Data-Driven Transport Geometry for Rare Events in Metastable Systems

Victor Kawasaki-Borruat [1] **Pierre Vandergheynst** [1]

Abstract

Rare-event transitions in metastable systems are expensive to sample because direct simulation spends most of its time in long-lived basins, and only rarely undergo transitions. We study whether Implicit Manifold-valued Diffusions (IMDs), learned directly from trajectory point clouds, can act as data-driven surrogates of the transport geometry underlying such transitions. Our central hypothesis is that when the training cloud is drawn from a biased dynamical ensemble (such as successful reactive trajectories or steady-state Langevin samples) the learned operator inherits both the support of the data and its sampling bias. On a curved transition tube and on the Müller-Brown landscape, IMDs trained on reactive data recover the dominant transition corridor and yield substantially higher short-horizon crossing probabilities than unbiased reference simulations, without access to the ambient drift. When trained on steady-state data, IMDs reproduce the coarse basin occupancy structure of the reference process. These results support IMDs as lightweight surrogates for transport geometry in metastable systems, while clarifying that they approximate coarse transport structure rather than exact kinetics or invariant laws.

1. Introduction

Rare events in stochastic physical systems are transitions between long-lived metastable states that are separated by energetic or entropic barriers, so direct simulation typically spends most of its time fluctuating inside a stable basin and only rarely observes a successful crossing (Kramers, 1940; Hänggi et al., 1990; Vanden-Eijnden et al., 2010). In the overdamped setting, this difficulty is already captured by classical barrier-crossing theory, while modern approaches such as transition-path theory emphasize that the central object is often not only a rate, but the low-dimensional transport structure organizing successful transitions through a high-dimensional landscape (Torrie & Valleau, 1977; Dellago & Bolhuis, 2008). Classical rare-event methods usually address this by explicitly biasing the dynamics or by working directly in path space.

Implicit Manifold-valued Diffusions (IMDs) offer a related yet different perspective. IMDs construct a diffusion process on the data's geometry by directly estimating its *infinitesimal generator* and associated *carré-du-champ* operators from a point cloud, without requiring explicit geometric primitives (Kawasaki-Borruat et al., 2026). In the original IMD setting, this recovers intrinsic diffusion on an implicitly sampled manifold. Here, we investigate the case of the training point cloud being drawn from a non-uniform dynamical ensemble, such as reactive trajectories (i.e. rare events) or steady-state Langevin samples. The corresponding IMD diffusion should then inherit not only the support of the data *but also its occupancy bias*. This also suggests an interpretation reminiscent of Bakry-Émery theory (Bakry et al., 2014), in which isotropic Brownian motion in a biased geometry (non-uniform medium) may be understood as overdamped Langevin-type dynamics in flat space.

The setting we target is common in modern computational physics: trajectory data may be available from simulation or experiment, while the underlying energy landscape (or drift) may be unknown or too expensive to use directly. In such cases, a learned stochastic surrogate that preserves coarse transport structure may be more valuable than an exact but costly simulator.

Contributions We study whether graph-based operator learning from trajectory point clouds can provide a useful surrogate for rare-event transport in metastable systems. Using IMDs trained on biased dynamical ensembles, we show empirically that reactive-trajectory point clouds induce transition-favoring learned dynamics, while point clouds sampled from Langevin steady-state distributions induce coarse occupancy-aware dynamics. On a curved transition tube and the Müller-Brown landscape (Müller & Brown, 1979), the learned processes approximate the dominant transition corridor without explicit access to the ambient drift or potential. We also demonstrate that IMDs can approxi-

[1]Signal Processing Laboratory 2, EPFL. Correspondence to: Victor Kawasaki-Borruat <victor.borruat@epfl.ch>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

mate coarse occupancy of Langevin steady-state ensembles, while still being agnostic to the potential. Moreover, we provide mathematical intuition towards a deeper and potentially rigorous theory of recovering steady-state ensembles through data-driven operators.

2. Related Work

Closest to our approach is graph-based operator learning under non-uniform sampling. Diffusion maps (Coifman & Lafon, 2006) established that local Markov kernels learned from point clouds can encode both intrinsic geometry and sampling density, which later work generalized to arbitrary Itô diffusions and target measures (Banisch et al., 2020). In the rare-event setting, Mahalanobis and target-measure diffusion-map methods have been used to approximate generators and compute committors from trajectory data (Evans et al., 2023; Sule et al., 2025). The transition-path theory of Gao et al. (2023) is also distinct, as we do not compute transition-path observable, but rather directly ask whether a learned diffusion can encode coarse reactive structure from the data. Our perspective differs in that we use the learned operator itself to induce a surrogate stochastic process, rather than primarily as a meshless PDE solver.

