Scalable Gromov-Wasserstein based comparison of biological time series

N. Kravtsova[1] R. L. McGee II[2] A. T. Dawes[1,3]

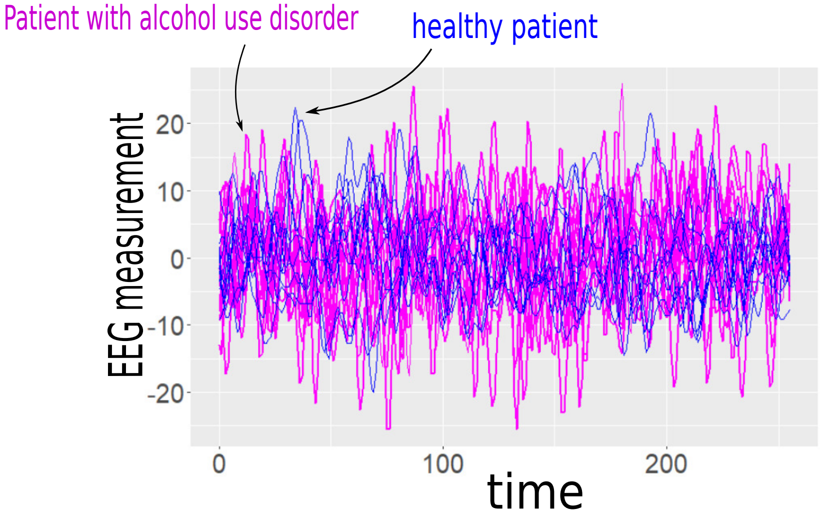
[1]Department of Mathematics
The Ohio State University

[2]Department of Mathematics and Computer Science
College of the Holy Cross

[3]Department of Molecular Genetics
The Ohio State University

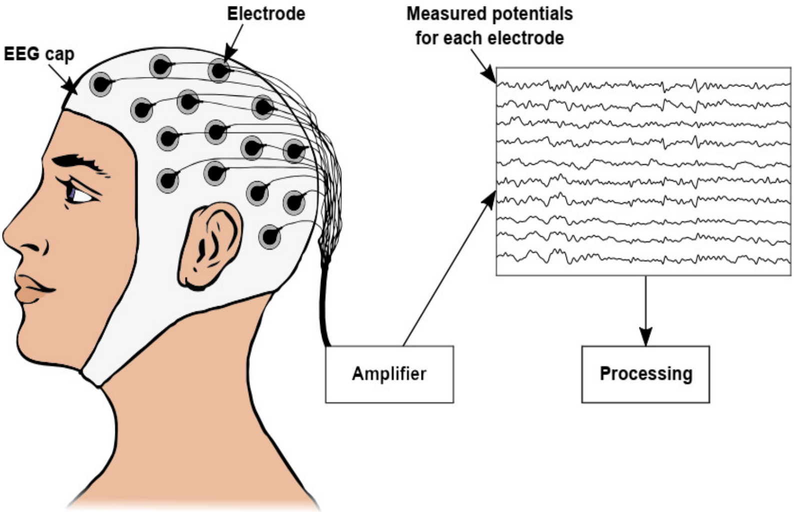
Paper: Kravtsova, McGee II, Dawes (2023), *Bull. Math. Biol.*

