Geometry Based Data Generation

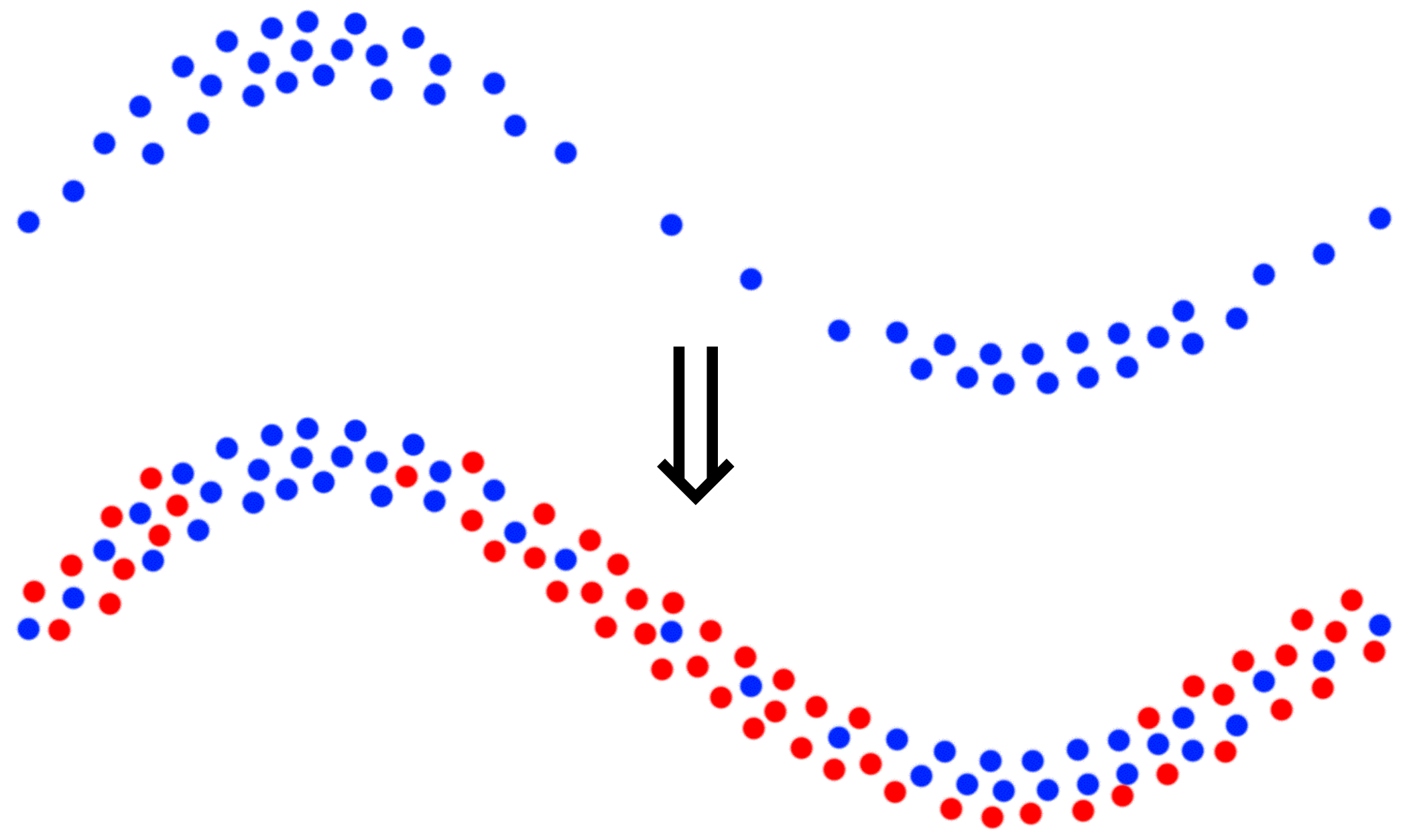
Ofir Lindenbaum [1,*], Jay S. Stanley III [2,*], Guy Wolf [1,†] & Smita Krishnaswamy [1,2,3,†]

[1] Applied Mathematics Program, [2] Comp. Bio. & Bioinformatics Program, [3] Depts. of Genetics & Computer Science, Yale University.
[*] Equal contributions; [†] equal contributions.