This also distinguishes our work from data-driven transition-path theory and recent neural committor-learning methods, which focus on estimating transition observables or reaction coordinates. It is likewise complementary to ML-guided path-sampling methods (Jung et al., 2023; Post & Hummer, 2025), which keep the original simulator in the loop, and to generative trajectory models such as DynamicsDiffusion (Petersen et al., 2023). Instead, we study whether a diffusion learned from biased trajectory clouds (similar to work by Lelièvre et al. (2024)) can directly inherit and reproduce coarse transport geometry.

3. Data-Induced Transport Geometry

3.1. Implicit Manifold-valued Diffusions

Implicit Manifold-valued Diffusions (IMDs) construct a diffusion process directly from a point cloud by estimating a graph-based infinitesimal generator and its *carré-du-champ* (CDC), thereby defining an ambient-space stochastic process whose local drift and noise encode intrinsic stochastic geometry (Kawasaki-Borruat et al., 2026). In particular, assuming the point cloud $X_N = \{x_i\}_{i=1}^N$ is drawn from a smooth submanifold $\mathcal{M}$ of ambient Euclidean space $\mathbb{R}^k$, we can construct a proximity graph whose nodes are X_N and edges are constructed from a cutoff kernel of fixed bandwidth $\epsilon > 0$,

$$K^{(\epsilon)}(x_i, x_j) = \mathbf{1}\{\|x_i - x_j\|_2^2 < \epsilon\} \equiv K_{i,j}^{(\epsilon)}. \tag{1}$$

We can define the node degree as

$$d_i := \sum_j K_{ij}^{(\epsilon)}, \tag{2}$$

which we use to define the α-biased walk

$$Q_{ij} := \frac{K_{ij}^{(\epsilon)}}{(d_i d_j)^\alpha} \tag{3}$$

for $\alpha \in [0, 1]$. Finally, we obtain the Markov transition matrix

$$P_{ij} := \frac{Q_{ij}}{\sum_k Q_{ik}} \tag{4}$$

which allows us to define the α-biased random walk graph Laplacian

$$L = \frac{P - I}{\epsilon} \tag{5}$$

(Chung, 1997; Coifman & Lafon, 2006).

Kawasaki-Borruat et al. (2026) work only in the case of $\alpha = 1$ which recovers geometry (Coifman & Lafon, 2006). They show that as $N \to \infty$, L converges to a second-order differential operator $c\mathcal{L}$ which is the (rescaled) *infinitesimal generator* of $B_t^{\mathcal{M}}$, Brownian motion on $\mathcal{M}$ (Stroock & Varadhan, 2007). We use the data driven generator L and compute its associated *carré-du-champ* (CDC) operator

$$\Gamma(f, g) := \frac{1}{2}\left(L(fg) - fLg - gLf\right), \tag{6}$$

which allow us to define *IMD Brownian motion*

$$dX_t = L(X_t)dt + \Gamma^{\frac{1}{2}}(X_t)dW_t, \tag{7}$$

where W_t is Brownian motion in $\mathbb{R}^k$, and $\Gamma(f) \equiv \Gamma(f, f)$.

The above construction was initially proposed as a method to simulate Brownian motion on data manifolds for exploration or sampling entire stochastic paths.

Remark 3.1. In practice, L is not defined on all of $\mathbb{R}^k$, and we thus use a nearest-neighbour (NN) or k-NN lookup (e.g. $L(X_t) \to \frac{1}{k}\sum_{i=1}^k L(x_i)$ where x_i are training points close to X_t) when numerically implementing IMDs.

3.2. Isotropy in Biased Geometry

In this work, we are interested in a different regime, in which the points in X_N are not sampled from the volume form of a manifold, but from an observational dynamical ensemble in $\mathbb{R}^k$. Inspired by Diffusion Maps (Coifman & Lafon, 2006), we choose $\alpha = \frac{1}{2}$ in all of our biased walks (3), suggesting that our graph Laplacian L captures the corresponding density bias in the large-sample limit.

