Mining Geometric Patterns in Protein Structures Using Graph-Based Lévy Diffusion Models

Eric Paquet[1,2] and Herna Viktor[2] and Wojtek Michalowski[3]

[1] National Research Council of Canada, Ottawa, Ontario, Canada
ericpaquet@nrc-cnrc.gc.ca
[2] School of Electrical Engineering and Computer Science, University of Ottawa. Ottawa, Ontario, Canada
hviktor@uottawa.ca
[3] Telfer School of Management, University of Ottawa. Ottawa, Ontario, Canada
wojtek@telfer.uottawa.ca

Abstract. The process of extracting intricate geometric patterns from biological structures poses a significant challenge, primarily due to the fact that protein conformations manifest non-Euclidian geometry, exhibit stringent structural constraints, and are characterized by multi-scale dependencies. Existing diffusion-based generative models frequently employ Gaussian perturbations, which have the potential to restrict the exploration of highly multimodal structural landscapes while concurrently compromising geometric consistency. To address this limitation, we present a graph-based Lévy diffusion framework for learning and generating complex geometric patterns in protein structures. Protein geometry is represented through graph Laplacian operators and incorporated directly into a geometry-aware score function. To improve exploration of the pattern space, conventional Gaussian noise is replaced with Lévy–Itō diffusion dynamics, enabling occasional long-range transitions while preserving local geometric structure. This combination of graph-based inductive biases and heavy-tailed stochastic exploration enables the generation of structurally plausible samples from complex biological distributions. Experiments on protein structures from the Protein Data Bank demonstrate the ability to accurately reconstruct protein distance maps while maintaining controllable generative diversity. The results indicate that our method effectively balances reconstruction fidelity and structural variability, suggesting a promising approach for pattern discovery and generative modeling in biological and other structured scientific datasets.

Keywords: Lévy distribution · bifractional Fokker–Planck equation · Lévy–Itō diffusion model · score function

1 Introduction

The discovery and generation of complex structural patterns in biological data are central challenges in computational biology and drug design. Protein structures are particularly difficult to model because they exhibit non-Euclidean geometry,

strong structural constraints, and dependencies spanning multiple spatial scales. Learning such patterns is important for applications including protein engineering, molecular design, and therapeutic discovery. Recent score-based diffusion models have achieved remarkable success in learning complex data distributions [4]. However, their application to geometric biological data remains challenging because standard diffusion processes rely on Gaussian perturbations that can distort geometric structure and limit exploration of highly multimodal pattern spaces. As a result, maintaining geometric consistency while preserving generative diversity remains an open problem for protein structure generation and related tasks.

In this paper, we propose a geometry-aware score function for use in such domains based on the graph Laplacian and a Lévy–Itō diffusion process. The Laplacian allows the score function to exploit the geometry, while the Lévy–Itō diffusion process maintains high fidelity while enabling diverse sampling. Specifically, the process allows for intermittent long-distance steps during diffusion, resulting in more efficient movements within the distribution of the geometry. As such, we introduce a method that bridges the gap between expressive generative modelling and the strict structural requirements of geometric data by directly incorporating inductive biases into the score estimation procedure. This model better respects intrinsic constraints and multi-scale dependencies during generation, while continuing to efficiently explore complex structural spaces.

The paper is structured as follows. Section 2 introduces the Lévy distribution, the bifractional Fokker–Planck equation, and the heat kernel. Section 3 derives the geometry-aware score function based on the graph Laplacian and Lévy–Itō diffusion stochastic differential equations (SDEs). Section 4 presents the experimental results, and Section 5 concludes the paper. All symbols appearing in this paper are defined in Appendix A.

2 The Lévy Distribution, the Bifractional Fokker–Planck Equation, and the Heat Kernel

Diffusion models generate samples by progressively transforming simple noise into structured data through a sequence of denoising steps. As a result, the choice of noise distribution strongly influences how effectively the model explores the underlying pattern space. Most score-based diffusion models assume Gaussian noise, leading to local and incremental updates throughout the sampling process. While effective in many settings, this assumption can be limiting for complex biological and geometric data, where the underlying distributions are highly constrained, multimodal, and characterized by long-range structural dependencies. Examples include protein structures, protein-ligand complexes, and molecular conformations, where effective pattern discovery and generation require exploration beyond local geometric variations.

