

Principal Curvatures Estimation with Applications to Single Cell Data

Yanlei Zhang^{1, 2, *}, Lydia Mezrag^{1, 2, *}, Xingzhi Sun³, Charles Xu^{3, 4}, Kincaid Macdonald³, Dhananjay Bhaskar^{3,4}, Smita Krishnaswamy^{2, 3, 4, 5, †}, Guy Wolf^{1, 2, 7, †} and Bastian Rieck^{6, 7, †}

¹Université de Montréal, Dept. of Math. & Stat.; ²Mila – Quebec AI Institute, Montréal, QC, CA

³Yale University, Dept. of Comp. Sci.; ⁴Dept. of Genetics; ⁵Applied Mathematics Program, New Haven, CT, USA

⁶Université de Fribourg, Department of Informatics, Fribourg, FR, CH

⁷Helmholtz Zentrum München, Institute of AI for Health, Munich, BY, DE

*Equal contribution; †Co-senior authors.

Abstract—The rapidly growing field of single-cell transcriptomic sequencing (scRNAseq) presents challenges for data analysis due to its massive datasets. A common method in manifold learning consists in hypothesizing that datasets lie on a lower dimensional manifold. This allows to study the geometry of point clouds by extracting meaningful descriptors like curvature. In this work, we will present *Adaptive Local PCA (AdaL-PCA)*, a data-driven method for accurately estimating various notions of intrinsic curvature on data manifolds, in particular principal curvatures for surfaces. The model relies on local PCA to estimate the tangent spaces. The evaluation of AdaL-PCA on sampled surfaces shows state-of-the-art results. Combined with a PHATE embedding, the model applied to single-cell RNA sequencing data allows us to identify key variations in the cellular differentiation.

Index Terms—Principal curvature, Gaussian curvature, single-cell, principal directions.

I. INTRODUCTION

The estimation of principal curvatures and principal directions is crucial in uncovering directional changes within data manifolds. Indeed, the mean and Gaussian curvatures of surfaces have been studied for several decades in computer graphics and some related areas (e.g., [1]–[5]). Recent methods have proposed to estimate curvature over data manifolds derived from point-cloud data via manifold learning techniques (e.g., [6]–[8]). However, achieving precision is challenging given the variations in data density and the necessity for high-quality samplings. To address this, various methods have been developed. Volume-based approaches like diffusion curvature by [9] and [10] heavily depend on accurate distance estimations. Laplace–Beltrami operator-based approaches, as explored in [11] and [9] encounter limitations

in accurately estimating curvature from small sample sizes. Second Fundamental Form-based approaches, as proposed in [11], demonstrate relatively high-quality curvature estimation for scalar curvature. However, they rely on fixed parameters for neighborhood selection. We introduce adaptability into the estimation process, addressing the challenges associated with variable data density and the absence of intrinsic curvature information by dynamically adjusting parameters based on the local properties of the manifold. This ensures robustness across diverse manifolds. Our **main contributions** are:

- We estimate the point-wise Gaussian curvature of point clouds and their underlying principal curvatures, i.e. *How much* the data curves and *in which directions* it curves the most.
- We dynamically adjust neighborhood scales for local PCA and curvature estimation based on the explained variance ratio. This ensures accurate predictions *without requiring hand-tuning* of the parameters.
- We demonstrate the fidelity of our method relative to ground truth (Gaussian and mean) curvatures on canonical 2-manifolds.
- We illustrate its application to single-cell data analysis, where principal curvatures suggest the directions of cell differentiation.

II. METHODS

For differential geometry preliminaries, we refer the reader to [12], as an extensive introduction to this topic would be beyond the scope of this work and its succinct presentation.