Yale

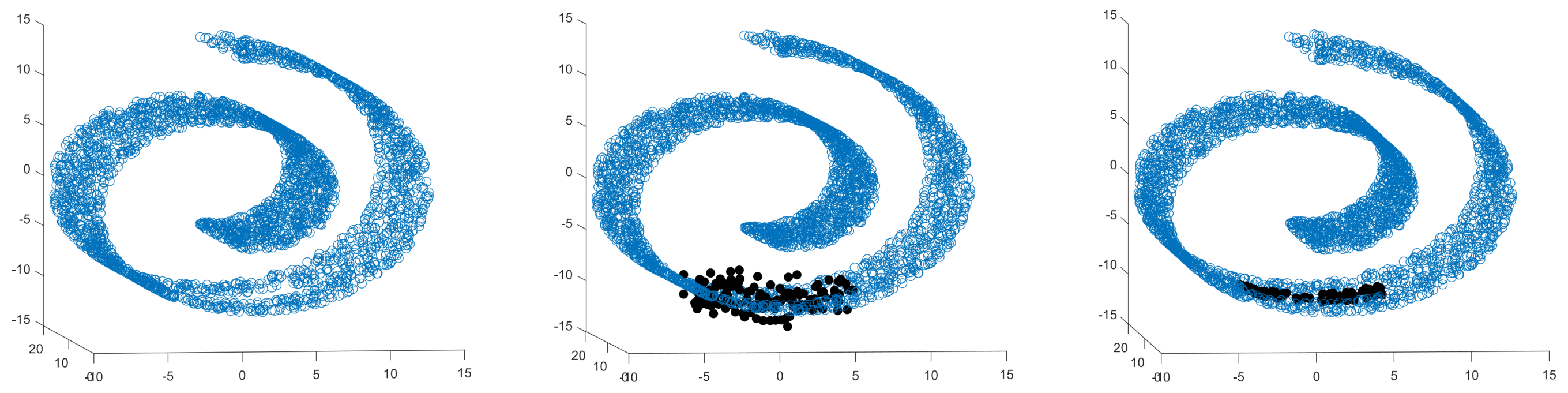
Motivation

Generate points uniformly from intrinsic data geometry / manifold:

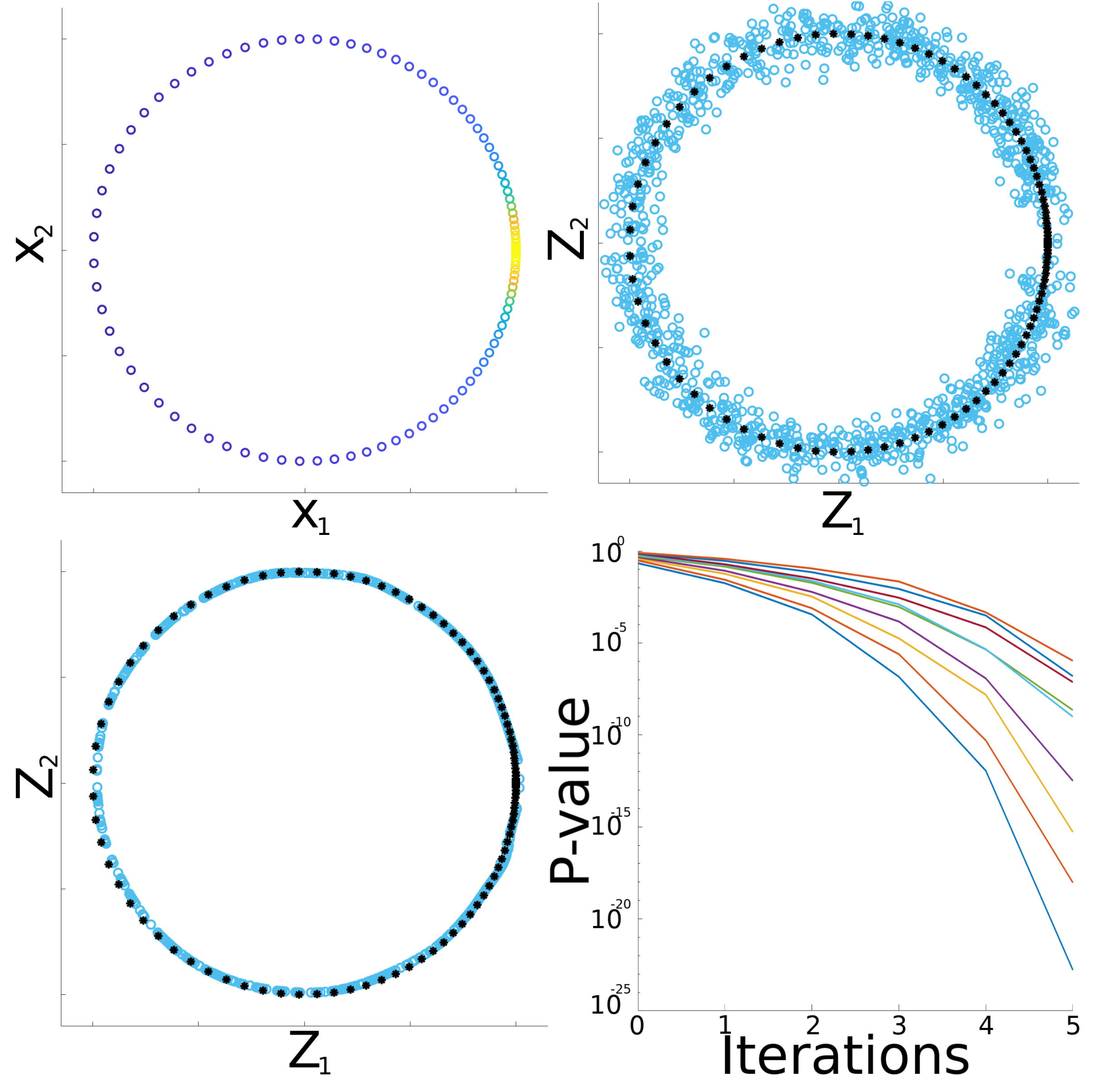


SUGAR

- Compute MGC [2] kernel: $\hat{\mathcal{K}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\boldsymbol{r}\in\text{data}} \frac{\mathcal{K}(\boldsymbol{x},\boldsymbol{r})\mathcal{K}(\boldsymbol{r},\boldsymbol{y})}{\text{density}(r)}$
- Define Markovian random walks: $\Pr[x \overset{1\text{ step}}{\rightsquigarrow} y] = \frac{\hat{\mathcal{K}}(\boldsymbol{x},\boldsymbol{y})}{\|\hat{\mathcal{K}}(\boldsymbol{x},\cdot)\|_1}$
- Initialize new points: $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y},\boldsymbol{\Sigma}_{\boldsymbol{y}})$, $y \in$ data
- Walk towards the data manifold: $\boldsymbol{x} \mapsto \sum_{\boldsymbol{y}\in\text{data}} \boldsymbol{y} \cdot \Pr[x \overset{t\text{ steps}}{\rightsquigarrow} y]$