Unlike the Gaussian distribution, which decreases exponentially for large values, the Lévy distribution [1] is only suppressed by a polynomial factor. This results in a heavy tail, meaning that large values have a non-negligible probability

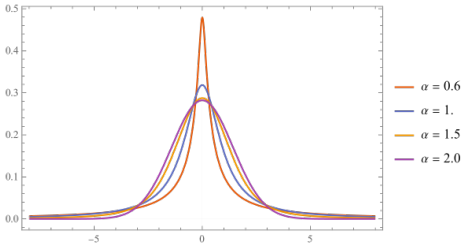

Fig. 1. Lévy distribution for various values of the stability index.

of occurring:

$$p_L(x;\alpha,\gamma,\mu,\sigma) \sim \left.\frac{C_\pm}{|x|^{1+\alpha}}\right|_{x\to\pm\infty}. \tag{1}$$

An advantage of incorporating heavy-tailed Lévy noise is that it allows for occasional large jumps during diffusion, enabling the sampler to move efficiently between distant regions of the geometric distribution, rather than settling for small local updates. This facilitates exploration of highly multimodal spaces and avoids mode collapse, leading to greater diversity in the generated samples.

Formally, the Lévy distribution does not have a closed form; rather, it is defined by its Fourier transform, or characteristic function:

$$p_L(x;\alpha,\gamma,\mu,\sigma) = \int_{-\infty}^{\infty} \varphi(k;\alpha,\gamma,\mu,\sigma)e^{-2\pi ikx}\,dk, \quad i=\sqrt{-1}, \quad \alpha\in(0,2]. \tag{2}$$

where

$$\varphi(k;\alpha,\gamma,\mu,\sigma) = \begin{cases} \exp\Big[ik\mu - |k\sigma|\Big(1+\dfrac{2i\gamma}{\pi}\operatorname{sgn}(k)\ln|k\sigma|\Big)\Big] & , \quad \alpha=1, \\ \exp\Big[ik\mu - |k\sigma|^\alpha\Big(1+i\gamma\tan\Big(\dfrac{\pi\alpha}{2}\Big) & \\ \qquad \times \operatorname{sgn}(k)\big(|k\sigma|^{1-\alpha}-1\big)\Big)\Big] & , \quad \alpha\neq 1. \end{cases} \tag{3}$$

The Lévy distribution has four parameters: a skewness parameter, a location parameter, a dispersion parameter, and a stability index. The parameters determine the degree of polynomial suppression. In our setting, these parameters provide direct control over the stochastic behaviour of the Lévy–Itō diffusion, allowing us to regulate the magnitude and asymmetry of perturbations during sampling and thereby optimise the trade-off between reconstruction fidelity and generative diversity. The Lévy distribution is illustrated in Fig. 1 for various values of the stability index, and the corresponding heavy-tailed distributions are shown in Fig. 2. The Lévy distribution reduces to the Gaussian distribution when the stability index is equal to two.

A diffusion process can be viewed as a random walk associated with a probability distribution. The resulting behaviour is given by the Kolmogorov–Chapman

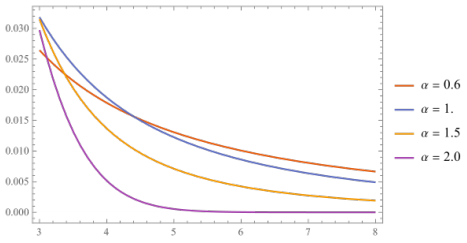

Fig. 2. The heavy tails of the Lévy distribution for various values of the stability index.

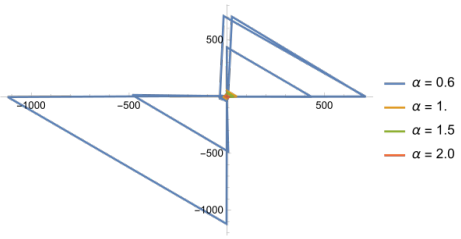

Fig. 3. Lévy random walks for various values of the stability index.