A. Local PCA

Our method starts with Local PCA as described in [13]. Given a point cloud $x_1, \dots, x_m$, we select a neighborhood $\mathcal{N}_{x_i, \epsilon_{\text{PCA}}} := \{x_j : 0 < \|x_j - x_i\| < \epsilon_{\text{PCA}}\}$ around each point x_i for a hyperparameter $\epsilon_{\text{PCA}} > 0$ that has to be determined. Each data matrix containing the neighbors of a point x_i is shifted to be centered around x_i to get a matrix $X_i = [x_{i1} - x_i, \dots, x_{iN_i} - x_i]$ where $N_i := |\mathcal{N}_{x_i, \epsilon_{\text{PCA}}}|$. Then, the columns of X_i are rescaled to $B_i = X_i D_i$ by applying

This research was partially funded by Mitacs Globalink Research Award IT40964 [L.M.]; Yale – Boehringer Ingelheim Biomedical Data Science Fellowship [D.B.]; Hightech Agenda Bavaria, Swiss State Secretariat for Education, Research and Innovation [B.R.]; Humboldt Research Fellowship, CIFAR AI Chair, NSERC Discovery grant 03267, FRQNT grant 343567 [G.W.]; CRM-Simons visiting professor award, NSF career grant 2047856 [S.K.]; and NSF grant DMS-2327211 [G.W., S.K.]. The content provided here is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. Correspondence to guy.wolf@umontreal.ca and smita.krishnaswamy@yale.edu

Algorithm 1 Adaptive Local PCA (AdaL-PCA)

Input: Point cloud data $x_1, \dots, x_m \in \mathbb{R}^3$, query point p , kernel function K with supports in $[0,1]$, data bound δ (maximum pairwise distance in data), ratio bound $\rho_0 \in (0, 1)$ for choosing size of PCA neighborhood.

for $r \in (0, 0.2\delta]$ **do**

$(\mathcal{N}_{p,r}, \mathbf{D}_r) \leftarrow \{(q, \|q - p\|) : 0 < \|q - p\| < r\}$

$\mathbf{X} \leftarrow \mathcal{N}_{p,r} - p$

$\mathbf{D} \leftarrow \text{diag}(\sqrt{K(\mathbf{D}_r/r)})$

$\mathbf{B} \leftarrow \mathbf{D}\mathbf{X}$

$\mathbf{U}\Sigma\mathbf{V}^T \leftarrow \text{SVD}(\mathbf{B})$

$\rho(r) \leftarrow \left\{ \sum_{i=1}^2 \sigma_i^2 / \sum_{i=1}^3 \sigma_i^2 : \sigma_i \in \Sigma \right\}$

end for

$\epsilon_{\text{PCA}} \leftarrow \max(\{r : \rho(r) > \rho_0\})$

$\tau \leftarrow \text{argmin}_r \{\rho(r)\}$

Return: $\epsilon_{\text{PCA}}, \tau$

a diagonal weighting matrix D_i to emphasize the importance of local data. Finally, SVD decomposition yields a numerical approximation of the tangent plane. Since, we are interested mainly in surfaces, the first two eigenvectors are selected as a basis for the local tangent space and the third one serves as a normal vector to the surface.

B. Adaptive Local PCA and parameter selection

AdaL-PCA uses the explained variance ratio for the first two singular values given by

$$\rho(r) := \frac{\sum_{i=1}^2 \sigma_i(r)^2}{\sum_{i=1}^3 \sigma_i(r)^2} \quad (1)$$

to select a suitable parameter ϵ_{PCA} . This ratio describes the fraction of data variance captured by the tangent plane approximated by the span of the first two singular vectors. We set a threshold γ for the ratio $\rho(r)$ and compute the largest r -neighborhood that explains a fraction γ of the data variance. That is,

$$\epsilon_{\text{PCA}} := \max \{r \mid \rho(r) > \gamma\}. \quad (2)$$

We use a similar method to select a radius τ_i for estimating the curvature around each data point x_i . In this case, we need a neighborhood large enough to capture the “bending” of the surface. This is done by computing the lowest value reached by the graph of the explained variance ratio $\rho(r)$,

$$\tau := \arg \min_r \{\rho(r)\}. \quad (3)$$

As illustrated in Fig.1 at a point p on a torus, this approach is motivated by the fact that as the τ -neighborhood increases past a certain threshold, the variance in data can no longer be explained by the selected tangent plane. We refer the reader to Algorithm 1 for a summary of AdaL-PCA’s key steps and emphasize that both ϵ_{PCA} and τ are adjusted at each data point to capture the local geometry.