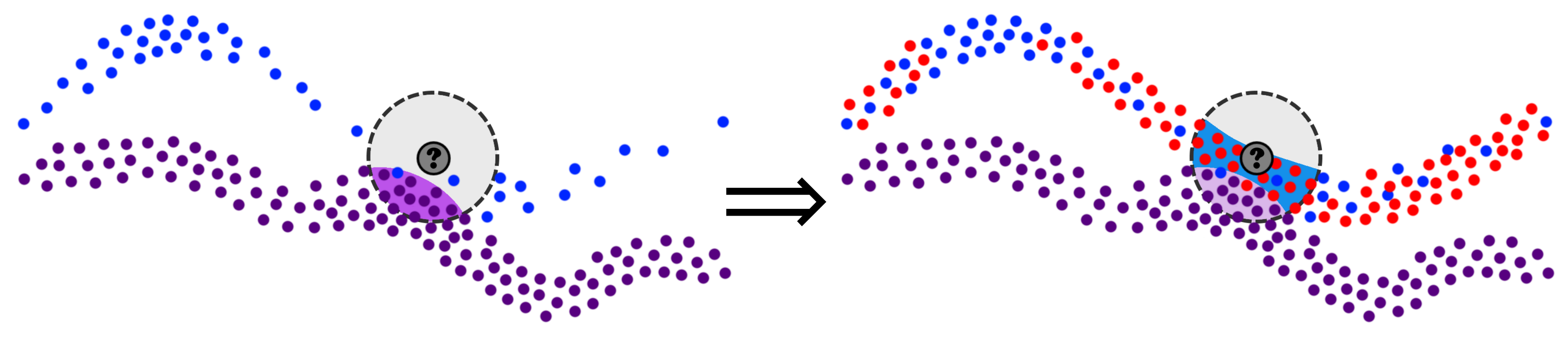


Manifold Equalization



SUGAR recovers undersampled regions according to manifold geometry.

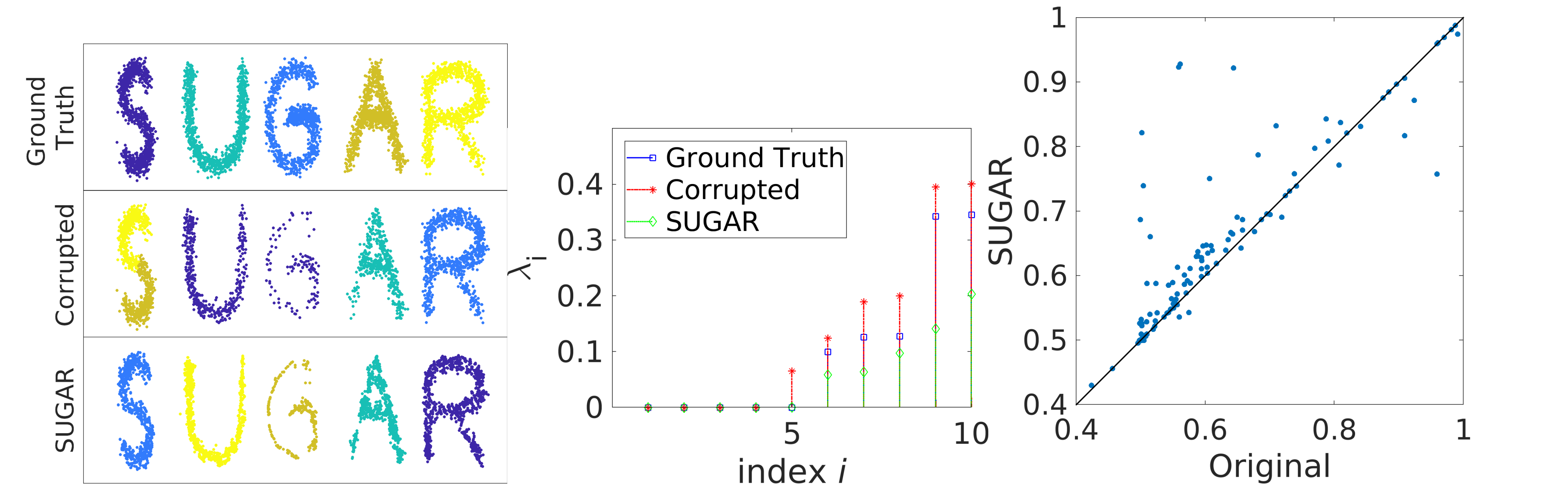
Classification



	k-NN			SVM			RUSBoost
	Orig	SMOTE	SUGAR	Orig	SMOTE	SUGAR	
ACP	0.67	0.76	**0.78**	0.77	0.77	**0.78**	0.75
ACR	0.64	0.73	**0.77**	0.78	0.78	**0.84**	0.81
MCC	0.66	0.74	**0.78**	0.78	0.78	**0.84**	0.80