equation. In the case of a Gaussian distribution, the Fokker–Planck equation [2] is obtained. If a Lévy distribution is used instead, the outcome is the fractional Fokker–Planck equation [2]. Fig. 3 illustrates that lower values of the index allow for more thorough exploration of the solution space. Such exploration is particularly desirable in biological applications, where important structural patterns may occupy only a small fraction of the overall configuration space. A non-Markovian random walk with memory can be achieved by introducing a fractional time derivative into the inhomogeneous fractional Fokker–Planck equation, resulting in a bifractional Fokker–Planck equation:

$$\partial_t^\beta u(x,t) - u(x,t_0)\frac{t^{-\beta}}{\Gamma(1-\beta)} = \kappa \Delta^{\alpha/2} u(x,t). \tag{4}$$

The fractional time derivative is defined as

$$\partial_t^\beta f(t) \triangleq \frac{1}{\Gamma(-\beta)} \int_{-\infty}^{t} \frac{f(\tau)}{(t-\tau)^{1+\beta}}\, d\tau, \tag{5}$$

whereas the fractional Laplacian is defined in terms of its Fourier transform:

$$\Delta^{\alpha/2} \cdot = \int_{-\infty}^{\infty} d\mathbf{k}\, \exp(-2\pi i \mathbf{k}\cdot\mathbf{x})\, \|\mathbf{k}\|^\alpha \cdot . \tag{6}$$

The fractional time becomes non-fractional when the fractional time index is 0.5. Notably, the stability index is related to the fractal dimension of the process, meaning that the bifractional Fokker–Planck equation is fractal by nature:

$$d_F = \alpha, \qquad 1 < \alpha \leq 2. \tag{7}$$

The bifractional heat kernel is the solution of the homogeneous bifractional Fokker–Planck equation [2]:

$$\partial_t^\beta K_{\alpha,\beta}(x,x_0;t) - K_{\alpha,\beta}(x,x_0;t_0)\frac{t^{-\beta}}{\Gamma(1-\beta)} - \kappa\Delta^{\alpha/2}K_{\alpha,\beta}(x,x_0;t) \tag{8}$$

$$= \delta(x-x_0)\delta(t-t_0). \tag{9}$$

The solution may be obtained from the spectral decomposition of the Laplacian:

$$K_{\alpha,\beta}(x,x_0;t) = \sum_{n=0}^{\infty}\phi_n(x)\phi_n(x_0)E_\beta\left(-t^\beta\lambda_n^{\alpha/2}\right), \tag{10}$$

$$\Delta\phi_n(x) = \lambda_n\phi_n(x). \tag{11}$$

The temporal evolution is determined by the Mittag–Leffler function,

$$E_\beta(x) \triangleq \sum_{k=0}^{\infty}\frac{x^k}{\Gamma(\beta k+1)}. \tag{12}$$

The corresponding discrete bifractional Fokker–Planck equation is obtained by replacing the continuous Laplacian with the graph Laplacian:

$$\partial_t^\beta \mathbf{u} - \mathbf{u}_0\frac{t^{-\beta}}{\Gamma(1-\beta)} = \kappa(\Delta_\mathcal{G})^{\alpha/2}\mathbf{u}, \tag{13}$$

$$\mathbf{u}_0, \mathbf{u} \in \mathbb{R}^N, \qquad \Delta_\mathcal{G} \in \mathbb{R}^{N\times N}. \tag{14}$$

The resulting spectral representation provides a natural multi-scale description of geometric structure, making it particularly suitable for learning complex patterns in graph-structured biological data. As demonstrated in [3], the graph Laplacian may be defined as

$$\Delta_\mathcal{G} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{1/2}, \qquad \mathbf{I},\mathbf{D},\mathbf{W} \in \mathbb{R}^{N\times N}, \tag{15}$$

$$\Delta_\mathcal{G} = (\Delta_\mathcal{G})^T \Rightarrow \Lambda \in \mathbb{R}^{N\times N}, \quad \Phi^T\Phi = \Phi\Phi^T. \tag{16}$$

The weights and degrees are determined by the Euclidean distance between the points forming the geometry:

$$\mathbf{W} = [w_{ij}], \qquad w_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2, \tag{17}$$

$$\mathbf{D} = \mathrm{diag}[d_i], \qquad d_i = \sum_j w_{ij}. \tag{18}$$

As the graph Laplacian is symmetric, its eigenvectors are orthogonal. This enables the fractional Laplacian to be evaluated in a computationally efficient manner:

$$(\Delta_\mathcal{G})^{\alpha/2} = \Phi\Lambda^{\alpha/2}\Phi^T, \qquad \mathcal{O}(|\mathcal{E}|). \tag{19}$$

