fBm-Based Generative Inpainting for the Reconstruction of Chromosomal Distances

Alexander Lobashev[1]
Alexander.Lobashev@skoltech.ru

Dmitry Guskov[1]
Dmitry.Guskov@skoltech.ru

Kirill Polovnikov[1]
kipolovnikov@gmail.com

[1]Skolkovo Institute of Science and Technology, Moscow 121205, Russia

Abstract

Fractional Brownian motion (fBm) features both randomness and strong scale-free correlations, challenging generative models to reproduce the intrinsic memory characterizing the underlying stochastic process. Here we examine a zoo of diffusion-based inpainting methods on a specific dataset of corrupted images, which represent incomplete Euclidean distance matrices (EDMs) of fBm at various memory exponents H. Our dataset implies uniqueness of the data imputation in the regime of low missing ratio providing the unique ground truth for the inpainting. We find that the conditional diffusion generation readily reproduces the built-in correlations of fBm paths in different memory regimes (i.e., for sub-, Brownian and super-diffusion trajectories), providing a robust tool for the statistical imputation at high missing ratio. As a biological application, we apply our fBm-trained diffusion model for the imputation of microscopy-derived distance matrices of chromosomal segments (Fluorescence In Situ Hybridization data) – incomplete due to experimental imperfections – and demonstrate its superiority over the standard approaches used in bioinformatics. Our source code is available at https://github.com/alobashev/fbm-inpainting-benchmark.

1 Introduction

Diffusion probabilistic models are gaining popularity in the field of generative machine learning due to their ability to synthesize diverse and high-quality images from the training distribution. The iterative denoising approach taken by diffusion (1; 2; 3) outperforms in quality of generated samples the previously used schemes (4), such as VAEs (5; 6) and GANs (7; 8; 9), and has demonstrated a distinctive potential in scalability (10). Recently, several conditional diffusion-based generation methods have been developed (11; 12; 13), allowing for effective inpainting of masked images using the pre-trained unconditional diffusion model. Still, whether the diffusion-based inpainting can learn and reproduce the intrinsic non-local dependencies in the pixels of the image drawn from a particular statistical ensemble has remained unaddressed. Furtheremore, recent studies by (14; 15) suggest that modern text-to-image generative diffusion models, such as Dalle-2 (10), Imagen (16), or StableDiffusion (17), tend to recall samples from their training databases, raising questions about their generalization capabilities and bringing up the copyright infringement concerns during the diffusion training process.

In this paper we consider a dataset of incomplete Euclidean distance matrices (EDMs) and propose to approach the EDM completion problem as the image inpainting via conditioning of the diffusion

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

generative models. Importantly, the possibility of existence of the ground truth of the inpainting in the EDM dataset allows one to evaluate the quality of the conditional generation at the instance level. At high missing ratio, however, the solution of EDM completion does not exist and one has to rely on the ensemble-level metrics such as Fréchet Inception Distance (FID). Here we ask: can the diffusion model learn the intrinsic correlations between the entries of the matrix when an ensemble of such matrices is given and statistically reproduce them upon the inpainting? In order to explore the modern generative models at this novel angle we consider the pairwise distances between the points of a discrete fractional Brownian process (fBm), the simplest Gaussian generalization of a Brownian motion with strong scale-free correlations. The built-in memory in the fBm process can induce a non-Brownian exponent of the second moment (also known as the mean-squared displacement of a particle undergoing the anomalous diffusion, (18)), which is translated into strong couplings between the pixels in the distance matrix.

Imputation of missing data has recently got a second wind with the development of high-throughput experimental techniques in chromosome biology. Diffusion models have been recently applied to generate and enhance protein and DNA datasets (19; 20; 21). Hi-C and FISH (Fluorescence In Situ Hybridization, (22)) experiments have provided significant new insights into the fractal (non-Brownian) folding of chromosomes (23; 22), despite the data being noisy and incomplete (24; 25). In particular, it was recently shown that the spatial organization of human chromosomes without loop-extruding complexes (cohesin motors) statistically resembles the ensemble of fractal trajectories with the fractal dimension $d_f = 3$ (26; 27; 28; 29; 30; 23). Such an ensemble – leaving aside the biophysical principles of such organization – can be modelled as trajectories of a subdiffusive fBm particle with $H = 1/3$ (27). This suggests an important statistical insight for the downstream data analysis (31; 32).

