Supporting Information for:

Similarity Metrics for Sub-Cellular Analysis of FRET Microscopy Videos

Michael J. Burke, Victor S. Batista*, Caitlin M. Davis*

Department of Chemistry, Yale University, New Haven, CT, 06520

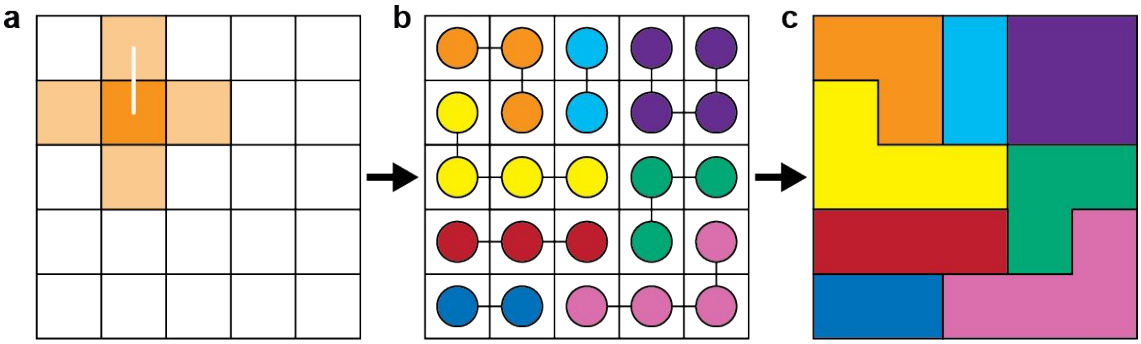

Figure S1. Local Hierarchical Agglomerative Clustering (L-HAC). a) Each neighboring pixel is identified and its similarity to the center pixel is calculated. b) The highest similarity to each neighbor is tracked and the connections are created for each pixel in the image. c) The connections are then averaged together to create new pixel groups. This process can be iterated with neighboring groups to further reduce the number of pixels in the image.

