two models that exhibit a similar overall distribution of gradient norms exhibit substantial differences when compared on a per-point basis.

Therefore, when one compares two free energy surfaces constructed in the same CV space but from independently trained models, one implicitly assumes that the change $\frac{\partial \mathbf{s}}{\partial \mathbf{x}}$ is negligible. However, we showed above how the retraining of an MLCV corresponds to a transformation of the corresponding $F(\mathbf{s})$. In this sense, biasing along an MLCV results in a low-dimensional representation of the potential energy surface that is only accessible via the exact model parameters of $\mathbf{s}$. In practice, these are rarely published, and it would be more useful to have a representation that corresponds to a methodology rather than a set of parameter values without physical meaning. We, therefore, suggest normalizing model gradients during biasing such that

$$\lambda^m \sqrt{\det J_{\mathbf{s}} J_{\mathbf{s}}^T} = 1. \tag{15}$$

The advantages of this approach are threefold: First, this effectively means that any sampled F is also an F_G, replicable without access to the exact model weights. Second, this simplifies the deployment of MLCVs for practitioners since the same simulation parameters affect the behavior of simulations in the same way across instances of independently trained equivalent MLCVs. Third, this alleviates numerical stability issues associated with model gradients nearing zero in basins. This is an issue common in MLCVs, as pointed out by Gökdemir and Rydzewski in their recent review on the field as a whole.[3]

Recently, methods to construct CVs using machine learning have become increasingly popular. However, to the authors' knowledge, it has not yet been addressed that retraining an MLCV constitutes a transformation of the corresponding FES. At least in theory, this means that any published FES in the space of MLCVs is not reproducible without access to the exact model parameters, which are rarely published. In this work, we borrow from the study of kinetics from enhanced sampling simulations to suggest a best practice to publish geometric free energy surfaces when working with MLCVs. This practice promotes reproducibility in the field by producing FESs that do not simply correspond to a set of parameters with limited physical meaning, but a methodology as a whole, including the physics imbued into it by its developers.

F.M.D. and M.S. acknowledge financial support from XtalPi. M.S. acknowledges funding from ht-MATTER UKRI Frontier Research Guarantee Grant No. EP/X033139/1. M.S. and F.M.D. thank Michael Bellucci and Marcello Sega for the support and stimulating discussions and Matteo Paloni and Aaron Finney for their feedback on an early draft of the manuscript.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Florian M. Dietrich: Conceptualization (equal); Data curation (equal); Formal analysis (lead); Investigation (equal); Methodology (equal); Software (lead); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Matteo Salvalaglio**: Conceptualization (equal); Data curation (equal); Formal analysis (supporting); Funding acquisition (lead); Investigation (equal); Methodology (equal); Project administration (lead); Software (supporting); Supervision (lead); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article. The package to train GNN models of $n(q6)$ discussed in Fig. 3 can be downloaded from https://github.com/mme-ucl/NNucleate. The notebooks implementing the toy models discussed in Figs. 1 and 2 are accessible at https://github.com/mme-ucl/MLCVs_FE.

APPENDIX: GENERALIZATION OF THE GAUGE INVARIANT SAMPLE WEIGHTS

As previously mentioned, Bal *et al.* derived an expression [Eq. (11)] for modifying the Boltzmann weights w to obtain geometric free energy surfaces from enhanced sampling simulations.

This expression can be derived by substituting $m = 1$ in the ensemble average in Eq. (10),

$$\lambda \mathrm{vol}(J_{\mathbf{s}}) = \lambda \sqrt{\det J_{\mathbf{s}} J_{\mathbf{s}}^T},$$

$$J_{\mathbf{s}} J_{\mathbf{s}}^T = \begin{bmatrix} \frac{\partial s}{\partial x_1} \cdots \frac{\partial s}{\partial x_n} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial s}{\partial x_1} \\ \vdots \\ \frac{\partial s}{\partial x_n} \end{bmatrix} = \sum_i^n \frac{\partial s}{\partial x_i}^2, \tag{A1}$$

$$\lambda \mathrm{vol}(J_{\mathbf{s}}) = \lambda \sqrt{\sum_i^n \frac{\partial s}{\partial x_i}^2} = \lambda \|\nabla s\| = \lambda^1 \det d,$$

where d^2 is a matrix containing the entries[40]

$$d_{ij}^2 = \nabla s_i \nabla s_j. \tag{A2}$$

Bal generalizes this expression for higher-dimensional CV spaces as $\lambda^m \det d$, where m is the dimensionality of the CV space. However, repeating this derivation with $m = 2$ yields