FISH imaging experiments produce datasets that represent matrices of pairwise distances between chromosomal loci on single cells, which are obtained in multiplex microscopy. Thus each matrix corresponds to internal distances within a given chromosomal segment in a given cell. Occasionally, some data in the matrices is masked due to experimental imperfections (biochemistry of the protocol) posing a real challenge for the methods of the downstream analysis. In particular, inference of features of the 3D organization at the single cell level is notoriously obscured by the sparsity of the dataset at hand (25). Here we for the first time propose to use the diffusion models for the inpainting of missing values and completion of experimentally-derived FISH matrices. For this aim we deploy the pre-trained fBm diffusion benchmark at $H = 1/3$, thus virtually taking into account the intrinsic correlations present in the fractal chromosome trajectories (27; 26).

2 Background

2.1 Euclidean distance matrices

In this paper we deal with $n \times n$ matrices A of squares of pairwise distances between n points $x_1, x_2, ..., x_n$ in the D-dimensional Euclidean space. For the purposes of this paper, we considered the case of $D = 3$. Such matrices $A = \{a_{ij}\}$ satisfying

$$a_{ij} = ||x_i - x_j||^2, \quad x_i \in \mathbb{R}^D \tag{1}$$

are called Euclidean distance matrices (EDM). Any uncorrupted (complete, noise-less and labelled (33)) EDM A allows for the unique reconstruction of the original coordinates $\{x_i\}$ up to rigid transformations (translations, rotations and reflections). From Eq. 1 it could be derived that the rank of EDM cannot be larger than $D + 2$. A case of incomplete distance matrix, where a particular set of pairwise distances in Eq. 1 is unknown, is a prominent setting of EDM corruption that we study in this paper.

Here we use the following greedy algorithm that checks for the uniqueness of the EDM completion of a given binary mask with known values B which we interpret as an adjacency matrix of some graph. We describe it for $D = 3$, however, it can be simply generalized for an arbitrary D. The algorithm sequentially chooses and adds vertices one by one to a subgraph, ensuring the growing subgraph at each step remains rigid. The key idea is that the coordinates of a new vertex in D dimensions can be uniquely determined, if it is connected to at least $D + 1 = 4$ vertices of the rigid subgraph. Following this idea, on the initial step (i) the algorithm identifies the maximal clique with not less than 4 vertices. As the clique has a complete EDM, there is the corresponding unique realization

in the metric space (all cliques are rigid). Then, (ii) the algorithm seeks and adds a new external vertex, maximally connected with the rigid subgraph, but having not less than 4 edges. This process continues (iii-iv...) until all vertices are included to the subgraph, or none can be further added. If all the vertices are eventually included, the algorithm confirms that the given graph B is rigid and the EDM completion of $\tilde{A}$ is unique.

Fractional Brownian motion

Fractional Brownian motion is one the simplest generalizations of Brownian motion that preserves Gaussianity of the process, but introduces strong memory effects (34). By definition, fBm is a Gaussian process $B_H(t)$ on the interval $[0, T]$ that starts at the origin, $B(0) = 0$, and has the following first two moments:

$$\langle B_H(t) \rangle = 0; \;\; \langle B_H(t) B_H(t') \rangle = \frac{1}{2} \left(t^{2H} + t'^{2H} - |t - t'|^{2H} \right) \tag{2}$$

and $0 < H < 1$ is the Hurst parameter (the memory exponent).

3 Experiments and Results

In this work, we test diffusion-based inpainting methods such as DDPM (35), DDNM (12), DDRM (13), and RePaint (11). These methods only need a pre-trained diffusion model as the generative prior, but we stress that DDNM, DDRM, and RePaint additionally require knowing the corruption operators at the generation. In our case, the corruption operator is a known corruption mask, which helps diffusion to inpaint. Methods DDNM and RePaint use a time-travel trick (also known as resampling in (11)) for better restoration quality, aimed at intense inpainting with a huge mask, but at small missing ratio μ when almost all pixels are known this trick may worsen the quality. By *DDPM inpainting*, we refer to the method introduced by (35), which is equivalent to the RePaint approach (11) without applying resampling steps. It was shown in (12) that DDNM generalizes DDRM and RePaint, but in our paper, we follow the convention that DDNM is a model with parameters, where the travel length and the repeat times are both set to 3, and for RePaint, we use a number of resamplings steps set to 10. The results of the inpainting are presented in the Table 1.