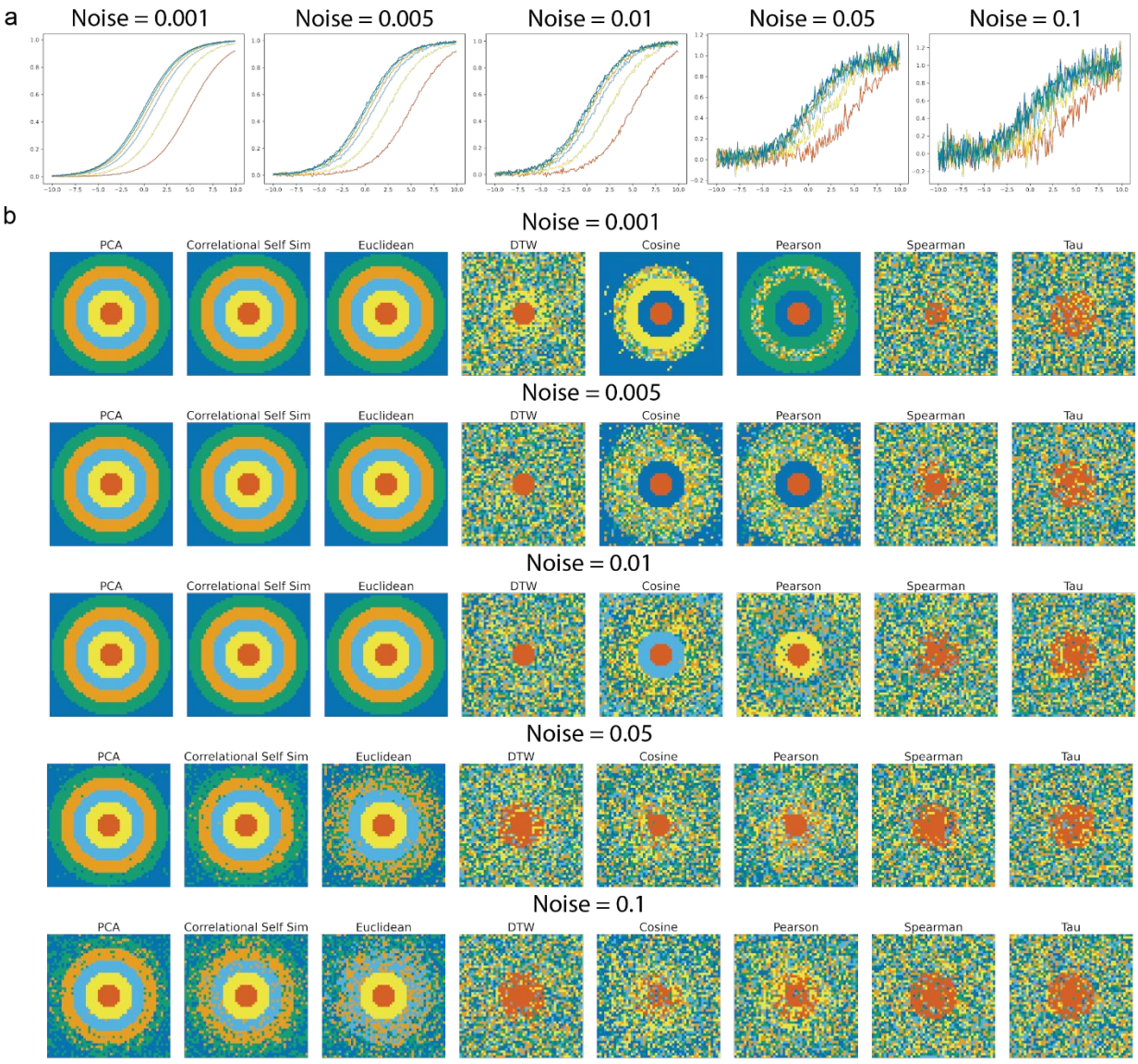

Figure S2. Clustering as a function of Gaussian noise. The data from Figure 2c was reanalyzed with varying Gaussian noise levels. A 51 x 51 pixel grid was subdivided into six regions, each simulating a sigmoid function whose parameters are varied across groups, as detailed in Table S1. a) Representative time series from each region of the grid, with color matching the respective region. Gaussian noise is increased from left to right. b) The clustering outcomes of each similarity metric, tested against simulated sigmoidal curves with varying midpoints at the specified Gaussian noise level.