$$\begin{aligned} J_{\mathbf{s}} J_{\mathbf{s}}^T &= \begin{bmatrix} \frac{\partial s_1}{\partial x_1} \cdots \frac{\partial s_1}{\partial x_n} \\ \frac{\partial s_2}{\partial x_1} \cdots \frac{\partial s_2}{\partial x_n} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial s_1}{\partial x_1} & \frac{\partial s_2}{\partial x_1} \\ & \vdots \\ \frac{\partial s_1}{\partial x_n} & \frac{\partial s_2}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} \|\nabla s_1\|^2 \sum_i^n \frac{\partial s_1}{\partial x_1} \frac{\partial s_2}{\partial x_1} \\ \sum_i^n \frac{\partial s_2}{\partial x_1} \frac{\partial s_1}{\partial x_1} \|\nabla s_2\|^2 \end{bmatrix} \\ &= \begin{bmatrix} \|\nabla s_1\|^2 (\nabla s_1 \cdot \nabla s_2) \\ (\nabla s_1 \cdot \nabla s_2) \|\nabla s_2\|^2 \end{bmatrix}, \end{aligned} \tag{A3}$$

$$\begin{aligned} \lambda \mathrm{vol}(J_{\mathbf{s}}) &= \lambda^2 \sqrt{\det J_{\mathbf{s}} J_{\mathbf{s}}^T} \\ &= \lambda^2 \sqrt{\|\nabla s_1\|^2 \|\nabla s_2\|^2 - (\nabla s_1 \cdot \nabla s_2)^2} \end{aligned} \tag{A4}$$

$$\begin{aligned} &\neq \lambda^2 \|\nabla s_1\| \|\nabla s_2\| - (\nabla s_1 \cdot \nabla s_2) \\ &= \lambda^2 \det d. \end{aligned} \tag{A5}$$

Therefore, the expression for the weights to obtain F_G from biased simulations that generalizes correctly to m-dimensions is Eq. (12).

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer Science + Business Media, LLC, 2006).

[2] G. A. Tribello, M. Ceriotti, and M. Parrinello, "Using sketch-map coordinates to analyze and bias molecular dynamics simulations," Proc. Natl. Acad. Sci. U. S. A. **109**, 5196 (2012).

[3] T. Gökdemir and J. Rydzewski, "Machine learning of slow collective variables and enhanced sampling via spatial techniques," Chem. Phys. Rev. **6**, 011304 (2025).

[4] Neha, V. Tiwari, S. Mondal, N. Kumari, and T. Karmakar, "Collective variables for crystallization simulations—From early developments to recent advances," ACS Omega **8**, 127 (2023).

[5] F. Noé and C. Clementi, "Collective variables for the study of long-time kinetics from molecular trajectories: Theory and methods," Curr. Opin. Struct. Biol. **43**, 141 (2017).

[6] H. Fu, H. Bian, X. Shao, and W. Cai, "Collective variable-based enhanced sampling: From human learning to machine learning," J. Phys. Chem. Lett. **15**, 1774 (2024).

[7] N. Naleem, C. R. A. Abreu, K. Warmuz, M. Tong, S. Kirmizialtin, and M. E. Tuckerman, "An exploration of machine learning models for the determination of reaction coordinates associated with conformational transitions," J. Chem. Phys. **159**, 034102 (2023).

[8] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," Comput. Geosci. **19**, 303 (1993).

[9] R. R. Coifman and S. Lafon, "Diffusion maps," Appl. Comput. Harmonic Anal. **21**(1), 5 (2006), a part of Special Issue: Diffusion Maps and Wavelets.

[10] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, "Systematic determination of order parameters for chain dynamics using diffusion maps," Proc. Natl. Acad. Sci. U. S. A. **107**, 13597 (2010).

[11] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map," J. Chem. Phys. **134**, 124116 (2011).

[12] W. Chen and A. L. Ferguson, "Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration," J. Comput. Chem. **39**, 2079 (2018).

[13] R. Ketkaew, F. Creazzo, and S. Luber, "Machine learning-assisted discovery of hidden states in expanded free energy space," J. Phys. Chem. Lett. **13**, 1797 (2022).

[14] J. M. L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary, "Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)," J. Chem. Phys. **149**, 072301 (2018).

[15] F. M. Dietrich, X. R. Advincula, G. Gobbo, M. A. Bellucci, and M. Salvalaglio, "Machine learning nucleation collective variables with graph neural networks," J. Chem. Theory Comput. **20**, 1600 (2024).

[16] M. M. Sultan and V. S. Pande, "tICA-metadynamics: Accelerating metadynamics by using kinetically selected collective variables," J. Chem. Theory Comput. **13**, 2440 (2017).

[17] V. Spiwok and P. Kříž, "Time-lagged t-distributed stochastic neighbor embedding (t-SNE) of molecular simulation trajectories," Front. Mol. Biosci. **7**, 132 (2020).