Table 1: Comparison of different inpainting methods in EDM completion. The FID is calculated between an ensemble of distance matrices, which are generated by the Davies-Harte algorithm (36), and reconstructed samples corresponding to three different sparsity values μ. The dimension of the InceptionV3 (37) embedding used for FID is 64. The rank measures the contribution of the first $r = 5$ singular values of the reconstructed matrix in the nuclear norm. Database search refers to the most similar element from the training database computed over known values.

Sparsity	Metrics	**RePaint**	**DDRM**	**DDNM**	**DDPM**	**Database search**
$\mu = 0.25$	RMSE ↓	0.49 ± 0.02	$\mathbf{0.170 \pm 0.017}$	0.211 ± 0.018	0.313 ± 0.023	1.12 ± 0.12
	FID ↓	0.0446 ± 0.0026	$\mathbf{0.013 \pm 0.0017}$	0.027 ± 0.0015	0.0235 ± 0.0019	1.225 ± 0.009
	Rank ↑	0.858 ± 0.025	0.853 ± 0.023	0.854 ± 0.022	0.849 ± 0.025	0.65 ± 0.05
$\mu = 0.5$	RMSE ↓	0.54 ± 0.04	$\mathbf{0.241 \pm 0.027}$	0.325 ± 0.027	0.55 ± 0.05	1.61 ± 0.18
	FID ↓	0.053 ± 0.003	$\mathbf{0.018 \pm 0.002}$	0.053 ± 0.002	0.0246 ± 0.0007	1.79 ± 0.01
	Rank ↑	0.86 ± 0.025	0.854 ± 0.025	0.853 ± 0.025	0.843 ± 0.030	0.63 ± 0.05
$\mu = 0.75$	RMSE ↓	0.68 ± 0.06	$\mathbf{0.42 \pm 0.04}$	0.56 ± 0.04	1.23 ± 0.18	1.97 ± 0.22
	FID ↓	0.075 ± 0.003	$\mathbf{0.034 \pm 0.003}$	0.116 ± 0.002	0.096 ± 0.003	1.25 ± 0.011
	Rank ↑	0.863 ± 0.025	0.854 ± 0.027	0.854 ± 0.027	0.82 ± 0.04	0.65 ± 0.05

As an application of the pre-trained fBm diffusion model, we discuss the results of imputation of missing data in single cell matrices of pairwise distances between chromosomal segments (see Figure

2) obtained from microscopy experiments, namely Fluorescence In Situ Hybridization, FISH (22). The dataset [1] represents the 3D coordinates of 30kb segments on a human chromosome 21 (the human colon cancer cell line, HCT116) measured using the multiplex microscopy. Noticeably, some of the coordinates are missed (nan values in the data). For our purposes of the restoration of missed coordinates we used the data with auxin that supposedly corresponds to the condition with no cohesin-mediated loops. We take the 2Mb-long segment from 28Mb to 30Mb for the analysis, the corresponding file name is "HCT116_chr21-28-30Mb_6h auxin.txt".