Equation above and the discrete counterpart of the heat equation are used to find the solution to the homogeneous bifractional heat kernel:

$$\mathbf{K}_{t;\alpha,\beta} = \sum_{n=0}^{N-1} \phi_n \phi_n^T E_\beta \left(-\kappa t^\beta \lambda_n^{\alpha/2}\right) \tag{20}$$

$$= \Phi E_\beta \left(-\Lambda^{\alpha/2} t^\beta\right) \Phi^T, \qquad \mathbf{K}_{t;\alpha,\beta} \in \mathbb{R}^{N\times N}. \tag{21}$$

This gives the closed-form transition density of the forward Lévy–Itō diffusion and provides the fundamental object needed to define the graph noise corruption process and derive the reverse-time score-based generative dynamics. This transition density forms the basis for learning and generating geometric patterns within the underlying structural distribution.

3 A Geometry-Aware Score Function

Recall that a diffusion model consists of a fixed forward process that adds noise to the data and a learned backward denoising process that iteratively removes noise from the data [5]. From a pattern-mining perspective, diffusion models learn the underlying distribution of complex structured data and generate new samples by progressively recovering geometric and relational patterns from noise. For Gaussian noise, the noising and denoising processes are given in their simplest form by the following SDEs:

$$d\mathbf{x}_t = \mathbf{x}_t dt + d\mathbf{w}_t, \tag{22}$$

$$d\mathbf{x}_t = \left[\mathbf{x}_t - \nabla_{\mathbf{x}_t} p_t(\mathbf{x}_t)\right] dt + d\mathbf{w}_t. \tag{23}$$

The score function cannot be evaluated directly; it must be evaluated using a score matching technique [6,8,12] that employs transition probabilities instead of non-tractable data distributions:

$$\mathbb{E}_{p(\mathbf{x}_t)} \left[\frac{1}{2} \left\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t)\right\|^2\right] \tag{24}$$

$$= \mathbb{E}_{p(\mathbf{x}_0) p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)} \left[\frac{1}{2} \left\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)\right\|^2\right] + \Omega, \tag{25}$$

$$\mathbf{x}_t = \left[\mathbf{x}_t - \mathbf{s}_\theta(\mathbf{x}_t, t)\right] dt + d\mathbf{w}_t. \tag{26}$$

The transition probability is given by a heat kernel, for which the gradient may be evaluated [11]:

$$p_{t|0}(\mathbf{x}_t|\mathbf{x}_0) \equiv K(x, x_0; t) = \sum_{n\in\mathbb{N}} \exp(-\lambda_n t)\phi_n(\mathbf{x}_0)\phi_n(\mathbf{x}_t), \tag{27}$$

$$\nabla_{\mathbf{x}_t} p_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = \frac{\sum_{n=0}^{N-1} \exp(-\lambda_n t)\phi_n(\mathbf{x}_0)\nabla_{\mathbf{x}_t}\phi_n(\mathbf{x}_t)}{\sum_{n=0}^{N-1} \exp(-\lambda_n t)\phi_n(\mathbf{x}_0)\phi_n(\mathbf{x}_t)}. \tag{28}$$

Replacing Gaussian noise with Lévy noise yields a Lévy–Itō diffusion process [14]. The SDEs for noising and denoising are similar in form to their Gaussian counterparts, but the score involves calculating the fractional Laplacian of the gradient of the bifractional heat kernel and then taking its inverse:

$$d\mathbf{x}_t = \mathbf{x}_t dt + d\mathbf{L}_{t;\alpha}, \tag{29}$$

$$d\mathbf{x}_t = \left[\mathbf{x}_t - \alpha \frac{\Delta^{(\alpha-2)/2}\nabla_{\mathbf{x}_t} p_{t|0;\alpha,\beta}(\mathbf{x}_t|\mathbf{x}_0)}{p_{t|0;\alpha,\beta}(\mathbf{x}_t|\mathbf{x}_0)}\right] dt + d\mathbf{L}_{t;\alpha}, \tag{30}$$

$$p_{t|0;\alpha,\beta}(\mathbf{x}_t|\mathbf{x}_0) \equiv K_{\alpha,\beta}(x, x_0; t) = \sum_{n=0}^{N-1} E_\beta\left(-\lambda_n^{\alpha/2} t^\beta\right) \phi_n(\mathbf{x}_0)\phi_n(\mathbf{x}_t). \tag{31}$$