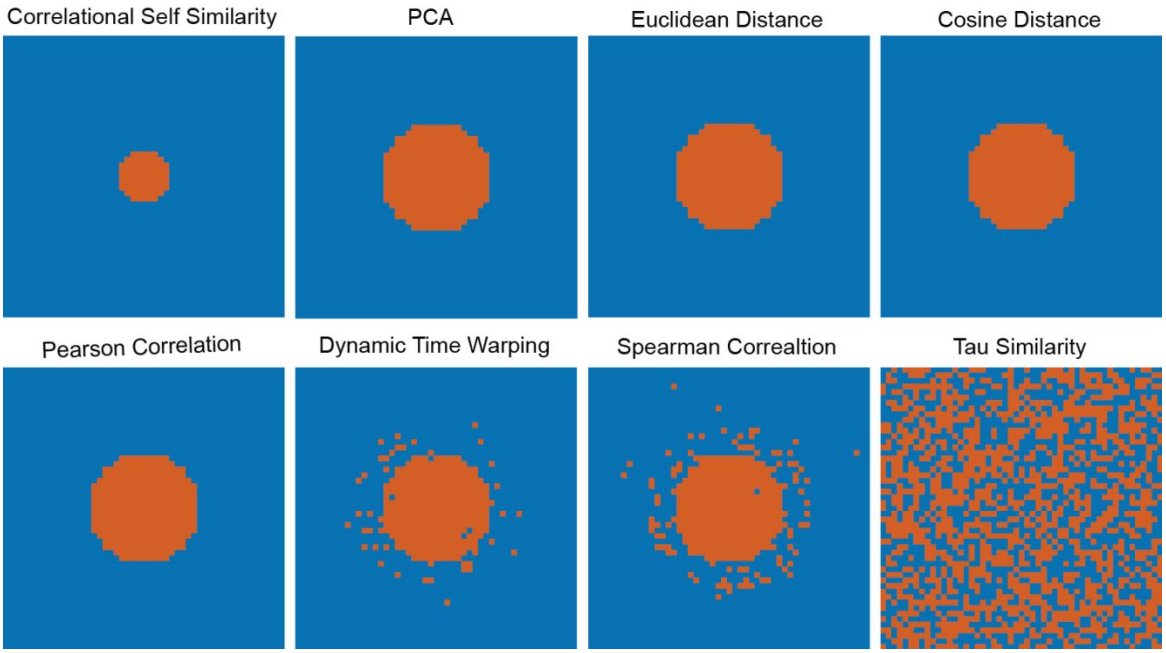

Figure S3. The data presented in main text Fig. 1A is clustered into two groups. This figure illustrates 2 group clustering outcomes for data derived from six groups of sigmoid functions, each characterized by a systematically shifted midpoint. The simulations span a time range from t = -10 to t = 10 with a step size (Δt) of 0.1. The sigmoid functions are arranged in concentric rings, with the innermost ring having a midpoint at t = 5. Subsequent outer rings have midpoints shifted to t = 2.5, 1.0, 0.5, 0.2, and 0. These represent decreasing incremental differences from the center to the outermost region of 25, 15, 5, 3, and 2 frames, respectively. Correlation Self-Similarity separates out the most different inner most group, which is a 25-frame midpoint shift. PCA, Euclidean Distance, Cosine Distance, and Pearson Correlation group the data from rings with 50- and 25-frame midpoint shifts together, favoring a more even division of the data into two groups. DTW and Spearman perfume a similar general grouping but with less precision. Tau similarity shows complete failure in this task.

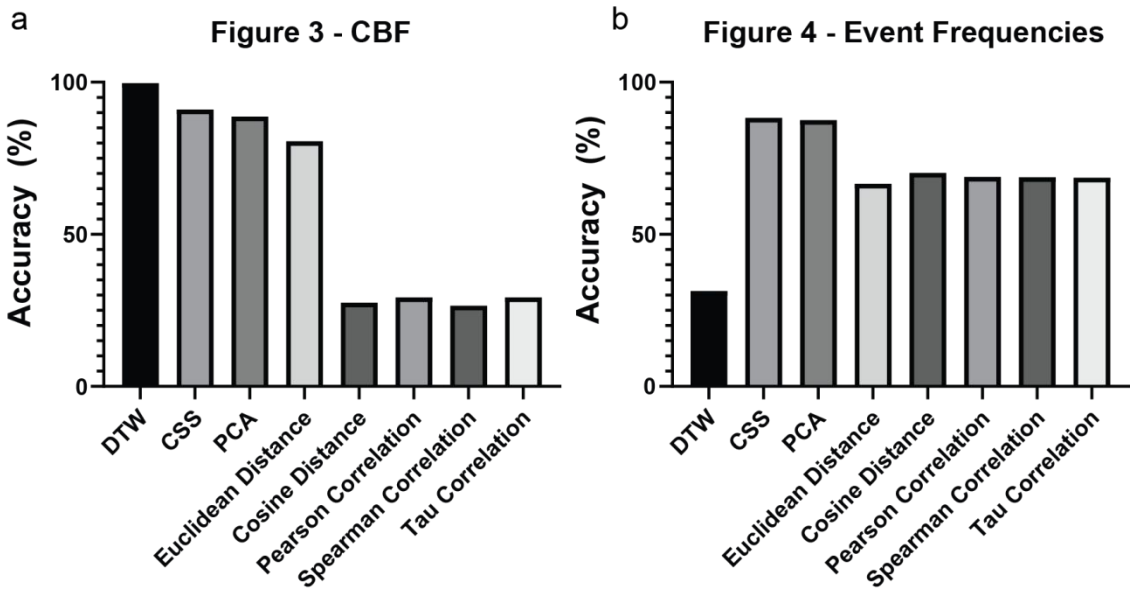

Figure S4. Accuracy. Accuracy values from main text (a) figure 3 and (b) figure 4 are plotted for easy comparison. a) Within the cylinder, funnel, bell dataset (Figure 3) DTW was the most accurate followed by CSS, then PCA, then Euclidean Distance. b) Within the event frequency dataset, CSS and PCA were the most accurate followed by a second group of Cosine Distance, Pearson Correlation, Spearman Correlation, Tau Correlation, and Euclidean Distance. DTW was the least accurate in classifying differences in event frequency.