Motivating example:
Medical time series with two classes



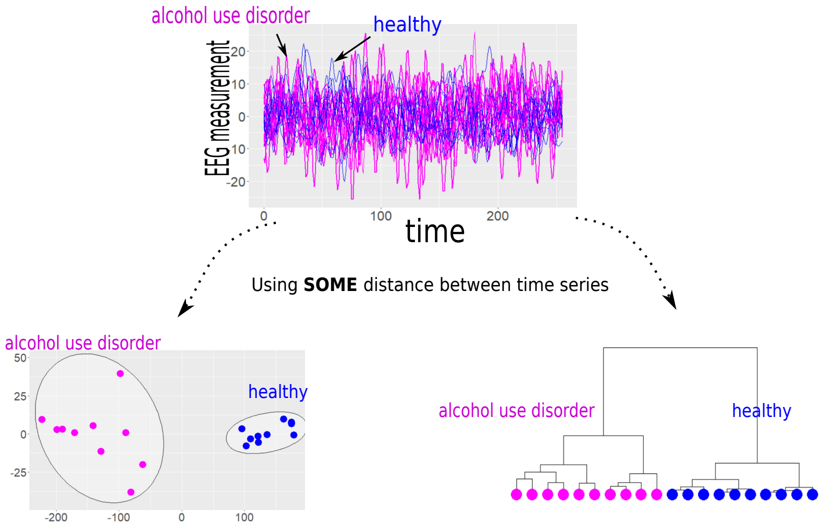
Dataset *smni9_eeg_data* from *UCI Machine Learning repository (Dua and Graph 2017)*

Motivating example: Electroencephalogram (EEG) process illustration

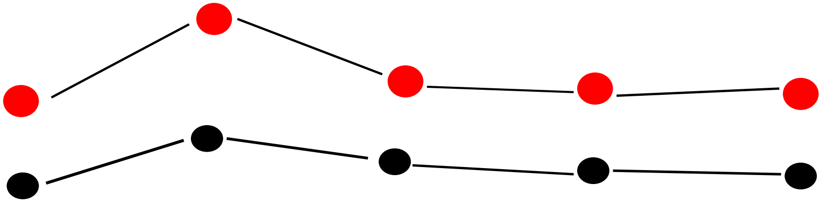


Picture from *Nagel 2019*

Motivating example:
Separate two classes (healthy vs. alcohol use disorder)

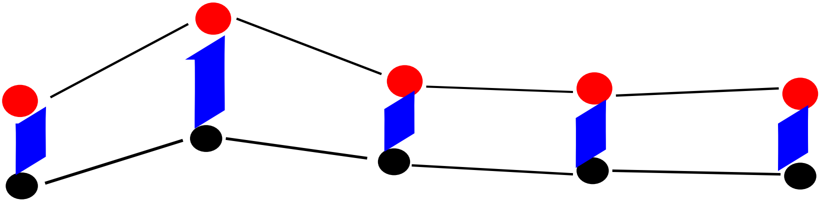


Distance between two trajectories



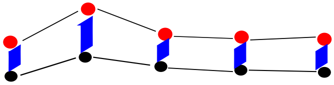
Distance between two trajectories

Euclidean

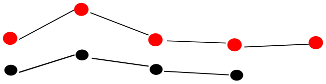


Distance between two trajectories

Euclidean

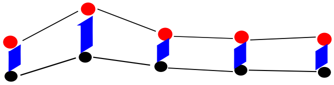


?

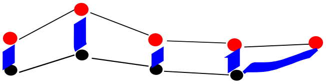


Distance between two trajectories

Euclidean

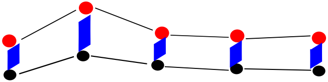


DTW

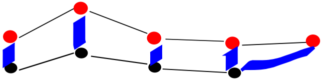


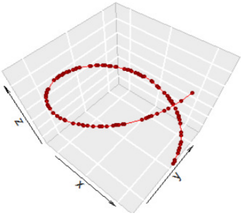
Distance between two trajectories

Euclidean

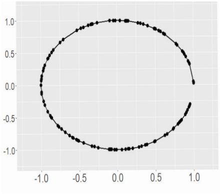


DTW



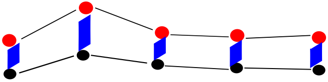


??