Motivated by this observation, we now adopt a Bakry-Émery-inspired viewpoint on weighted measures. The main

idea consists of biasing the geometry (through the Laplacian), letting transport dynamics (i.e. drift) emerge without being explicitly enforced. Consider the Lebesgue measure dx on $\mathbb{R}^k$, and a potential $U : \mathbb{R}^k \to \mathbb{R}$. Then, the Langevin SDE driven by U

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dB_t, \tag{8}$$

has steady-state

$$\pi(x) \propto e^{-\beta U(x)}, \tag{9}$$

and is generated by

$$L := -\nabla U \cdot \nabla + \beta^{-1}\Delta, \tag{10}$$

where $\beta > 0$ is inverse temperature, and $\Delta = \text{div}\nabla$ is the Laplacian operator.

From Section 3.1, we notice that if the samples $x_i \overset{iid}{\sim} dx$ over a compact subspace $\mathcal{B} \subset \mathbb{R}^k$, then the graph Laplacian L tends to the rescaled Laplacian operator $c\Delta$ on $\mathcal{B}$. [1]

However, following Zosso et al. (2026), if the samples are *non-uniformly distributed* $x_i \sim \pi$ with π as in Eq. (9), then the produced graph Laplacian L_π should formally correspond to $-c\left(\nabla U \nabla - \beta^{-1}\Delta\right)$ on $\mathcal{B}$, i.e. it develops a deterministic drift term $-\nabla U$ in the biased geometry.[2] Therefore, the produced IMD generator L_π from non-uniform samples is posited to follow the geometry of the given dataset: if the regions are sparse, then the IMD Brownian motion will follow it naturally without needing knowledge of the drift $-\nabla U$.

We make the argument more thorough in Appendix A.

Two cases are especially relevant. In the first, the dataset consists of successful reactive trajectories between metastable basins, so the cloud is concentrated on the transition region. In the second, the dataset consists of long-time samples from a stationary Langevin process, so the cloud reflects the equilibrium occupancy of the underlying landscape. In either case, the learned IMD graph operator L should be expected to inherit not only the support of the sampled set, but also the non-uniform density induced by the data-generating process.

This motivates our working hypothesis: IMDs can serve as data-driven surrogates for transport by learning a density-biased effective geometry directly from observational data. In this view, the role of the dataset is not merely to reveal where trajectories live, but also to deform the operator that governs how they move.

[1] The assumed compactness of $\mathcal{B}$ is required for the existence of steady-state distributions.

[2] We note that our L_π is their L_μ.

4. Experiments

We now provide numerical experiments, the results of which support the previous discussion. We will concentrate on Langevin-based dynamics from Eq. (8), and show that we can obtain Langevin-like behaviour from drift-free IMDs.

4.1. Curved Transition Tube

In this section we evaluate the capability of IMDs on a 2D potential which concentrates strongly around two basins, connected by a non-trivial sinusoidal tube

$$U(x_1, x_2) = (x_1^2 - 1)^2 + \lambda(x_2 - c\sin(\omega x_1))^2, \tag{11}$$

where the first term attracts dynamics towards $x_1 = \pm 1$ and the second term constrains the x_2-coordinate around a sine curve of amplitude $c > 0$ and frequency $\omega > 0$ with strength $\lambda > 0$.

Baseline With parameters $\lambda = 250, \omega = 2, c = 0.7, \beta = 5$, an Euler-Maruyama discretization of Eq. (8) under the potential (11) yields a transition rate from $x_1 = -1$ to $x_1 = 1$ of

$$\rho_{\text{Langevin}} = 0.087, \tag{12}$$

depicted in Fig. 1. We consider a transition successful when the trajectory exits a neighbourhood of radius 0.1 around an attractor, and crosses to a neighbourhood of radius 0.1 of the other attractor.

Dataset We then construct an IMD on a downsampled subset of 5000 *reactive trajectory points*, which are all points with first coordinate $x_1 \in [-0.9, 0.9]$.

Results Upon computation of L, Γ on our reactive cloud and simulation, we compute IMD Brownian motion from Eq. (7). We report an expected higher short-horizon crossing probability along the learned geometry of

$$\rho_{\text{IMD}} = 0.11, \tag{13}$$

which we depict in Fig. 2. We summarize our comparison of transition rates and vertical residual error in Table 1.

Table 1. We observe that IMDs produce a higher empirical transition probability ρ within a lower horizon T, while also keeping a lower residual error along the x_2-axis.