The fractional score function is analytically intractable. Consequently, the gradient of an eigenvector at a given point is approximated by the expectation of the gradient on the first neighbourhood surrounding the point [15]:

$$\nabla\phi_i \Rightarrow \langle \nabla\phi_i \rangle_{j\in\mathcal{N}(i)} \approx \Delta_{\mathcal{G}}\phi_i. \tag{32}$$

Then the fractional score function may be evaluated as

$$\Delta^{(\alpha-2)/2}\nabla_{\mathbf{x}_t} p_{t|0;\alpha,\beta}(\mathbf{x}_t|\mathbf{x}_0) \Rightarrow \Delta^{\alpha/2} p_{t|0;\alpha,\beta}(\mathbf{x}_t|\mathbf{x}_0) \tag{33}$$

$$= \frac{\sum_{n=0}^{N-1} E_\beta(-\lambda_n^{\alpha/2} t^\beta)\phi_n(\mathbf{x}_0)\Delta^{\alpha/2}\phi_n(\mathbf{x}_t)}{\sum_{n=0}^{N-1} E_\beta(-\lambda_n^{\alpha/2} t^\beta)\phi_n(\mathbf{x}_0)\phi_n(\mathbf{x}_t)}. \tag{34}$$

The discrete fractional score function [9] is obtained similarly:

$$\mathbf{S}_{t;\alpha,\beta} = \Delta_{\mathcal{G}}^{(\alpha-2)/2}\nabla_{\mathcal{G}}\mathbf{K}_{t;\alpha,\beta} \tag{35}$$

$$\Rightarrow \Delta_{\mathcal{G}}^{(\alpha-2)/2}\Delta_{\mathcal{G}}\mathbf{K}_{t;\alpha,\beta} = \Delta_{\mathcal{G}}^{\alpha/2}\mathbf{K}_{t;\alpha,\beta}, \tag{36}$$

$$\mathbf{S}_{t;\alpha,\beta} = \frac{\Phi E_\beta(-\Lambda^{\alpha/2} t^\beta)\Lambda^{\alpha/2}\Phi^T}{\Phi\left(E_\beta(-\Lambda^{\alpha/2} t^\beta)\right)^2\Phi^T}. \tag{37}$$

To ensure numerical stability, the denominator is replaced with Penrose's pseudo-inverse:

$$\Delta_{\mathcal{G}}^{(\alpha-2)/2}\nabla_{\mathcal{G}}\mathbf{K}_{t;\alpha,\beta} \approx \left(\Phi E_\beta(-\Lambda^{\alpha/2} t^\beta)\Lambda^{\alpha/2}\Phi^T\right)\left(\Phi E_\beta(-\Lambda^{\alpha/2} t^\beta)\Phi^T\right)^+. \tag{38}$$

The forward noising SDE and reverse denoising SDE are integrated using the exponential Euler–Maruyama method [7]:

$$\mathbf{X}_{t+\Delta t} = \left[\mathbf{A}(t)\mathbf{X}_t + \mathcal{L}(\mathbf{X}_t)\right]\Delta t + \mathbf{B}(\mathbf{X}_t, t)\Delta\mathbf{L}_{t;\alpha}, \tag{39}$$

$$\mathbf{X}_{t+\Delta t} = \exp[\mathbf{A}(t)\Delta t]\left[\mathbf{X}_t + \mathcal{L}(\mathbf{X}_t)\Delta t + \mathbf{B}(\mathbf{X}_t, t)\Delta\mathbf{L}_{t;\alpha}\right]. \tag{40}$$