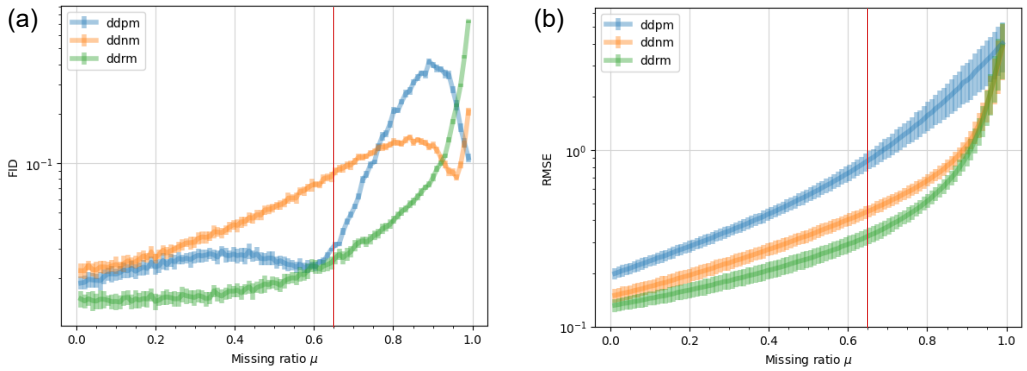

Figure 1: (a) FID and (b) RMSE for three diffusion-based inpainting methods: DDPM (35), DDRM (13) and DDNM (12). The metrics are computed as functions of sparsity μ of originally incomplete EDMs of fBm trajectories. The Hurst parameter of the corresponding fBm trajectories is $H = 1/2$. The errors of RMSE are computed using a sample of 2000 inpainted distance matrices. The errors of FID for each μ are computed by randomly drawing (100 times) sub-samples with 90% of matrices and computing the values of FID for each sub-sample; then the mean and the standard deviation of these values is taken. At $\mu \approx 0.65$ the uniqueness of EDM completion is lost (vertical red line).

First, using the raw data we reproduce the fractal scaling of chromosomal folding (23; 26; 27), i.e. $\langle x^2(s)\rangle^{1/2} \sim s^{1/3}$ (see more details in the supplementary material.) Thus, to inpaint the missing data we decided to use the diffusion model trained on the fBm ensemble with $H = 1/3$ (the fractal dimension is the inverse of the Hurst parameter). Figure 2 shows the resulting matrices obtained using various inpainting methods run on a particular FISH dataset (cell 343). To measure the performance

Table 2: Reconstruction of chromatin (FISH) distance matrices. Average metrics over 670 single cells are shown

Methods	Ens. mean	NN	DDPM	RePaint	DDNM	DDRM
RMSE, nm	147.4 ± 5.2	111.4 ± 3.2	97.2 ± 2.5	98.3 ± 2.7	85.1 ± 2.8	$\mathbf{84.2 \pm 2.8}$
Rank	0.752 ± 0.025	0.79 ± 0.03	0.79 ± 0.04	0.82 ± 0.04	0.82 ± 0.03	0.82 ± 0.03

of the DDPM inpainting in comparison to other methods we chose 670 cells (out of 7380 cells) in the dataset that have exactly 15 missing rows and columns. This corresponds to sparsity $\mu' = 0.29$. In order to compute RMSE (Table 2) we additionally dropped 10 rows and columns from the matrices resulting in $\mu = 0.63$. We then imputed the missing distances using the standard bioinformatics approaches (nearest neighbor, ensemble mean) and various diffusion-based inpainting methods (DDRM, DDNM, DDPM, RePaint). The nearest neighbor approach relies on filling the unknown distances with the nearest neighbour in the same matrix (cell). The ensemble mean approach fills the missing entry with the corresponding average over the cells where this element is known. The average RMSE was computed for each imputed cell over the known values in the dropped 10 columns and rows. Note that since the *entire* rows and columns are missing in FISH distance matrices, precise EDM completion algorithms such as FISTA (38) or trajectory optimization (OPT) are not applicable here. Figure 1 illustrates the performance of various inpainting methods as a function of the missing ratio μ. Around $\mu \approx 0.65$, the EDM loses the uniqueness of its completion, leading to higher FID

[1] Data is publicly available at https://github.com/BogdanBintu/ChromatinImaging