	REACTIVE DATASET	IMD
HORIZON T	50	5
$\mathbb{P}$(crossing before T)	0.087	**0.11**
VERTICAL ERROR	0.024 ± 0.03	$\mathbf{0.013 \pm 0.01}$

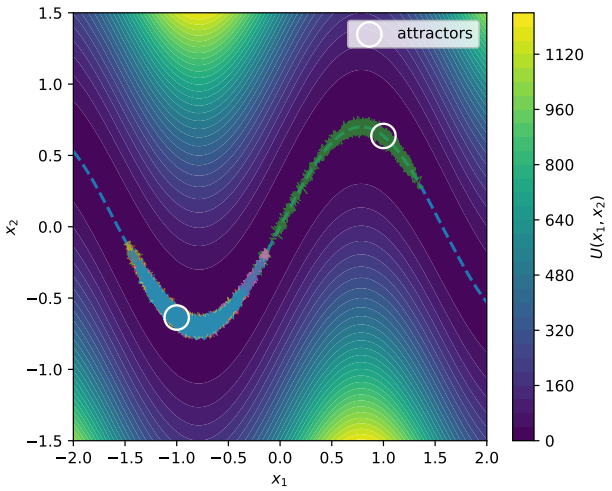

Figure 1. Baseline 1000 Langevin trajectories of step size 10^{-3}. We use one color per trajectory.

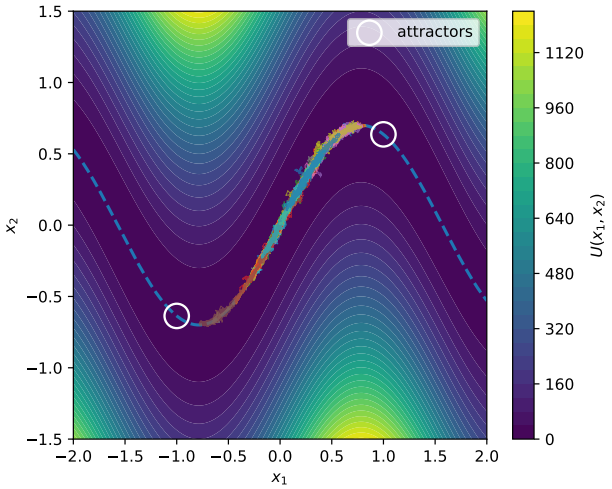

Figure 2. 1000 IMD Brownian Motions of step size 10^{-3}, effectively sampling from the S-shaped reaction tube in less time. One color per trajectory.

4.2. Müller-Brown Experiments

We now turn to the Müller-Brown potential $V(x_1, x_2)$, a canonical benchmark for metastable dynamics and rare-event transport (Müller & Brown, 1979). The landscape contains two endpoint minima connected through an ordered sequence of critical regions consisting of a first saddle, an intermediate minimum, and a second saddle, as shown in Fig. 3. This makes it a useful testbed for asking whether IMDs reflect the dominant *transport geometry* of the transition, rather than merely its endpoints.

Setup We consider transitions that start in the top-left basin (source) and terminate in the bottom-right basin (target). A trajectory is deemed successful if it leaves a neighbourhood of radius 0.3 around the source and reaches a neighbourhood of radius 0.2 around the target, as depicted in Fig. 3. The physically relevant transition pathway passes through the nearby saddle, the intermediate minimum, and the final saddle before reaching the target.

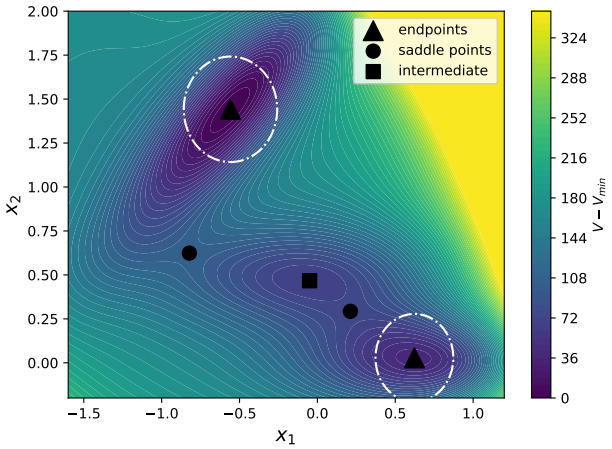

Figure 3. The Müller-Brown potential with annotated critical points and chosen acceptance radii.

4.3. Reactive Trajectory Data

We first study IMDs trained only on successful reactive transitions.