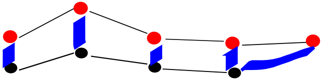


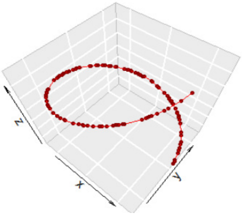
Distance between two trajectories

Euclidean

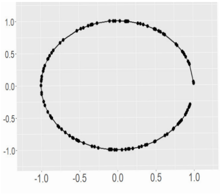


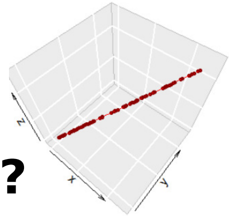
DTW



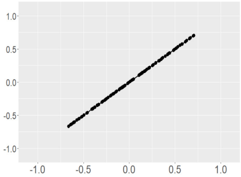


??



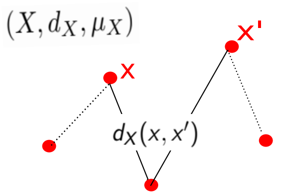


??

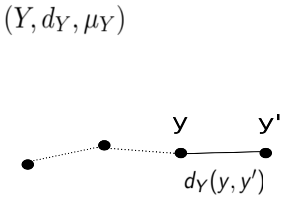


Define distance between time series based on Gromov-Wasserstein distance of *Mémoli 2011*

View two trajectories as metric measure spaces:



$\mu_X = (1/5, \ 1/5 \ \ 1/5 \ \ 1/5 \ \ 1/5)$



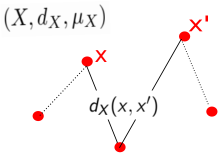
$\mu_Y = (1/4 \ \ 1/4 \ \ 1/4 \ \ 1/4)$

NOTE: Even if both trajectories lie in the same space (e.g. $\mathbb{R}^2$), this technique purposely ignores it

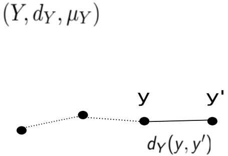
Define distance between time series based on Gromov-Wasserstein distance of *Mémoli 2011*

Mémoli 2011 defines the $p \in [1, \infty)$ **Gromov-Wasserstein distance** between metric measure spaces by

$$GW(X, Y) := \frac{1}{2} \inf_{\mu \in \mathcal{C}(\mu_X, \mu_Y)} \left(\int_{X \times Y} \int_{X \times Y} |d_X(x, x') - d_Y(y, y')|^p \, d\mu(x, y) d\mu(x', y') \right)^{1/p}$$



$\mu_X = (1/5, \ 1/5 \ \ 1/5 \ \ 1/5 \ \ 1/5)$



$\mu_Y = (1/4 \ \ 1/4 \ \ 1/4 \ \ 1/4)$

Note: Non-convex program

Define distance between time series based on Gromov-Wasserstein distance of *Mémoli 2011*

To overcome non-convexity issue, two main approaches exist:

1. Regularize *GW* objective (*Peyré, Cuturi, & Solomon 2016*)

   Disadvantages: Still non-convex

   Advantages: Convenient gradient descent (used in *Demetci et al. 2022* for bio application)

Define distance between time series based on Gromov-Wasserstein distance of *Mémoli 2011*

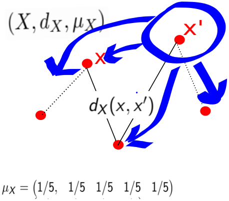
To overcome non-convexity issue, two main approaches exist:

2. Replace *GW* it's lower bounds (*Mémoli 2011*, *Chowdhury & Mémoli 2019*)

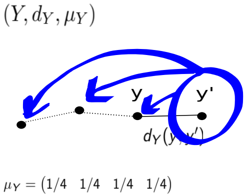
   Advantages: - Convex programs → can be solved exactly!
   - Amenable to statistical analysis (*Weitkamp et al. 2022*)

   Disadvantages: how far is given lower bound from actual *GW*?