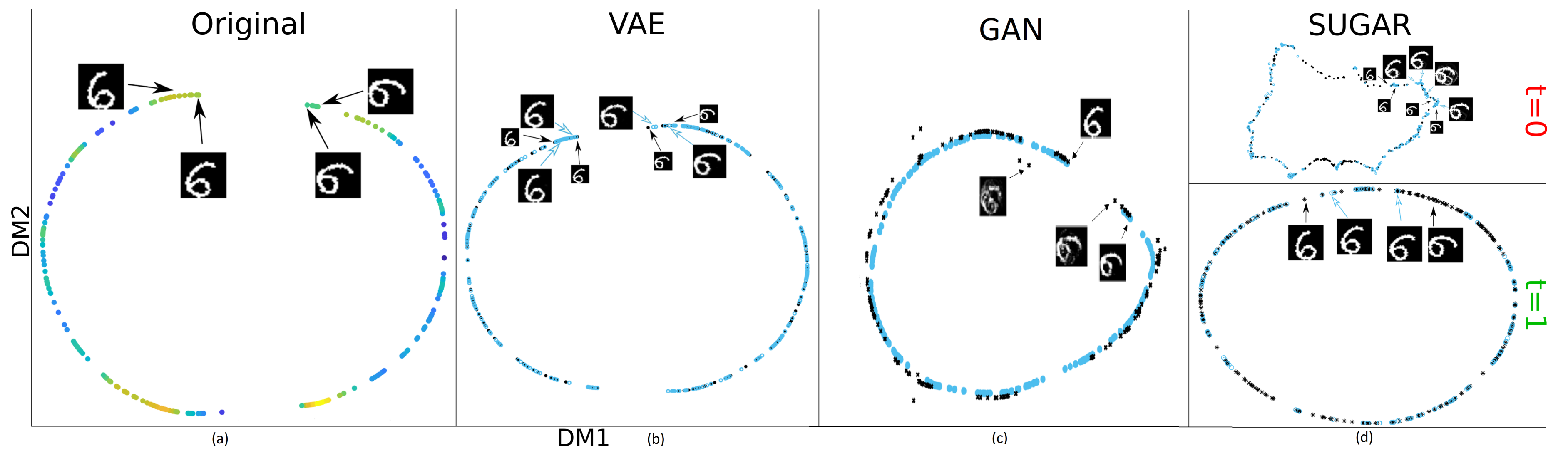
61 Imbalance datasets- Average class precision (ACP), class recall (ACR), and the Matthews correlation coefficient (MCC) for kNN and kernel SVM classifiers (using 10-fold cross validation) before / after SMOTE and SUGAR, and for RUSBoost classification.