Baseline Using overdamped Langevin dynamics with inverse temperature $\beta = 0.5$, time horizon $T = 50$, and step size 10^{-4}, we observe a rare transition rate

$$\rho_{\mathrm{MB}} = 0.004 \tag{14}$$

over 1000 simulated trajectories.

Dataset From the successful Langevin transitions, we construct a reactive point cloud containing 9953 points. This cloud is concentrated on the observed transition corridor and therefore encodes the geometry of successful transport rather than the full metastable occupancy structure of the ambient process.

Results We then simulate IMD Brownian motion (7) on this reactive cloud using the same step size 10^{-4} but over a shorter horizon $T = 2.0$. The resulting IMD trajectories exhibit a much higher short-horizon transition probability,

$$\rho_{\mathrm{IMD}} = 0.506. \tag{15}$$

This shows that the learned operator strongly favors motion along the observed transition channel. Representative trajectories are shown in Fig. 4. While most trajectories follow the expected ordered corridor through the intermediate

minimum, we also observe occasional off-corridor paths. Occasional off-corridor excursions remain, but they do not alter the main qualitative finding: the learned diffusion robustly recovers the dominant reactive corridor and should therefore be interpreted as a coarse transport surrogate rather than a faithful kinetic replica.

Table 2 compares occupancy rates near the critical regions. The IMD trajectories spend substantially less time in the source basin and more time in the intermediate minimum and off-path regions than the reference reactive trajectories. We therefore interpret the reactive-cloud IMD not as a faithful surrogate for metastable residence statistics, but rather as a data-driven model of the transition geometry itself. In other words, training on successful reactive paths produces a learned diffusion that efficiently traverses the observed channel, but does not preserve the original occupancy law along that channel.

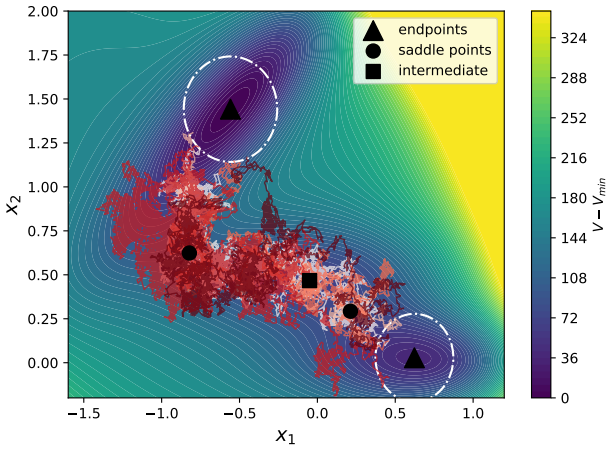

Figure 4. Five example IMD trajectories computed using the reactive dataset. We notice one unphysical trajectory (dark red) which exits the expected trajectory between critical points by avoiding the saddle points and intermediate local minimum.

Table 2. Occupancy rates of IMDs trained on the reactive pathways. We immediately notice a discrepancy between the true reactive trajectories and IMDs, the latter encouraging faster transitions between basins.

OCCUPANCIES	REACTIVE TRAJECTORIES	IMD
SOURCE ▲	0.774	0.267
INTERMEDIATE ■	0.019	0.223
TARGET ▲	0.19	0.093
OUTSIDE	0.009	0.4050

4.4. Steady-State Occupancy Geometry

We next consider a different regime, in which the training data are sampled from a steady-state Langevin ensemble rather than from successful reactive paths. The purpose of this experiment is not to test rare-event transport directly as in Section 4.3, but to determine whether IMDs trained on equilibrium trajectory clouds recover the coarse basin occupancy structure of the underlying stationary process.

Baseline & Dataset We generate Langevin trajectories on the Müller-Brown landscape using initializations in all three minima and collect a steady-state cloud of 7,345,750 points after pruning, while preserving the empirical basin proportions. For computational efficiency, we then subsample this cloud to 10,000 points and construct the IMD using the fully connected Gaussian kernel

$$K^{(\epsilon)}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{\epsilon^2}\right) \tag{16}$$

which is more stable in this highly sparse regime. This was also done in the original IMD work (Kawasaki-Borruat et al., 2026) for real-world datasets.