Applied to the Lévy–Itō forward and backward equations, this results in

$$\mathbf{X}_{t+\Delta t} = \mathbf{X}_t\Delta t + \Delta\mathbf{L}_{t;\alpha} \Rightarrow \exp[\mathbf{I}(t)\Delta t]\left[\mathbf{X}_t + \Delta\mathbf{L}_{t;\alpha}\right], \tag{41}$$

$$\mathbf{X}_{t-\Delta t} = \left[\mathbf{X}_t - \alpha\mathbf{S}_{t;\alpha,\beta}\right]\Delta t + \Delta\mathbf{L}_{t;\alpha} \Rightarrow \exp[-\mathbf{I}(t)\Delta t]\left[\mathbf{X}_t + \alpha\mathbf{S}_{t;\alpha,\beta}\Delta t + \Delta\mathbf{L}_{t;\alpha}\right], \tag{42}$$

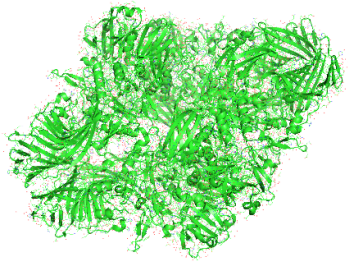

Fig. 4. The backbone of protein 1DP0 of the PDB.

where the Lévy process is given by

$$\Delta \mathbf{L}_{t;\alpha} \sim \Big[(\Delta t)^{1/\alpha} p_L(\alpha, 0, 0, 1)\Big]_{N\times N}. \tag{43}$$

These equations define the forward and reverse diffusion processes used to learn, reconstruct, and generate complex geometric patterns in protein structures.

4 Experimental Results

We evaluated the proposed framework on protein structures from the Protein Data Bank (PDB) [13]. These structures exhibit complex geometric organization, long-range dependencies, and substantial variability across folds and structural classes, making them a challenging benchmark for pattern learning and generative modeling. The experiments were designed to assess the ability of the proposed graph-based Lévy diffusion framework to capture, reconstruct, and generate biologically relevant geometric patterns while preserving structural consistency. An example of the proteins used is shown in Fig. 4, corresponding to PDB entry 1DP0. Protein structures were represented by pairwise alpha-carbon distance matrices [10], which provide a compact and rotation-invariant representation of protein geometry. As structured geometric patterns, these matrices encode both local interactions and long-range dependencies, enabling the proposed graph-based diffusion model to learn, reconstruct, and generate biologically meaningful structural patterns across multiple scales. The distance matrix corresponding to protein 1DP0 is shown in Fig. 5. We selected proteins with a range of amino acid lengths to demonstrate the method's versatility. The distribution of amino acid numbers is shown in Fig. 6.

The time scale for the diffusion process spans the range from 0 to 10. Noising was simulated with a step size of 0.05, resulting in 200 steps over the full time horizon. Denoising was simulated using a larger step size of 0.5, requiring only 20 steps to traverse the same interval. This reduction in denoising steps was enabled by the improved stability of the exponential Euler–Maruyama integration scheme.

During denoising, the stability index was increased from a low value, which promotes exploration via heavy-tailed fluctuations, to a high value, which pro-

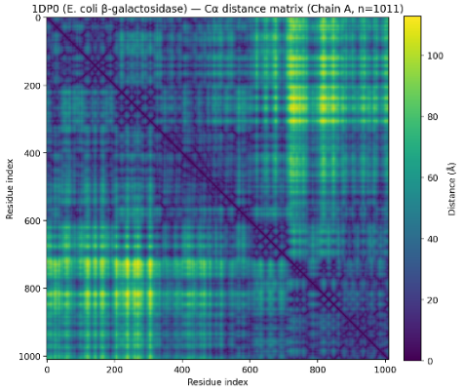

Fig. 5. The distance matrix associated with protein 1DP0 from the PDB.

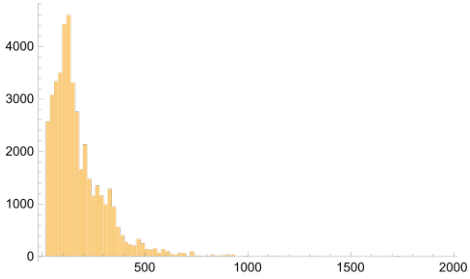

Fig. 6. The distribution of amino acid numbers in the dataset.

gressively suppresses large jumps and yields near-Gaussian, small-step updates:

$$\alpha(t) = \alpha_{\min} + \frac{\alpha_{\max} - \alpha_{\min}}{T} t. \tag{44}$$