Define distance between time series based on Gromov-Wasserstein distance of *Mémoli 2011*



Local distribution of distance at x':
distribution of $d_X(x', \cdot)$

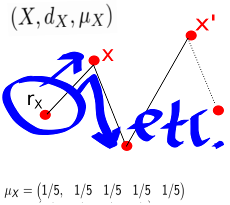


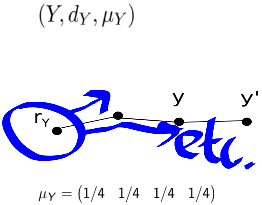
Local distribution of distance at y':
distribution of $d_Y(y', \cdot)$

The Third Lower Bound (*Mémoli 2011*, *Chowdhury & Mémoli 2019*) would compare **ALL** local distributions

Define distance between time series based on Gromov-Wasserstein distance of *Mémoli 2011*

We propose to pick **ONE** particular local distribution:
local distribution at the start of the trajectory





Our program reads: for any $p \in [1, \infty)$,

$$GW_\tau(X, Y) := \inf_{\mu \in \mathcal{C}(\mu_X, \mu_Y)} \left(\int_{X \times Y} |d_X(r_X, x) - d_Y(r_Y, y)|^p \, d\mu(x, y) \right)^{1/p}$$

Properties of GW_τ distance between time series

The object

$$GW_\tau(X,Y) := \inf_{\mu \in \mathcal{C}(\mu_X,\mu_Y)} \left(\int_{X\times Y} |d_X(r_X,x) - d_Y(r_Y,y)|^p \, d\mu(x,y) \right)^{1/p}$$

satisfies:

1. GW_τ is an upper bound of GW.
   Open question: $TLB \leq GW \leq GW_\tau$

Properties of GW_τ distance between time series

The object

$$GW_\tau(X, Y) := \inf_{\mu \in \mathcal{C}(\mu_X, \mu_Y)} \left(\int_{X \times Y} |d_X(r_X, x) - d_Y(r_Y, y)|^p \, d\mu(x, y) \right)^{1/p}$$

satisfies:

1. GW_τ is an upper bound of GW.
   Open question: $TLB \leq GW \leq GW_\tau$
2. GW_τ is equivalent to Wasserstein distance between *local distributions of distance (Mémoli 2011)* at the start of each trajectory

Properties of GW_τ distance between time series

The object

$$GW_\tau(X, Y) := \inf_{\mu \in \mathcal{C}(\mu_X, \mu_Y)} \left(\int_{X \times Y} |d_X(r_X, x) - d_Y(r_Y, y)|^p \, d\mu(x, y) \right)^{1/p}$$

satisfies:

1. GW_τ is an upper bound of GW.
   Open question: $TLB \leq GW \leq GW_\tau$
2. GW_τ is equivalent to Wasserstein distance between *local distributions of distance (Mémoli 2011)* at the start of each trajectory
3. GW_τ is a metric on the space of (certain) equivalence classes of trajectories

Properties of GW_τ distance between time series

The object

$$GW_\tau(X,Y) := \inf_{\mu \in \mathcal{C}(\mu_X,\mu_Y)} \left(\int_{X \times Y} |d_X(r_X, x) - d_Y(r_Y, y)|^p \, d\mu(x,y) \right)^{1/p}$$

satisfies:

1. GW_τ is an upper bound of GW.
   Open question: $TLB \leq GW \leq GW_\tau$
2. GW_τ is equivalent to Wasserstein distance between *local distributions of distance (Mémoli 2011)* at the start of each trajectory
3. GW_τ is a metric on the space of (certain) equivalence classes of trajectories
4. Similar construction is defined for graphs in *Le, Ho, & Yamada 2022*