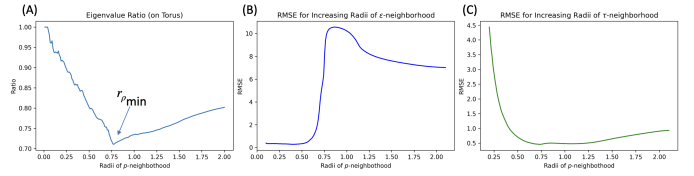


Fig. 1. Comparison of the explained variance ratio of the top two singular values and accuracy (RMSE) of Gaussian curvature estimation w.r.t. increasing radii of ϵ -neighborhood and τ -neighborhood around p on torus.

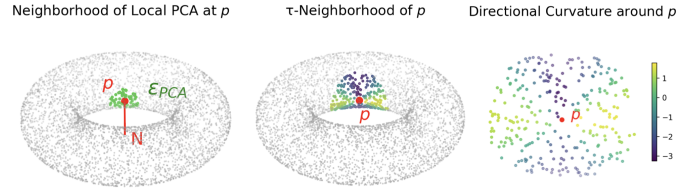


Fig. 2. Directional curvatures in an ϵ -PCA neighborhood of p .

Algorithm 2 Estimation for Principal Curvature, Gaussian Curvature, and Mean Curvature

Input: Point cloud data $x_1, \dots, x_m \in \mathbb{R}^3$, query point p , kernel function K with supports in $[0,1]$, the pair $(\epsilon_{\text{PCA}}, \tau)$, the percentage $p \in (0, 1)$ of total number of points for which the largest (smallest) directional curvature κ_1 (κ_2) is computed.

$(\mathcal{N}_{p,\epsilon_{\text{PCA}}}, \mathbf{D}_{\epsilon_{\text{PCA}}}) \leftarrow \{(q, \|q - p\|) : 0 < \|q - p\| < \epsilon_{\text{PCA}}\}$

$\mathbf{X} \leftarrow \mathcal{N}_{p,\epsilon_{\text{PCA}}} - p$

$\mathbf{D} \leftarrow \text{diag}(\sqrt{K(\mathbf{D}_{\epsilon_{\text{PCA}}}/\epsilon_{\text{PCA}})})$

$\mathbf{B} \leftarrow \mathbf{D}\mathbf{X}$

$\mathbf{U}\Sigma\mathbf{V}^T \leftarrow \text{SVD}(\mathbf{B})$

$\mathbf{O} \leftarrow \mathbf{U}[:, 3, :]$

$\mathcal{N}_{p,\tau} \leftarrow \{q : 0 < \|q - p\| < \tau\}$

for $q \in \mathcal{N}_{p,\tau}$ **do**

$v_q \leftarrow q - p$

$\kappa_q \leftarrow 2(\mathbf{O}[2] \cdot v_q) / \|v_q\|^2$ {by equation 4}

$w_q \leftarrow K(v_q/\tau)$

end for

$\mathbf{C} \leftarrow \text{sort}\{(\kappa_q, w_q) : \kappa_q \text{ in ascending order}\}$

$k \leftarrow \text{int}(p \cdot \text{len}(\mathbf{C}))$

$\kappa_1 \leftarrow \text{sum}(\{\kappa_q \cdot w_q : (\kappa_q, w_q) \in \mathbf{C}[:, k]\}) / \text{sum}(\{w_q\})$