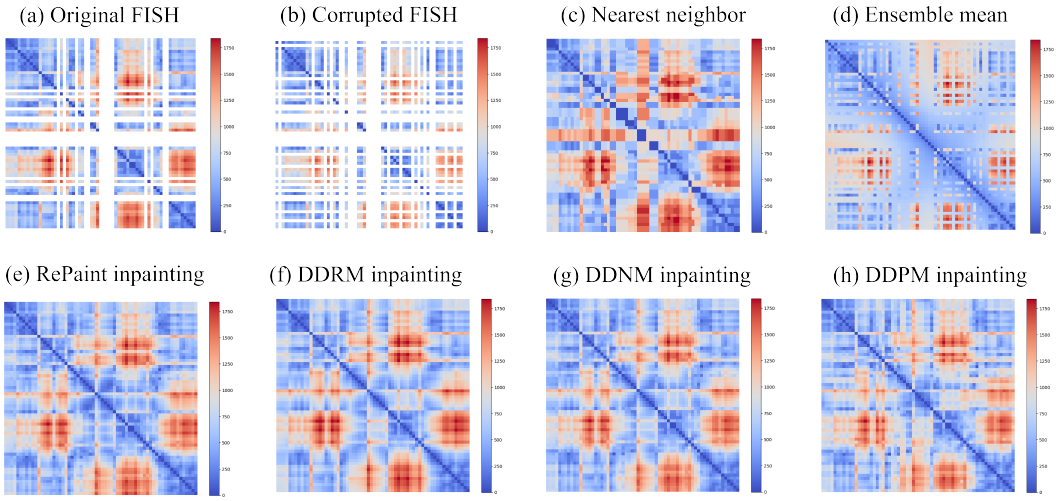

Figure 2: Inpainting of chromosome distance matrices from a FISH experiment. (a) Original experimental matrix with 15 missing rows and columns. (b) Corrupted experimental matrix with 10 rows and columns additionally masked. The masking is needed in order to evaluate RMSE of various inpainting methods at the known (masked) values. The resulting sparsity is $\mu = 0.63$. (c)-(h) Inpainting methods, as indicated. For the ensemble mean (d) the average value is taken over 670 single cell distance matrices, where the corresponding matrix element is known. Diffusion-based methods (e)-(h) exploit the pre-trained diffusion model with the Hurst parameter $H = 1/3$. The colorbars show the range of pairwise distances between chromosomal loci in nm. The data is shown for cell 343.

values as the methods sample different plausible solutions. At $\mu = 1$, where no information is provided for inpainting the methods exhibit contrasting behaviors. DDPM becomes unconditional sampling of EDM matrices. DDNM method, which has repaint steps, performs best in short region but then diverges in terms of FID. The DDRM, which depends on additional a pseudoinverse operator during the sampling, also monotonically diverge in FID. Among all method only DDPM is sensitive to the loss of uniqueness at $\mu \approx 0.65$.

Consistently with numerical experiments, the Table 2 demonstrates that the DDRM inpainting trained on the fBm benchmark is superior over other diffusion-based and bioinformatics approaches. It should be noted that it has a comparable RMSE and rank with DDNM inpainting, while RePaint and DDPM behave slightly worse (the resulting RMSE is more than 10% larger). This is to be compared with other approaches, such as filling the missing distances using the nearest neighbor pixel from the same matrix (NN) or using the average over the cells where this matrix element is known (Ens. mean). These clearly naive approaches behave significantly worse both in the rank and RMSE. This observation highlights that our diffusion-based inpainting shows evidently better performance on a biological dataset than canonical bioinformatics approaches.

4 Conclusion

We apply the unconditional DDPM model on the dataset of euclidean distance matrices of the fractional Brownian motion with memory exponent $H = 1/3$ for the problem of EDM completion, exploring diffusion inpainting methods at various sparsity parameters. We show that the diffusion-based inpainting not only learns the latent representation of the distance matrices, but also manages to properly reproduce the statistical features of the fBm ensemble (the memory exponent).

We observe that DDRM inpainting performs better in terms of FID and RMSE metrics, whereas the FID metric of DDPM indicates a loss of uniqueness in the ground truth for inpainting.

Application of the diffusion pretrained on fBm ensemble for the microscopy-derived dataset of pairwise spatial distances between chromosomal segments demonstrates its superiority in reconstructing the missing distances over the standard approaches widely used in bioinformatics. We thus expect that other chromosomal datasets obtained in high-throughput experiments (such as Hi-C) that can be represented as matrices would benefit from the proposed approach.

References

[1] Ho, J., A. Jain, P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[2] Song, Y., J. Sohl-Dickstein, D. Kingma, et al. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[3] Sohl-Dickstein, J., E. Weiss, N. Maheswaranathan, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. 2015.

[4] Dhariwal, P., A. Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. 2021.