Properties of GW_τ distance between time series

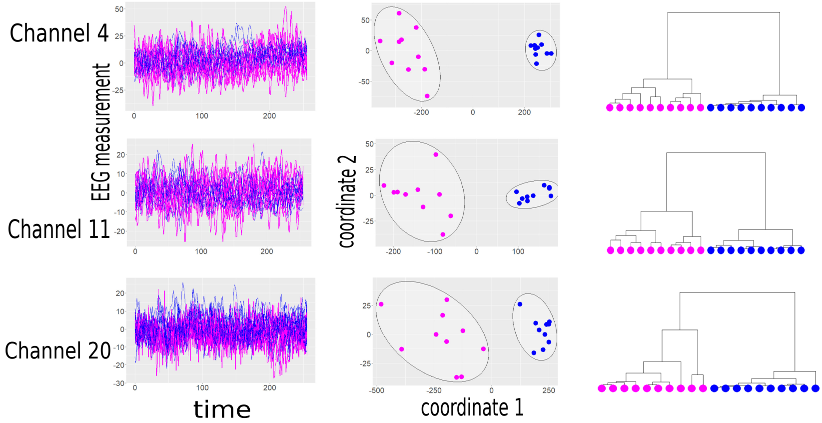
The object

$$GW_\tau(X,Y) := \inf_{\mu \in \mathcal{C}(\mu_X,\mu_Y)} \left(\int_{X\times Y} |d_X(r_X,x) - d_Y(r_Y,y)|^p \, d\mu(x,y) \right)^{1/p}$$

satisfies:

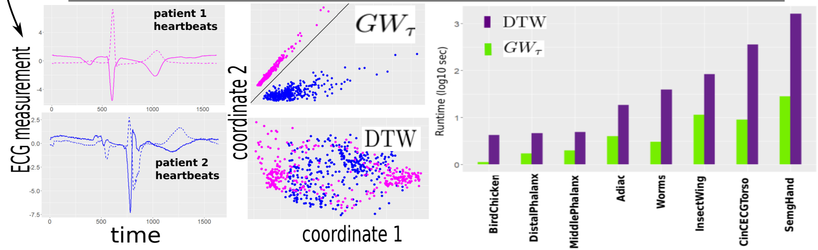
1. GW_τ is an upper bound of GW.
   Open question: $TLB \le GW \le GW_\tau$
2. GW_τ is equivalent to Wasserstein distance between *local distributions of distance (Mémoli 2011)* at the start of each trajectory
3. GW_τ is a metric on the space of (certain) equivalence classes of trajectories
4. Similar construction is defined for graphs in *Le, Ho, & Yamada 2022*
5. Can be computed in linear time (time series of the same length) or quadratic time (different lengths)

Performance of $GW_{\mathcal{T}}$ distance between time series: EEG dataset

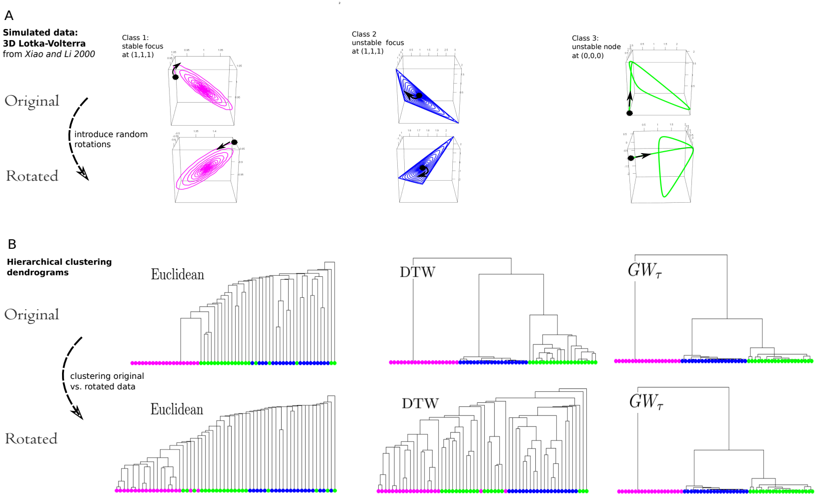


Performance of GW_τ distance between time series: 1-Nearest Neighbor classification of UCR Time Series Classification Archive data (*Dau et al. 2018*)