Results To probe the learned steady-state geometry away from a trivial initialization, we initialize 100 IMD trajectories in the neighbourhood of each minimum and run them with step size 10^{-4} up to horizon $T = 50$. After a burn-in period of 40,000 steps, we thin the resulting trajectories and estimate basin occupancies. Table 3 shows that the IMD approximately reproduces the dominant coarse occupancy structure of the steady-state dataset: the source and target basin masses are approximated reasonably well, although the intermediate minimum remains overrepresented relative to the reference distribution. We posit that this discrepancy may be due to the sparsity of the training point cloud, and the inherent very low presence of inter-basin points. Moreover, we highlight that all samples produced by IMDs remain "in-distribution", in that they remain suitably close to the empirical steady-state basins (Table 3), further supporting our effective biased geometry hypothesis.

Table 3. Coarse basin occupancies for IMDs trained on the steady-state Langevin cloud. The learned IMD approximates the dominant source/target mass split, while overestimating the occupancy of the intermediate minimum. Note that in this case, being outside may also simply mean being between basins.

	STEADY-STATE DATASET	IMD
SOURCE ▲	0.3323	0.333
INTERMEDIATE ■	0.0009	0.063
TARGET ▲	0.6597	0.603

This experiment supports the interpretation that IMDs trained on non-uniform steady-state clouds inherit a density-biased effective geometry; or run through an inhomogeneous medium. At the same time, the discrepancy in the intermediate basin indicates that the learned process should be viewed as an approximation of the coarse equilibrium geometry, rather than as an exact surrogate for the invariant law.

5. Limitations & Future Work

Our results should be interpreted with several limitations in mind. First, IMDs trained on reactive trajectories learn a transition-favoring surrogate rather than the original physical dynamics, and therefore do not preserve metastable residence statistics or the full occupancy structure of the reference process. Second, although the learned trajectories typically follow the dominant corridor, we observe occasional off-path excursions, indicating that the method is not yet a fully faithful simulator of transition dynamics. Third, our experiments are limited to low-dimensional benchmark landscapes, and the behavior of the method in higher-dimensional rare-event settings remains to be established. Finally, the weighted-geometry interpretation underlying our approach is presently formal and heuristic, not a complete theoretical guarantee.

A rigorous proof of Section 3.2 is first necessary to solidify our framework, which we then hope to apply to real-world data. We believe that IMDs could allow for a lightweight method to check for rare events (e.g. neural or molecular dynamics) as either a pre-processing or for better understanding of the underlying geometry. Moreover, producing transition-path theoretical observables using IMDs would be an another interesting direction to further ground the validity of IMDs in rare-event sampling and simulation.

6. Conclusion

We presented Implicit Manifold-valued Diffusions as an invitation to simulate rare-event systems by inheriting the biased geometry of datasets taken from dynamical systems. In one case, we demonstrated that we are accurately able to recover dominant reactive paths, opening possibilities of using IMDs for rare event generation or inpainting. Secondly, we show that we are able to heuristically approximate coarse basin occupancy through an implicit geometry bias, without knowledge of the given Langevin dynamics. We hope that this work can provide a new method to further explore rare-event analysis of implicit dynamical systems.

Acknowledgements

The authors acknowledge funding from the Swiss National Science Foundation (SNSF), grant number SNF 200021-232277.

Impact Statement

This paper presents work whose goal is to advance the field of data-driven rare event analysis. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Banisch, R., Trstanova, Z., Bittracher, A., Klus, S., and Koltai, P. Diffusion maps tailored to arbitrary non-degenerate itô processes. *Applied and Computational Harmonic Analysis*, 48(1):242–265, 2020.

Chung, F. R. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Dellago, C. and Bolhuis, P. G. Transition path sampling and other advanced simulation techniques for rare events. *Advanced computer simulation approaches for soft matter sciences III*, 221:167–233, 2008.

Evans, L., Cameron, M. K., and Tiwary, P. Computing committors in collective variables via mahalanobis diffusion maps. *Applied and Computational Harmonic Analysis*, 64:62–101, 2023.

Fukushima, M., Oshima, Y., and Takeda, M. *Dirichlet forms and symmetric Markov processes*, volume 19. Walter de Gruyter, 2011.

Gao, Y., Li, T., Li, X., and Liu, J.-G. Transition path theory for langevin dynamics on manifolds: Optimal control and data-driven solver. *Multiscale Modeling & Simulation*, 21(1):1–33, 2023.

García Trillos, N., Gerlach, M., Hein, M., and Slepčev, D. Error estimates for spectral convergence of the graph laplacian on random geometric graphs toward the laplace–beltrami operator. *Foundations of Computational Mathematics*, 20(4):827–887, 2020.