$\kappa_2 \leftarrow \text{sum}(\{\kappa_q \cdot w_q : (\kappa_q, w_q) \in \mathbf{C}[:, -k]\}) / \text{sum}(\{w_q\})$

$K_p \leftarrow \kappa_1 \cdot \kappa_2$

$H_p \leftarrow \kappa_1 + \kappa_2$

Return: $\kappa_1, \kappa_2, K_p, H_p$

C. Curvature Estimation

The directional curvatures $\kappa_i(T)$ at a point x_i in a direction T are approximated (see for instance [14]) by

$$\kappa_i(T) \approx \frac{2N \cdot T}{\|T\|^2} + O(t). \quad (4)$$

Here T is replaced by the entries of X_i in the proper τ_i -neighborhood and N is the orthonormal vector to the frame obtained by local PCA. The principal curvatures κ_1 and κ_2 correspond, respectively, to the highest and lowest values of the directional curvatures.¹ In practice, we select a percentage (20%) of the highest (respectively, lowest) curvatures and average them (using a Gaussian kernel) to approximate κ_1 (respectively, κ_2). By selecting the directional vectors T corresponding to the highest curvatures κ_T , this averaging yields principal directions, while we obtain Gaussian curvature by the product $\kappa_1\kappa_2$. The time complexity of our current implementation is $O(n_\tau(m^2 + \log n_\tau))$, where n_τ is an upper bound on the cardinality of the τ -neighborhoods (in general $n_\tau \ll m$). This can be improved significantly in practice with fast PCA algorithms for scalability [15], [16].

III. RESULTS AND DISCUSSION

Our main contribution is the estimation of the principal curvatures and principal directions. We mainly focus on the application of the principal curvatures and principal directions to biological data and identify key properties and changes in the geometry of these datasets.

We validate the accuracy of our principal curvature estimation by computing Gaussian curvature on toy datasets. Moreover, we apply our estimation of principal curvature and Gaussian curvature for single-cell RNA sequencing data (scRNA-seq).² Gaussian curvature gives the “intensity” for the differentiation of cell states, and principal directions give the directions for the split of the cell lineages.

A. Estimation on Sampled Surfaces

We compare AdaL-PCA’s estimates of Gaussian curvature against two contemporary methods, Hickok & Blumberg [10] and Diffusion Curvature [9]. We also quantify AdaL-PCA’s recovery of ground-truth mean curvature as a validation of the fidelity of its principal curvatures.

To assess the ability of various models to recover the Gaussian and mean curvatures, we generate datasets from three canonical 2-dimensional manifolds: the torus, ellipsoid, and the hyperbolic paraboloid (saddle). Tables I and II were generated from 5000 points sampled uniformly from these surfaces. To study the robustness of each method to noise, we corrupt each point:

$$\tilde{x}_i = x_i + \epsilon_i$$

where each $\epsilon_1, \dots, \epsilon_N \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma)$ and σ between 0.0 and 0.5.

Note that both Hickok & Blumberg’s method and Diffusion Curvature require manually specified parameters, which must be tuned for each dataset. By contrast, AdaL-PCA’s heuristics adapt the method to each dataset. This results in an improved performance observed in Fig. I and Fig. II. Diffusion Curvature is an unsigned measure of local curvature for point clouds sampled from a manifold. Although it differs from Gaussian

¹Note that this is sometimes taken as a definition of principal curvatures.

²Implementation details and some examples can be found at <https://github.com/LydiaMez/AdaL-PCA.git>.