UCR dataset name	# classes	t.s. length	train size	test size	Euclidean error	GW_τ error	DTW error
CinCECGTorso	4	1639	40	1380	0.1029	0.1290*	0.3493
InsectWingbeatSound	11	265	220	1980	0.4384	0.5995*	0.6449
DistalPhalanxOutlineAgeGroup	3	80	400	139	0.3741	0.3237*	0.2302
Worms	5	900	181	77	0.5455	0.5325*	0.4156
Adiac	37	176	390	391	0.3887	0.3555**	0.3964
BirdChicken	2	512	20	20	0.4500	0.1500**	0.2500
MiddlePhalanxOutlineAgeGroup	3	80	400	154	0.4805	0.4740**	0.5000
SemgHandMovementCh2	6	1500	450	450	0.6311	0.3933**	0.4156

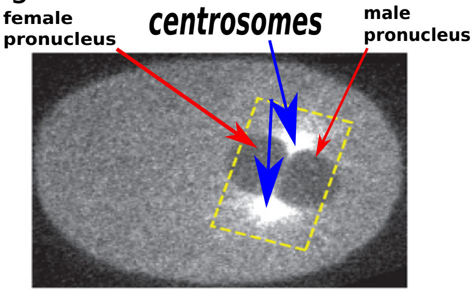


Performance of GW_τ distance between time series: 3D Lotka-Volterra dynamical system (*Xiao & Li 2000*) simulated data



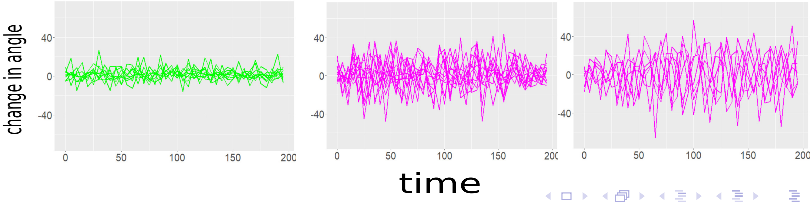
Performance of GW_τ distance between time series: data from Dawes lab (*Ignacio et al. 2022*)