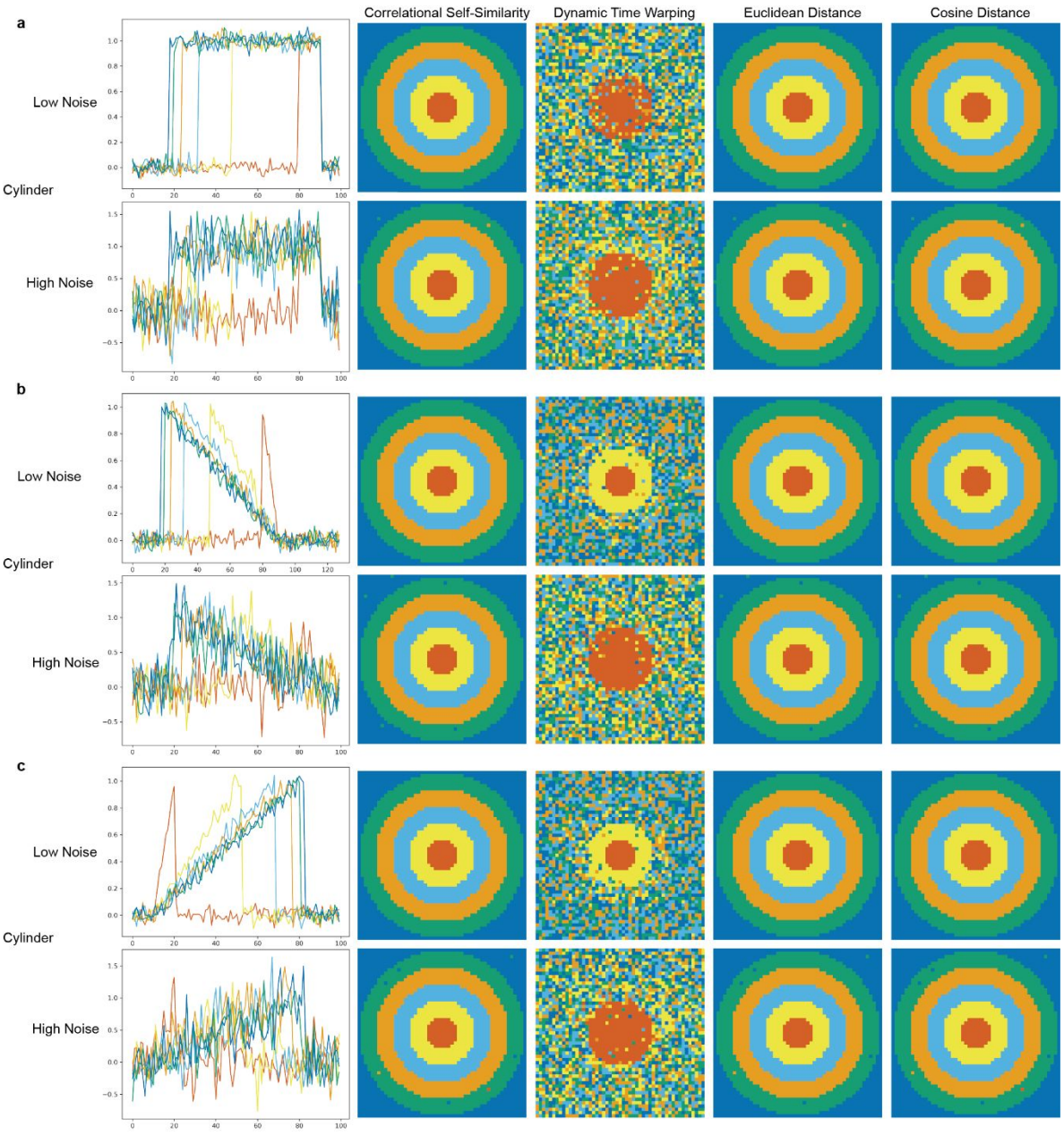

Figure S5. Comparative analysis of clustering performance across different similarity metrics for function types with varied noise levels. This figure illustrates the clustering capabilities of various similarity metrics when analyzing variations of the same function each of the three different mathematical function types, (a) cylinder, (b) funnel, and (c) bell, in the CBF data set. Each subjected to alterations in their parameters under low and high noise conditions. The first column for each type displays time-series data for the functions, with each line color representing a specific region from the clustered results. The other columns show the visualization of clustering results.