Clustering



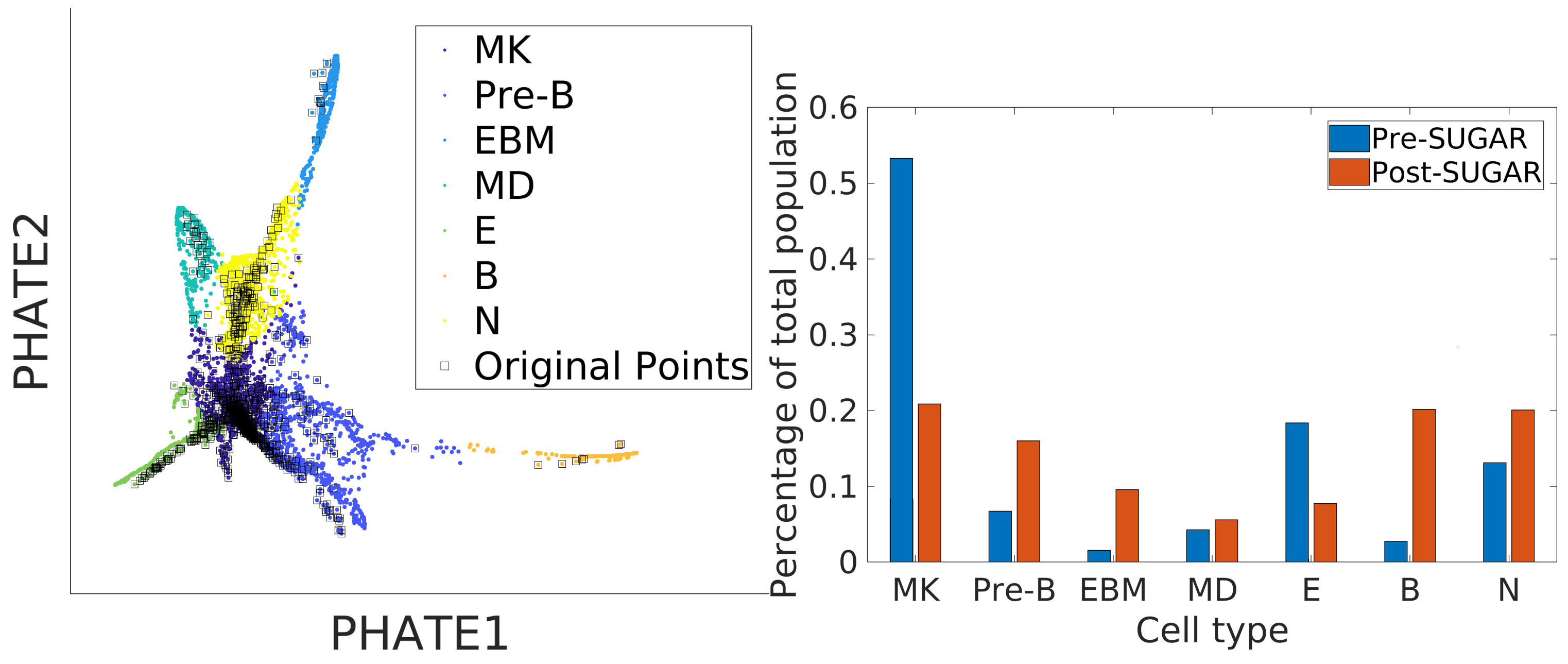
SUGAR enhances clustering accuracy **Left, top:** A ground truth SUGAR manifold was generated. **Left, middle:** Non uniform sampling of the ground truth manifold leads to inaccurate spectral clusters. **Left, bottom:** Spectral clusters are restored following SUGAR augmentation. **Middle:** The eigenvalues of the ground truth, corrupted, and SUGAR corrected graphs reveal that SUGAR restored the disconnected letters of the graph, evidenced by the multiplicity of the 0 eigenvalue. **Right:** Rand index scores of k-means before and after applying SUGAR on 115 datasets from [1].