C. elegans zygote

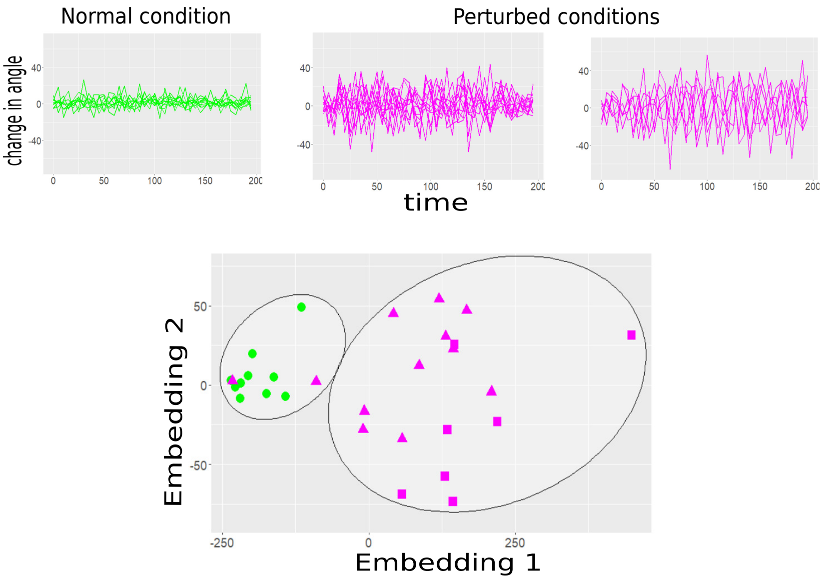


Normal condition

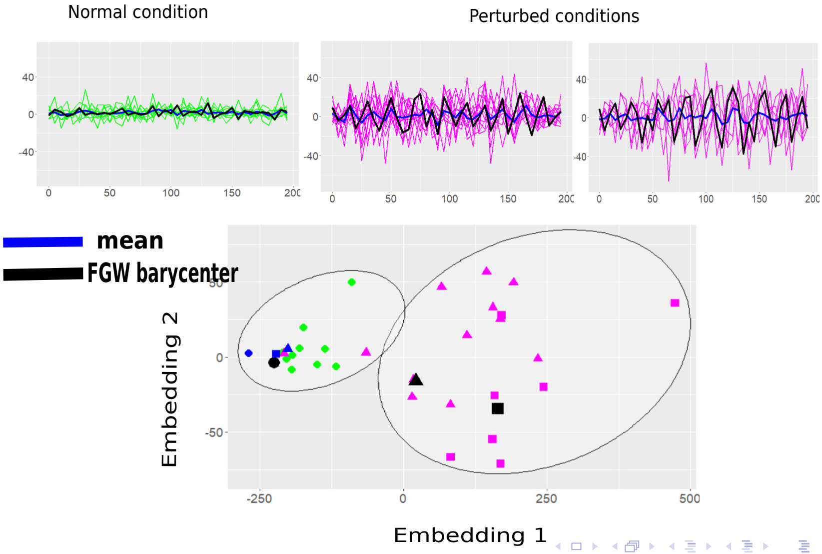
Perturbed conditions



Embedding result



Averaging result: on the use of Fused Gromov-Wasserstein barycenters of *Vayer et al. 2020*



References

1. Dua, D., & Graff, C. (2017). *UCI machine learning repository*
2. Nagel, Sebastian. (2019) *10.15496/publikation-37739*
3. Mémoli, F. (2011). *Found. Comput. Math.*, 11 (4)
4. Peyré, G., Cuturi, M., Solomon, J. (2016). *ICML*, 48
5. Chowdhury, S., & Mémoli, F. (2019). *Inf. Inference*, 8(4)
6. Weitkamp, C. A., Proksch, K., Tameling, C., & Munk, A. (2022). *J. Am. Stat. Assoc.*
7. Dau, H.A., Keogh, E., Kamgar, K., Yeh, C.-C.M., Zhu, Y., Gharghabi, S., ... Hexagon-ML (2018, October). *The UCR time series classification archive*
8. Xiao, D., & Li, W. (2000) *J. Diff. Equ.*, 164 (1)
9. Ignacio, D. P., Kravtsova, N., Henry, J., Palomares, R. H., & Dawes, A. T. (2022). *Cytoskeleton*
10. Coffman, V. C., McDermott, M. B., Shtylla, B., & Dawes, A. T. (2016) *Mol. Biol. Cell* 27(22)
11. Vayer, T., Chapel, L., Flamary, R., Tavenard, R., & Courty, N. (2020) *Algorithms*, 13 (9)

Acknowledgements and Funding

Paper: Kravtsova, McGee II, Dawes (2023), *Bull. Math. Biol.*

Authors thank:

Anonymous reviewer

Members of the Dawes Lab:

Will Burks
Ashley Castelloe
Andrew Cohen
Arkaprovo Ghosal
David Ignacio
Graham Mackey
Liam O'Brien
Shayne Plourde
Nethan Ramachandran
Caroline Tatsuoka

Members of the Chamberlin Lab

Work supported by National Institutes of Health (NIH) award to Adriana T. Dawes

National Institutes of Health

Upcoming SMB 2023 presentations from Dawes Lab

Thursday at 6:00, Archie Griffin Ballroom:

Liam O'Brien *Changes in Approximate Symmetries of a Parametrized Turing Pattern*

Poster ID MFBM-10

Caroline Tatsuoka *Data Driven Modeling of Biological Systems with Deep Neural Networks*

Poster ID MFBM-17

End of Presentation

Thank you!

Questions?