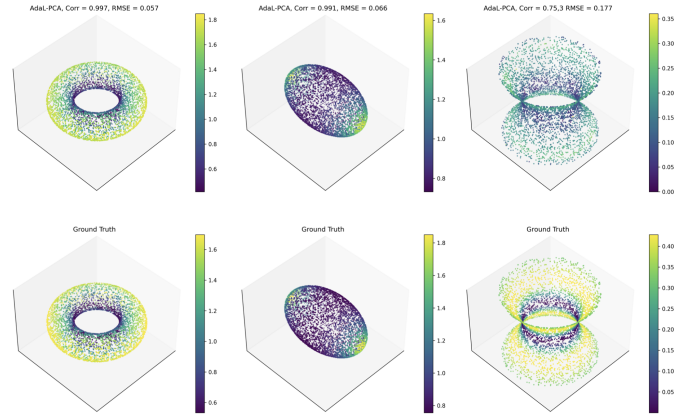


Fig. 3. Comparison of AdaL-PCA against ground truth for mean curvature on three toy datasets. Corr stands for Pearson correlation and RMSE stands for the root means squared error.

TABLE I
Root Mean Square Error (RMSE) and Energy Distance (Eng. Dist) of Gaussian curvature estimation for different noise levels. For RMSE and Eng. Dist, smaller is better.

data	noise	Ours		H. & B.	
		RMSE	Eng.Dist	RMSE	Eng.Dist
Torus	0.0	0.462	0.462	1.302	0.646
	0.1	1.391	0.725	7.489	3.076
	0.2	2.023	1.026	15.914	4.772
	0.3	2.056	1.071	19.143	5.171
	0.4	2.060	1.076	19.971	5.168
	0.5	2.048	1.059	19.944	5.041
Ellipsoid	0.0	0.430	0.251	0.388	0.361
	0.1	0.849	0.277	6.730	3.407
	0.2	0.564	0.832	15.647	5.075
	0.3	1.760	1.296	20.135	5.576
	0.4	1.988	1.565	21.007	5.541
	0.5	2.061	1.643	20.852	5.391
Hyperbolic paraboloid	0.0	0.293	0.321	2.025	1.154
	0.1	0.400	0.405	4.032	1.981
	0.2	0.567	0.538	10.077	3.588
	0.3	0.673	0.674	12.829	4.065
	0.4	0.753	0.757	13.532	4.069
	0.5	0.776	0.787	13.230	3.908

curvature, numerical experiments detailed in [9] suggest a correlation. Therefore, we report only the Pearson correlation for diffusion curvature.

B. Curvature estimation for single-cell data

We apply our model to single-cell data for cell state differentiation direction discovery. We use RNA sequencing data for human embryonic stem cells available at [17], collected over 27 days during which cells start as embryonic stem cells and then progressively differentiate into different cellular lineages. Low-dimensional manifold visualization of this data using PHATE (Fig. 4, A) shows that embryonic cells (days 0-3, displayed in red) branch into two lineages: endoderm

TABLE II
Pearson Correlation Coefficient (Pearson Corr.) of Gaussian curvature estimation for different noise levels.

data	noise	Ours	Diffusion Curvature
Torus	0.0	0.996	0.445
	0.1	0.865	0.270
	0.2	0.633	0.304
	0.3	0.550	0.308
	0.4	0.440	0.273
	0.5	0.408	0.243
Ellipsoid	0.0	0.988	0.149
	0.1	0.325	0.057
	0.2	0.124	0.002
	0.3	-0.153	0.017
	0.4	-0.131	0.044
	0.5	-0.018	0.048
Hyperbolic paraboloid	0.0	0.747	0.398
	0.1	0.603	0.333
	0.2	0.481	0.282
	0.3	0.428	0.336
	0.4	0.386	0.342
	0.5	0.363	0.327