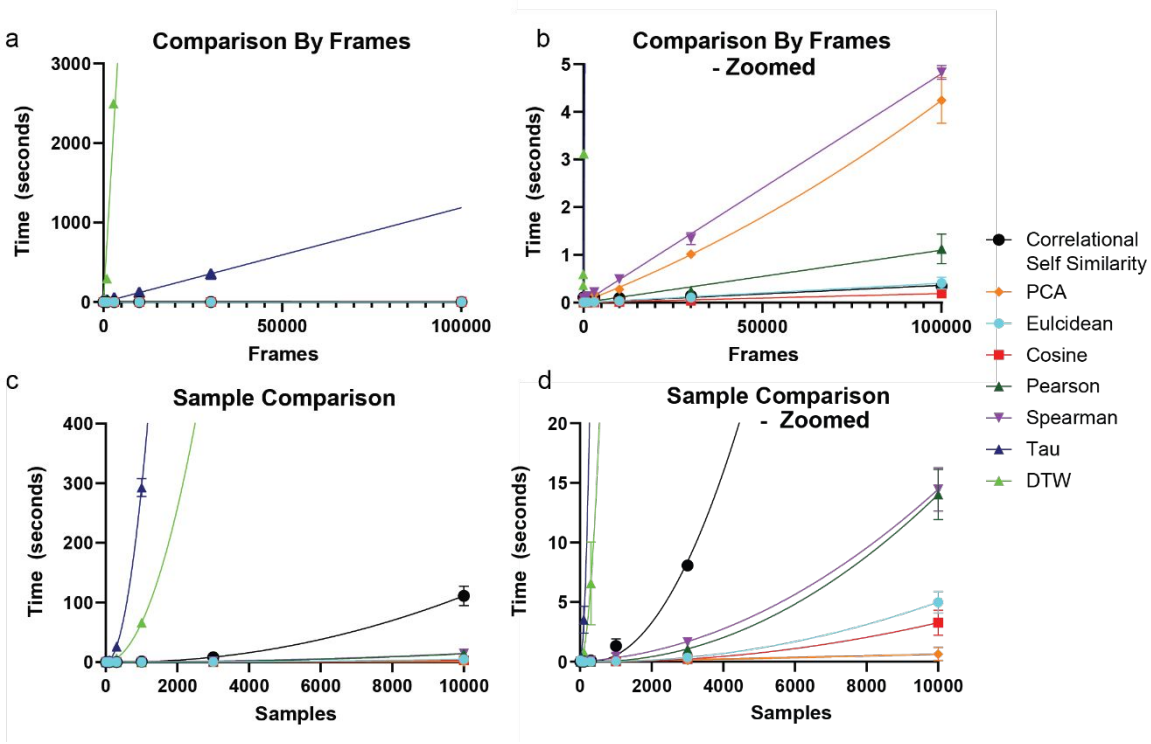

Figure S6. Time scaling with different amounts of samples and frames This figure illustrates how long it takes to form the adjacency matrix for each similarity metric. This depends on the number of samples and frames. There is always some overhead in time spent assigning variables, but generally the methods scale linearly with respect to the number frames (a and b) and quadratically with respect to number of samples (c and d). For both comparisons the other parameter was held at a constant value of 300. DTW and Kendall's Tau were not calculated for the higher sample and frame comparisons due to their poor time scaling. Both DTW and Kendall's Tau (a and c) are significantly slower than the other methods, which must be zoomed in on to distinguish (b and d). Euclidean, Cosine, and Correlational Self Similarity scale the best with respect to number of frames. PCA scales the best in regard to number of samples.