[5] Vahdat, A., J. Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[6] Diederik, P., M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, pages 8780–8794. 2014.

[7] Karras, T., S. Laine, M. Aittala, et al. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119. 2020.

[8] Brock, A., M. Hernán. Large scale gan training for high fidelity natural image synthesis, 2018. ArXiv, `https://arxiv.org/abs/1809.11096`.

[9] Goodfellow, I., J. Pouget-Abadie, M. Mirza, et al. Generative adversarial nets. In *Advances in neural information processing systems*. 2014.

[10] Ramesh, A., P. Dhariwal, A. Nichol, et al. Hierarchical text-conditional image generation with clip latents, 2022. ArXiv, `https://arxiv.org/abs/2204.06125`.

[11] Lugmayr, A., M. Danelljan, A. Romero, et al. Repaint: Inpainting using denoising diffusion probabilistic models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[12] Wang, Y., J. Yu, J. Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023.

[13] Kawar, B., M. Elad, S. Ermon, et al. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

[14] Carlini, N., J. Hayes, M. Nasr, et al. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270. 2023.

[15] Somepalli, G., V. Singla, M. Goldblum, et al. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058. 2023.

[16] Saharia, C., W. Chan, S. Saxena, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[17] Rombach, R., A. Blattmann, D. Lorenz, et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695. 2022.

[18] Metzler, R., J. Klafter. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1):1–77, 2000.

[19] Watson, J., D. Juergens, N. Bennett, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[20] Ingraham, J., M. Baranov, Z. Costello, et al. Illuminating protein space with a programmable generative model. *Nature*, pages 1–9, 2023.

[21] Wang, Y., J. Cheng. Hicdiff: single-cell hi-c data denoising with diffusion models. *bioRxiv*, pages 2023–12, 2023.

[22] Bintu, B., L. J. Mateo, J. H. Su, et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413):eaau1783, 2018.

[23] Lieberman-Aiden, E., N. L. Van Berkum, L. Williams, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[24] Imakaev, M., G. Fudenberg, R. McCord, et al. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999–1003, 2012.

[25] Galitsyna, A. A., M. Gelfand. Single-cell hi-c data analysis: safety in numbers. *Briefings in bioinformatics*, 22(6):bbab316, 2021.

[26] Polovnikov, K. E., H. Brandao, S. Belan, et al. Crumpled polymer with loops recapitulates key features of chromosome organization. *Physical Review X*, 13(4):041029, 2023.

[27] Polovnikov, K., S. Nechaev, M. V. Tamm. Effective hamiltonian of topologically stabilized polymer states. *Soft Matter*, 14(31):6561–6570, 2018.

[28] —. Many-body contacts in fractal polymer chains and fractional brownian trajectories. *Physical Review E*, 99(3):032501, 2019.

[29] Polovnikov, K., M. Gherardi, M. Cosentino-Lagomarsino, et al. Fractal folding and medium viscoelasticity contribute jointly to chromosome dynamics. *Physical Review Letters*, 120(8):088101, 2018.

[30] Tamm, M. V., K. Polovnikov. Dynamics of polymers: classic results and recent developments. Order, Disorder and Criticality: Advanced Problems of Phase Transition Theory. *World Scientific*, 2018.

[31] Polovnikov, K., A. Gorsky, S. Nechaev, et al. Non-backtracking walks reveal compartments in sparse chromatin interaction networks. *Scientific Reports*, 10(1):1–11, 2020.

[32] Ulianov, S., V. V. Zakharova, A. A. Galitsyna, et al. Order and stochasticity in the folding of individual drosophila genomes. *Nature communications*, 12(1):41, 2021.

[33] Dokmanic, I., R. Parhizkar, J. Ranieri, et al. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

[34] Mandelbrot, B., J. W. Van Ness. Fractional brownian motions, fractional noises and applications. *SIAM Rev.*, 10(4):422–437, 1968.

[35] Sohl-Dickstein, J., E. Weiss, N. Maheswaranathan, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[36] Davies, R. B., D. S. Harte. Tests for hurst effect. *Biometrika*, 74(1):95–101, 1987.

[37] Szegedy, C., V. Vanhoucke, S. Ioffe, et al. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826. 2016.

[38] Beck, A., M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.