The efficiency of the denoising process was evaluated using two metrics: the relative Frobenius error, which is equivalent to the relative root-mean-square error,

$$\delta_F = \frac{\|\mathbf{D} - \mathbf{D}_0\|_F}{\|\mathbf{D}_0\|_F}, \tag{45}$$

$$\|\mathbf{D}\|_F = \sqrt{\mathrm{tr}(\mathbf{D}\mathbf{D}^T)} = \sqrt{\sum_{i,j=1}^{N} d_{ij}^2}, \tag{46}$$

and the absolute error:

$$\delta_{|\cdot|} = \frac{\frac{1}{N^2}\sum_{i,j=1}^{N} |d_{ij} - d_{0;ij}|}{\frac{1}{N^2}\sum_{i,j=1}^{N} |d_{0;ij}|}. \tag{47}$$

Table 1. Relative Frobenius error and relative absolute error for various values of the stability index and the time index.

α	β	$\delta_F(\%)$	$\delta_{\lvert\cdot\rvert}(\%)$
2	1	0.214 ± 0.083	0.217 ± 0.083
1.75	1	0.340 ± 2.032	0.325 ± 2.039
1.5	1	2.529 ± 10.517	2.522 ± 10.917
1.25	1	15.648 ± 26.653	15.742 ± 27.562
2	0.75	0.064 ± 0.019	0.056 ± 0.017
1.75	0.75	0.331 ± 2.569	0.308151 ± 2.65334
1.5	0.75	2.523 ± 10.088	2.495 ± 10.397
1.25	0.75	15.970 ± 27.624	16.135 ± 28.707
2	0.5	0.064 ± 0.018	0.056 ± 0.016
1.75	0.5	0.501 ± 3.914	0.482 ± 4.013
1.5	0.5	2.433 ± 9.629	2.416 ± 9.968
1.25	0.5	16.344 ± 27.580	16.509 ± 28.612

The calculations were conducted on a computing node comprising two Intel Xeon Gold 6130 CPUs with 16 cores operating at 2.1 GHz, 192 GB of RAM, and two Nvidia V100 GPUs with 32 GB of RAM each. The code was written in Mathematica. The results are presented in Table 1.

Regardless of the fractional time index, reconstruction performance deteriorated when the stability index diminished. Optimal performance occurred when the stability index was equal to 2, corresponding to a Wiener process. For the fractional time index, performance improved substantially when the stability index was equal to 0.5, i.e., when the process effectively becomes Markovian. When both the stability and fractional time indices were equal to 2 and 0.5, respectively, the reconstruction error was only $0.064\% \pm 0.018$ for the Frobenius error and $0.056\% \pm 0.016$ for the absolute error. The high degree of similarity between these two metrics suggests that outliers were minimal. For comparison, the largest error was generally within 15–16% for a stability index of 1.25.

Although the error was lower with a Wiener process and non-fractional time, the uncertainty remained extremely low, suggesting overfitting. Conversely, using a stability index below 2, together with fractional time, allows explicit control of the fidelity-versus-diversity trade-off by increasing uncertainty while keeping reconstruction error low. For example, setting the stability index to 1.75 and the fractional time index to 1 yielded a Frobenius error of $0.340\% \pm 2.032$, indicating high-fidelity reconstruction with meaningful sample variability. If greater diversity is required, the stability index can be lowered to 1.5, further increasing uncertainty and broadening exploration of the solution space.

The experiments reveal a clear trade-off between reconstruction fidelity and pattern-space exploration. Higher stability indices produce highly accurate reconstructions but generate relatively limited variability. In contrast, lower stability indices introduce heavier-tailed perturbations that broaden exploration of the structural pattern space, resulting in more diverse generated geometries while

maintaining acceptable reconstruction accuracy. From a pattern-mining perspective, the proposed framework learns geometric structure through graph spectral representations while using heavy-tailed diffusion dynamics to explore rare and potentially informative regions of the structural distribution. The combination of graph-based geometric priors and Lévy diffusion therefore provides a unified mechanism for both pattern representation and pattern generation in complex biological data.