Hänggi, P., Talkner, P., and Borkovec, M. Reaction-rate theory: fifty years after Kramers. *Reviews of Modern Physics*, 62(2):251–341, April 1990. doi: 10.1103/RevModPhys.62.251.

Jung, H., Covino, R., Arjun, A., Leitold, C., Dellago, C., Bolhuis, P. G., and Hummer, G. Machine-guided path sampling to discover mechanisms of molecular self-organization. *Nature Computational Science*, 3(4):334–345, 2023.

Kawasaki-Borruat, V., Grotehans, C., Vandergheynst, P., and Gosztolai, A. Diffusion processes on implicit manifolds, 2026.

Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4): 284–304, April 1940. ISSN 0031-8914. doi: 10.1016/S0031-8914(40)90098-2.

Lelièvre, T., Santet, R., and Stoltz, G. Improving sampling by modifying the effective diffusion. preprint. *arXiv preprint arXiv:2410.00525*, 33, 2024.

Müller, K. and Brown, L. D. Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theoretica chimica acta*, 53(1): 75–93, March 1979. ISSN 1432-2234. doi: 10.1007/BF00547608.

Petersen, M., Roig, G., and Covino, R. Dynamicsdiffusion: Generating and rare event sampling of molecular dynamic trajectories using diffusion models. 2023.

Post, M. and Hummer, G. Ai-guided transition path sampling of lipid flip-flop and membrane nanoporation. *Nature Communications*, 2025.

Stroock, D. W. and Varadhan, S. S. *Multidimensional diffusion processes*. Springer, 2007.

Sule, S., Evans, L., and Cameron, M. Sharp error estimates for target measure diffusion maps with applications to the committor problem. *Applied and Computational Harmonic Analysis*, pp. 101803, 2025.

Torrie, G. M. and Valleau, J. P. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of computational physics*, 23(2): 187–199, 1977.

Vanden-Eijnden, E. et al. Transition-path theory and path-finding algorithms for the study of rare events. *Annual review of physical chemistry*, 61:391–420, 2010.

Zosso, M., Hariri, A., Kawasaki-Borruat, V., Berlureau, P.-G., and Vandergheynst, P. Geometry-induced diffusion on graphs: A learnable weighted laplacian for spectral gnns, 2026.

A. A formal argument for weighted Laplacian convergence

Here we present a formal argument supporting the discussion of Section 3.2. We highlight that the following is *not* a direct corollary of García Trillos et al. (2020), since their analysis is carried out for the isotropic kernel; the argument below indicates what continuum operator should arise after diffusion-maps α-renormalization.

A.1. IMD Framework

The IMD framework proves convergence of the random walk graph Laplacian in the case where the data points $\{x_i\}_{i=1}^N$ are iid distributed according to the volume form dvol_g of a compact boundaryless manifold $\mathcal{M}$ with metric g. The authors then construct a random walk graph Laplacian using the kernel from Eq. (1) but use the normalization of $\alpha = 1$. This is the argued construction to be "geometry only" in Diffusion Maps (Coifman & Lafon, 2006), in which even data sampled from a distribution with a non-constant density with respect to the volume form recovers the Laplace-Beltrami operator.

In (Kawasaki-Borruat et al., 2026), the data is assumed to have constant density ($\rho(x_i) = 1$) for the sampling distribution $\mu(x) = \rho(x)\text{dvol}_g$. The main proof technique invokes the construction from García Trillos et al. (2020), which also proves that different normalizations of the random walk constructed in (3) lead to different limit operators on $\mathcal{M}$. Here, we will formally argue that applying the α-renormalization Diffusion Map of Eq. (3) in the García Trillos et al. (2020) construction formally suggests a weighted Laplacian of the form (10), which is interpretable as a Bakry-Emery Laplacian in a weighted geometry $(\mathcal{M}, e^{-V(x)}\text{dvol}_g(x))$ (Bakry et al., 2014; Zosso et al., 2026).