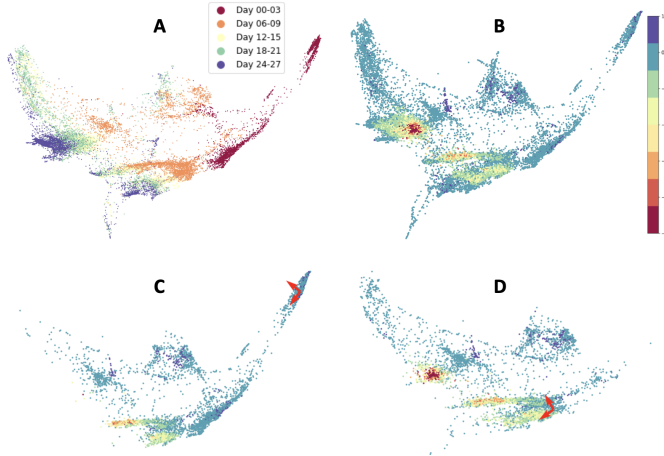


Fig. 4. Gaussian curvature and principal directions of embryonic stem cell differentiation. (A) PHATE visualization of scRNA-seq data color-coded by time intervals. (B) PHATE plot colored by Gaussian curvature values. (C, D) Principal directions at different stages of development of cells.

(upper split) and ectoderm (lower split) around day 6. Further differentiation occurs during days 12-27. This is reflected in Fig. 4B with relatively constant zero curvature values at days 0-3 and a transition into a region of high variations in curvature. We observe starting from day 3 a transition into very low negative values of curvature and then a rapid progression into higher values close to zero as we approach day 27. This is consistent with the fact that the region 0-3 days corresponds to the stem state and the region 12-27 to the differentiated state. In addition to the signed curvature that provides a better appreciation of the cellular differentiation

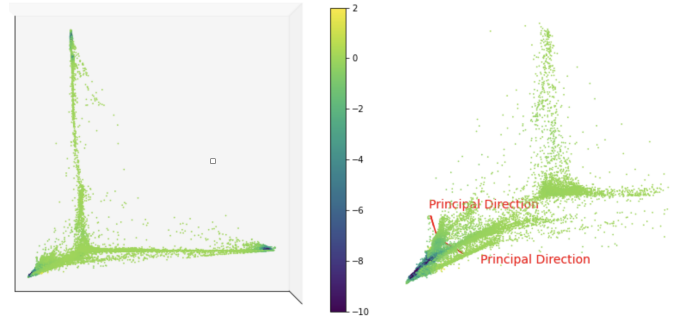


Fig. 5. Gaussian curvature and principal directions on IPSC dataset using AdaL-PCA.

into several lineages (cell types), the principal directions in Fig. 4C, D obtained from projecting the three-dimensional principal directions using the PHATE embedding allow us to track the state towards which the cells differentiate, adding directional information.

We estimated the curvature of a publicly available single-cell induced pluripotent stem cell (iPSC) reprogramming. In this dataset, mass cytometry is used to quantitatively measure 33 protein biomarkers in 2005 mouse fibroblast cells induced to undergo reprogramming into stem cell state. Low-dimensional PHATE visualization of this data shows fibroblasts progressing to a point of divergence where two lineages emerge, one that successfully undergoes reprogramming and another that undergoes apoptosis (cell death). Our model correctly identifies the initial branching point as having negative values of Gaussian curvature indicating saddle-like divergent paths out of the branching point (Fig. 5). Moreover, the principal directions on the diverging branch correctly identify the directions in which the cell lineages diverge.

IV. CONCLUSION

We introduced Adaptive Local PCA (AdaL-PCA), a novel method for estimating intrinsic curvature on data manifolds, with a focus on principal curvatures and directions. By dynamically adjusting neighborhood scales based on the explained variance ratio, AdaL-PCA provides robust and accurate curvature estimates without requiring manual parameter tuning. This adaptability effectively handles variations in data density and the lack of prior curvature information, making it ideal for complex, diverse datasets. We validated AdaL-PCA on synthetic surfaces, demonstrating its ability to recover Gaussian and mean curvatures even in noisy settings. Additionally, we applied it to human embryonic single-cell RNA sequencing data, revealing key directions of cellular differentiation and providing biologically meaningful insights. These results highlight AdaL-PCA's potential in both geometric data analysis and practical applications like single-cell studies. Future work may extend the method to higher dimensions for scalar curvature estimation and improve efficiency by integrating neural network-based local distribution estimation or exploring alternative local PCA frameworks.