[18] L. Bonati, G. Piccini, and M. Parrinello, "Deep learning the slow modes for rare events sampling," Proc. Natl. Acad. Sci. U. S. A. **118**, e2113533118 (2021).

[19] P. Tiwary and B. J. Berne, "Spectral gap optimization of order parameters for sampling complex molecular systems," Proc. Natl. Acad. Sci. U. S. A. **113**, 2839 (2016).

[20] J. Rydzewski, "Spectral map: Embedding slow kinetics in collective variables," J. Phys. Chem. Lett. **14**, 5216 (2023).

[21] P. Kang, E. Trizio, and M. Parrinello, "Computing the committor with the committor to study the transition state ensemble," Nat. Comput. Sci. **4**, 451 (2024).

[22] D. Trapl, I. Horvacanin, V. Mareska, F. Ozcelik, G. Unal, and V. Spiwok, "Anncolvar: Approximation of complex collective variables by artificial neural networks for analysis and biasing of molecular simulations," Front. Mol. Biosci. **6**, 25 (2019).

[23] J. Zhang, L. Bonati, E. Trizio, O. Zhang, Y. Kang, T. Hou, and M. Parrinello, "Descriptor-free collective variables from geometric graph neural networks," J. Chem. Theory Comput. **20**, 10787 (2024).

[24] Z. Zou and P. Tiwary, "Enhanced sampling of crystal nucleation with graph representation learnt variables," J. Phys. Chem. B **128**, 3037 (2024).

[25] Z. Zou, D. Wang, and P. Tiwary, "A graph neural network-state predictive information bottleneck (GNN-SPIB) approach for learning molecular thermodynamics and kinetics," Digital Discovery **4**, 211 (2025).

[26] J. G. Kirkwood, "Statistical mechanics of fluid mixtures," J. Chem. Phys. **3**, 300 (1935).
[27] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. I. Nonpolar gases," J. Chem. Phys. **22**, 1420 (1954).
[28] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," J. Comput. Phys. **23**, 187 (1977).
[29] B. Roux, "The calculation of the potential of mean force using computer simulations," Comput. Phys. Commun. **91**, 275 (1995).
[30] A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[31] A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: A smoothly converging and tunable free-energy method," Phys. Rev. Lett. **100**, 020603 (2008).
[32] W. E and E. Vanden-Eijnden, "Metastability, conformation dynamics, and transition pathways in complex systems," in *Multiscale Modelling and Simulation*, edited by S. Attinger and P. Koumoutsakos (Springer, Berlin, Heidelberg, 2004), pp. 35–68.
[33] E. Vanden-Eijnden and F. A. Tal, "Transition state theory: Variational formulation, dynamical corrections, and error estimates," J. Chem. Phys. **123**, 184103 (2005).
[34] C. Hartmann and C. Schütte, "Comment on two distinct notions of free energy," Physica D **228**, 59 (2007).
[35] C. Hartmann, J. C. Latorre, and G. Ciccotti, "On two possible definitions of the free energy for collective variables," Eur. Phys. J. Spec. Top. **200**, 73 (2011).
[36] Imagine describing an ion-pair interaction with the distance r as the collective variable. In this case, the shell dr_1 contains a smaller fraction of configurations than the shell dr_2, with $r_1 < r_2$.
[37] A. Mikhalev and I. V. Oseledets, "Rectangular maximum-volume submatrices and their applications," Linear Algebra Appl. **538**, 187 (2018).
[38] Or, analogously, as the product of the singular values obtained by singular value decomposition of the Jacobian $\mathrm{vol}(J_s) = \prod_{i=1}^{m} \sigma_i$.
[39] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, Proceedings of the Machine Learning Research*, edited by G. Lebanon and S. V. N. Vishwanathan (PMLR, San Diego, CA, 2015), Vol. 38, pp. 192–204.
[40] K. M. Bal, S. Fukuhara, Y. Shibuta, and E. C. Neyts, "Free energy barriers from biased molecular dynamics simulations," J. Chem. Phys. **153**, 114118 (2020).
[41] A. France-Lanord, H. Vroylandt, M. Salanne, B. Rotenberg, A. M. Saitta, and F. Pietrucci, "Data-driven path collective variables," J. Chem. Theory Comput. **20**, 3069 (2024).
[42] Note that this expression differs from the two-dimensional generalization proposed by Bal *et al.* as detailed in the Appendix. Bal's proposed expression captures only partially the volume change for $m > 1$.
[43] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön, *Machine Learning—A First Course for Engineers and Scientists* (Cambridge University Press, 2022).
[44] A. R. Finney and M. Salvalaglio, "A variational approach to assess reaction coordinates for two-step crystallization," J. Chem. Phys. **158**, 094503 (2023).