A.2. Formal Construction of Langevin Generator

In our case, we are not considering distributions on a Riemannian manifold $\mathcal{M}$, but rather a probability distribution on a compact subspace $\mathcal{B} \subset \mathbb{R}^k$. We furthermore assume vanishing boundary conditions for test functions $f, g \in \mathcal{C}^\infty(\mathcal{B}, \mathbb{R})$. The framework from García Trillos et al. (2020) requires an ϵ-bandwidth kernel

$$\eta\left(\frac{\|x_i - x_j\|^2}{\epsilon}\right), \tag{17}$$

which we can automatically renormalize

$$w_{ij}^{(\alpha)} \propto \frac{\eta\big(\|x_i - x_j\|^2/\varepsilon\big)}{\hat{p}_{\epsilon,N}(x_i)^\alpha \hat{p}_{\epsilon,N}(x_j)^\alpha},$$

for $\hat{p}_{\epsilon,N}(x_j) := \frac{\eta(x_i,x_j)}{\sum_i \eta(x_i,x_j)}$. Here $\widehat{p}_{\epsilon,N}$ denotes the empirical kernel density estimate used in the finite-sample renormalization, while p denotes the continuum sampling density. Thus the replacement $\hat{p}_{\epsilon,N} \rightsquigarrow p$ in the formulas below is a formal large-sample/small-bandwidth passage.

Following their intuition, we observe that this random walk operator has a limiting bilinear form (a Dirichlet form D_α (Fukushima et al., 2011)) which admits the continuous representation

$$D_\alpha(f,g) := \int_{\mathcal{B}} \langle \nabla f, \nabla g \rangle p^{2-2\alpha} \text{dvol}_g, \tag{18}$$

which uniquely defines a second-order symmetric operator Δ_α such that

$$D_\alpha(f,g) = -\int_{\mathcal{B}} \Delta_\alpha f g p^{2-2\alpha} \text{dvol}_g, \tag{19}$$

which is explicitly given as

$$\Delta_\alpha f = \frac{1}{p^{2-2\alpha}} \text{div}\left(p^{2-2\alpha}\nabla f\right). \tag{20}$$

The reference papers (García Trillos et al., 2020; Kawasaki-Borruat et al., 2026) only consider isotropic walks, matching the normalization with $\alpha = 1$ from Diffusion Maps. Now we observe that for $\alpha = \frac{1}{2}$, we obtain by Theorem 1 of (Coifman & Lafon, 2006)

$$\Delta_{\frac{1}{2}} f = \beta \Delta f + \nabla \log p \cdot \nabla f. \tag{21}$$

If the sampling density is of the form $p(x) \propto e^{-\beta U(x)}$, then $\nabla \log p = -\beta \nabla U$, so that

$$\beta^{-1}\Delta_{1/2} f = \beta^{-1}\Delta f - \nabla U \cdot \nabla f. \tag{22}$$

This is exactly the overdamped Langevin generator from Eq. (10). Therefore, the continuum operator associated with the $\alpha = \frac{1}{2}$ renormalization matches the Langevin generator up to the expected scalar time-rescaling by β^{-1}.

A.3. Bakry-Emery Perspective

Recent work in graph-based learning suggests viewing drift-diffusion as arising from a weighted geometry (Zosso et al., 2026). Let

$$\mu(x) = e^{-\beta U(x)}\, dx,$$

and define the Dirichlet form (Fukushima et al., 2011)

$$D_\mu(f, g) := \int_B \mu\, \nabla f \cdot \nabla g\, dx.$$

Assuming suitable boundary conditions, integration by parts gives

$$D_\mu(f, g) = -\int_B f \operatorname{div}(\mu \nabla g)\, dx = -\int_B \mu f\, (\Delta g + \nabla \log \mu \cdot \nabla g)\, dx.$$

Hence the associated symmetric generator is

$$L_\mu g := \Delta g + \nabla \log \mu \cdot \nabla g.$$

Since $\mu(x) \propto e^{-\beta U(x)}$, we have

$$\nabla \log \mu = -\beta \nabla U,$$

and therefore

$$L_\mu g = \Delta g - \beta \nabla U \cdot \nabla g.$$

Equivalently,

$$\beta^{-1} L_\mu g = \beta^{-1}\Delta g - \nabla U \cdot \nabla g,$$

which coincides with the Langevin generator in Eq. (10). Thus, the weighted-geometry viewpoint recovers the same drift-diffusion operator, again up to the scalar time-rescaling factor β^{-1}.

A.4. Interpretation of the IMD generator

We thus interpret the graph-based learned IMD generator from Eq. (7) as formally converging, in the infinite-sample limit, to the weighted operator $\Delta_{1/2}$, or equivalently to the overdamped Langevin generator up to a positive scalar time-rescaling. This offers a good understanding of why we *do not need to know the drift to simulate IMDs*, as it is already incorporated in the biased generator. We are simply letting it emerge from the sample density.