5 Conclusions

We have introduced a graph-based Lévy diffusion framework for learning and generating complex geometric patterns in protein structures. By integrating graph Laplacian representations into a geometry-aware score function, the proposed approach preserves structural relationships while enabling efficient exploration through heavy-tailed Lévy–Itō dynamics. Experimental results demonstrate that the method accurately captures geometric patterns while maintaining meaningful diversity in generated samples. These findings suggest that heavy-tailed graph diffusion provides an effective mechanism for exploring complex structural pattern spaces and generating biologically plausible protein geometries. More broadly, this work highlights the potential of graph-based diffusion models for pattern discovery and generative modeling in structured scientific data. Future work will focus on scalable graph diffusion architectures and applications to larger biological networks and other graph-structured datasets.

Acknowledgment The authors thank the Artificial Intelligence for Design Challenge program of the National Research Council of Canada for funding.

References

1. P. Lévy, "Sur les intégrales dont les éléments sont des variables aléatoires indépendantes," Ann. Sc. Norm. Super. Pisa, vol. 3(3–4), pp. 337–366, 1934.
2. B. West, M. Bologna, and P. Grigolini, Physics of Fractal Operators. Berlin, Germany: Springer, 2012.
3. M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," IEEE Signal Process. Mag., vol. 34(4), pp. 18–42, 2017.
4. J. R. Banavar, A. Giacometti, T. X. Hoang, A. Maritan, and T. Škrbić, "A geometrical framework for thinking about proteins," Proteins: Struct. Funct. Bioinform., vol. 93, pp. 145–159, 2025.
5. Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," arXiv:2011.13456, 2011.
6. Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," Adv. Neural Inf. Process. Syst., vol. 34, pp. 1415–1428, 2021.

7. S. Särkkä and A. Solin, Applied Stochastic Differential Equations. Cambridge, UK: Cambridge University Press, 2019.
8. P. Vincent, "A connection between score matching and denoising autoencoders," Neural Comput., vol. 23(7), pp. 1661–1674, 2011.
9. J. Jo, S. Lee, and S. J. Hwang, "Score-based generative modeling of graphs via the system of stochastic differential equations," Proc. ICML, pp. 10362–10383, 2022.
10. M. Hoffmann and F. Noé, "Generating valid Euclidean distance matrices," arXiv:1910.03131, 2019.
11. V. De Bortoli et al., "Riemannian score-based generative modelling," Proc. NeurIPS, pp. 2406–2422, 2022.
12. J. Hyun Lim et al., "Score-Based Diffusion Models in Function Space," J. Mach. Learn. Res., vol. 26(158), pp. 1–62, 2025.
13. S. K. Burley et al., "RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy," Nucleic Acids Res., vol. 47(D1), pp. D464–D474, 2019.
14. Y. Yifeng and Y. Lu, "Diffusion Models with Heavy-Tailed Targets: Score Estimation and Sampling Guarantees," arXiv:2601.06715, 2026.
15. F. Aliniaeifard, V. Wang, and S. van Willigenburg, "Deletion-contraction for a unified Laplacian and applications," Linear Algebra Appl., vol. 691, pp. 50–95, 2024.

A List of Symbols

Symbol	Description
α	Lévy stability index. Controls the heaviness of the distribution tail and the frequency of long-range jumps.
β	Fractional time index of the bifractional Fokker–Planck equation.
γ	Skewness parameter of the Lévy distribution.
μ	Location parameter of the Lévy distribution.
σ	Scale parameter of the Lévy distribution.
$p_L(x;\alpha,\gamma,\mu,\sigma)$	Lévy probability density function.
$\Delta_{\mathcal{G}}$	Graph Laplacian matrix.
$\mathbf{W}$	Graph weight matrix.
$\mathbf{D}$	Graph degree matrix.
Λ	Diagonal matrix of graph Laplacian eigenvalues.
Φ	Matrix of graph Laplacian eigenvectors.
$\mathbf{K}_{t;\alpha,\beta}$	Discrete bifractional heat kernel matrix.
$\mathbf{S}_{t;\alpha,\beta}$	Geometry-aware score matrix.
$\mathbf{x}_0$	Original data sample.
$\mathbf{x}_t$	Noisy sample at diffusion time t.
$\mathbf{L}_{t;\alpha}$	Lévy process.
Δt	Numerical integration time step.
δ_F	Relative Frobenius reconstruction error.
$\delta_{\lvert\cdot\rvert}$	Relative absolute reconstruction error.