MNIST Rotations



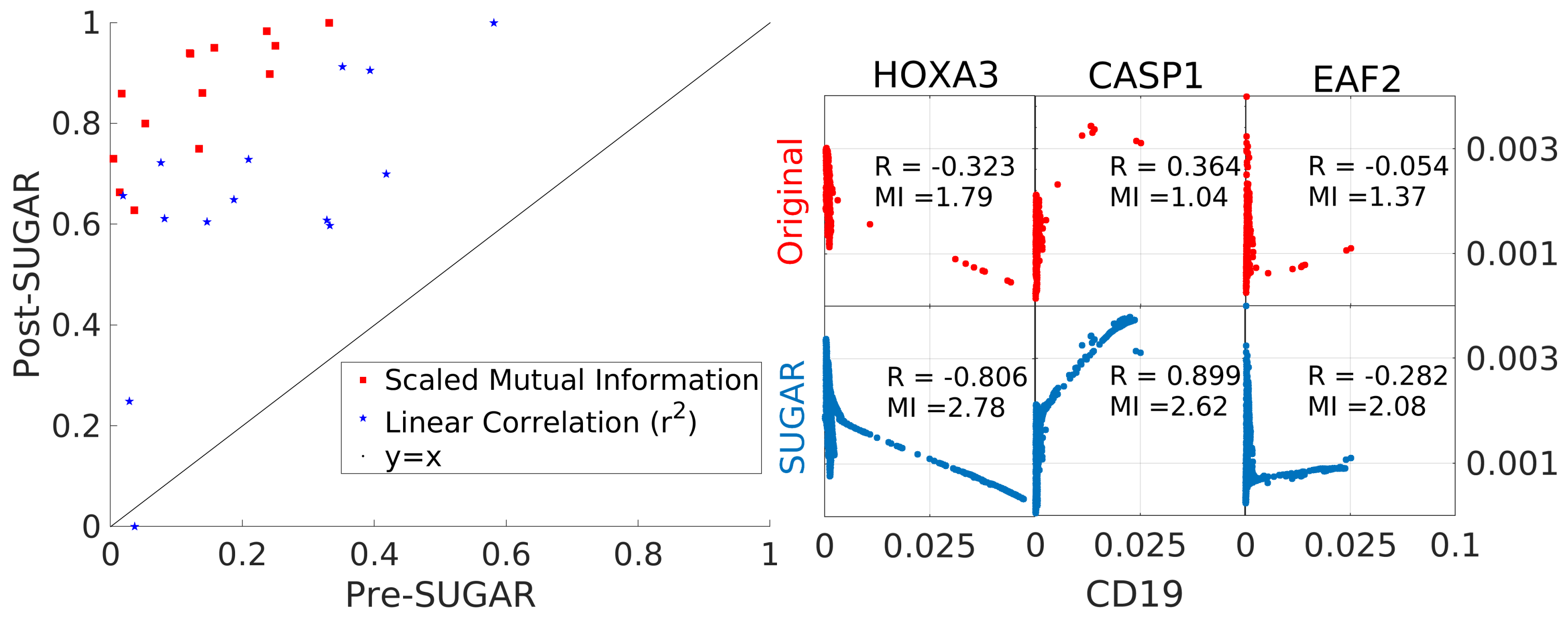
SUGAR recovers undersampled regions according to manifold geometry.

Hypothetical Population Exploration [4]



SUGAR illuminates hypothetical cell types in data. [4] explore a continuum model of early hematopoeisis. SUGAR generated new points in undersampled cell types and restored an absent late-B cell population. **Left:** PHATE embedding of the data. **Right:** Cell type distribution before and after SUGAR.

Canonical Gene-Gene Relationships [4]



SUGAR recovers gene-gene relationships. **Left:** Intra-module mutual information and linear correlation of modules discussed in [4]. **Right:** Two canonically correlated genes for B cell maturation (CD19), HOXA3 (negative) and CASP1 (positive), are recovered by SUGAR. [4] focus on EAF2 as a marker for neutrophils and monocytes, however, data after SUGAR illuminates a correlation in late B cells shown in [3].

References

[1] J. Alcalá-Fdez, L. Sanchez, S. Garcia, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, et al. Keel: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3):307–318, 2009.

[2] A. Bermanis, M. Salhov, G. Wolf, and A. Averbuch. Measure-based diffusion grid construction and high-dimensional data discretization. *Applied and Computational Harmonic Analysis*, 40(2):207–228, 2016.

[3] Y. Li, Y. Takahashi, S.-i. Fujii, Y. Zhou, R. Hong, A. Suzuki, T. Tsubata, K. Hase, and J.-Y. Wang. Eaf2 mediates germinal centre b-cell apoptosis to suppress excessive immune responses and prevent autoimmunity. *Nature communications*, 7:10836, 2016.

[4] L. Velten, S. F. Haas, S. Raffel, S. Blaszkiewicz, S. Islam, B. P. Hennig, C. Hirche, C. Lutz, E. C. Buss, D. Nowak, et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol*, 19:271–281, 2017.

krishnaswamylab.org

November 23, 2018

smita.krishnaswamy@yale.edu