REFERENCES

- [1] Y. L. Xin, "Curvature estimates for submanifolds with prescribed gauss image and mean curvature," *Calculus of Variations and Partial Differential Equations*, vol. 37, no. 3, pp. 385–405, Mar 2010. [Online]. Available: <https://doi.org/10.1007/s00526-009-0268-8>
- [2] T. Surazhsky, E. Magid, O. Soldea, G. Elber, and E. Rivlin, "A comparison of gaussian and mean curvatures estimation methods on triangular meshes," in *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, vol. 1, 2003, pp. 1021–1026 vol.1.
- [3] P. Flynn and A. Jain, "On reliable curvature estimation," in *Proceedings CVPR '89: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1989, pp. 110–116.
- [4] C. Zhao, D. Zhao, and Y. Chen, "Simplified gaussian and mean curvatures to range image segmentation," in *Proceedings of 13th International Conference on Pattern Recognition*, vol. 2, 1996, pp. 427–431 vol.2.
- [5] S. Rusinkiewicz, "Estimating curvatures and their derivatives on triangle meshes," in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.* IEEE, 2004, pp. 486–493.
- [6] G. Bella, G. Bruno, and A. Santoro, "Ball pivoting algorithm and discrete gaussian curvature: A direct way to curved nanographene circularly polarized luminescence spectral simulation," *FlatChem*, vol. 42, p. 100567, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2452262723000995>
- [7] J. Hou, Y. Li, Z. Sun, H. Wang, M. Lu, J. Hu, and H. Wu, "A cooperative regulation method for greenhouse soil moisture and light using gaussian curvature and machine learning algorithms," *Computers and Electronics in Agriculture*, vol. 215, p. 108452, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168169923008402>
- [8] W. He, Z. Jiang, C. Zhang, and A. M. Sainju, "Curvanet: Geometric deep learning based on directional curvature for 3d shape analysis," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2214–2224. [Online]. Available: <https://doi.org/10.1145/3394486.3403272>
- [9] D. Bhaskar, K. MacDonald, O. Fasina, D. S. Thomas, B. Rieck, I. Adelstein, and S. Krishnaswamy, "Diffusion curvature for estimating local curvature in high dimensional data," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: <https://openreview.net/forum?id=AYII8AkvD1e>
- [10] A. Hickok and A. J. Blumberg, "An intrinsic approach to scalar-curvature estimation for point clouds," *arXiv preprint arXiv:2308.02615*, 2023.
- [11] D. Sriharan, S. Wang, and S. Hormoz, "Computing the riemannian curvature of image patch and single-cell rna sequencing data manifolds using extrinsic differential geometry," *Proceedings of the National Academy of Sciences*, vol. 118, no. 29, p. e2100473118, 2021.
- [12] M. P. do Carmo, *Differential geometry of curves & surfaces*, 2nd ed. Dover Publications, Inc., Mineola, NY, 2016.
- [13] A. Singer and H.-T. Wu, "Vector diffusion maps and the connection laplacian," *Communications on pure and applied mathematics*, vol. 65, no. 8, pp. 1067–1144, 2012.
- [14] G. Taubin, "Estimating the tensor of curvature of a surface from a polyhedral approximation," in *Proceedings of IEEE International Conference on Computer Vision.* IEEE, 1995, pp. 902–907.
- [15] S. L. Shishkin, A. Shalaginov, and S. D. Bopardikar, "Fast approximate truncated svd," *Numerical Linear Algebra with Applications*, vol. 26, no. 4, p. e2246, 2019.
- [16] C. Xu, W. Xu, and K. Jing, "Fast algorithms for singular value decomposition and the inverse of nearly low-rank matrices," *National Science Review*, vol. 10, no. 6, p. nwad083, 2023.
- [17] K. Moon, "Embryoid body data for phate," *Mendeley Data*, vol. 1, 2018.