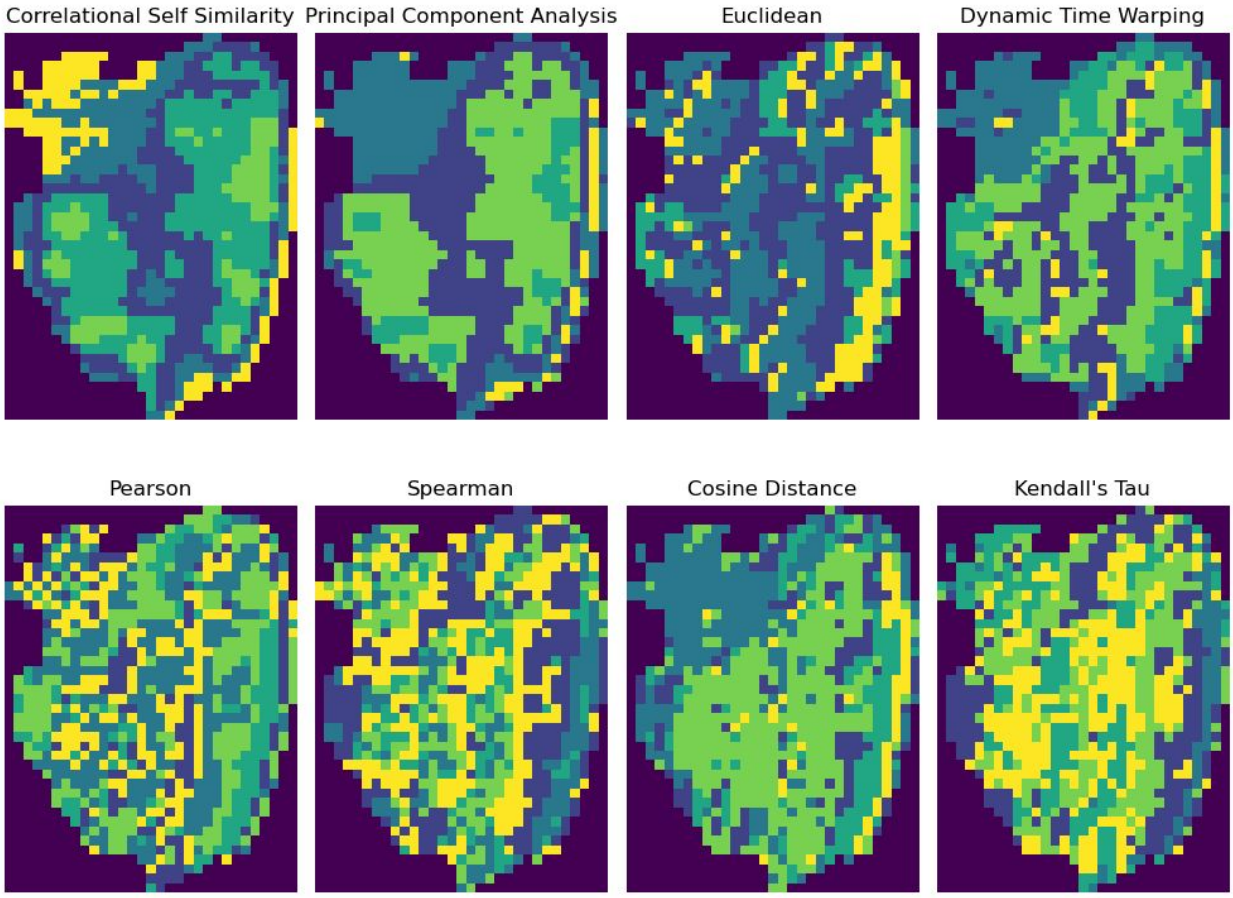

Figure S7. Clustering Result for U1A-SL2 Binding Dynamics for Different Similarity Metrics. The clustering results for each similarity metrics are shown. Correlational Self-Similarity and Principal Component Analysis both show the Cajal bodies with high similarity.

Table S1. Parameters for each simulation in Fig 1.

		Function	Region	Parameters				Range				Noise Std Dev
				Xo	L	k	b	Initial Point	End Point	Step Size	Total Points	
Sigmoid	Midpoint - (Fig 1 A, B and D i)	$f(x) = \frac{L}{1 + e^{-k(x - x_0)}} + b$		10	1	0.5	0	-10	10	0.1	200	0.05
				3								
				1								
				0.5								
				0.2								
				0								
	Rate - (Fig 1 D ii)	$f(x) = \frac{L}{1 + e^{-k(x - x_0)}} + b$		0	1	5	0	-10	10	0.1	200	0.05
						2.5						
						1						
						0.75						
						0.6						
						5						

		Function	Region	Parameters			Range				Noise Std Dev
				A	k	b	Initial Point	End Point	Step Size	Total Points	
Exponential Decay	Amplitude - (Fig 1D iii)	$f(x) = \boldsymbol{A}e^{-kx} + b$		3	1	0	0	5	0.1	50	0.05
				2							
				1.5							
				1.25							
				1.1							
				1							
	Rate - (Fig 1D iv)	$f(x) = Ae^{-\boldsymbol{k}x} + b$		1	3	0	0	5	0.1	50	0.05
					2						
					1.5						
					1.25						
					1.1						
					1						

Table S2. Formulation and time complexity of similarity metrics

Name	Measures	Explicit Formulation	Time Complexity
Euclidean Distance	The distance between two points in N-Dimensional space	$\sqrt{\sum (x_i - y_i)^2}$	O(n)
Cosine Similairty	The angle between two points in N-dimensional space	$\frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\|\|\mathbf{Y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$	O(n)
Correlational Self-Similarity	The distribution of numbers in the vector	$\frac{\sum_{n=0}^{w} Corr\ (F * G)[n]}{Max\ (\sum_{n=0}^{w} Corr\ (F * F)[n], \sum_{n=0}^{w} Corr\ (G * G)[n])}$ $Corr(F * G)[n] = \sum_{m=0}^{N-1} F[m]g[m+n]$	O(wn)
Dynamic Time Warping	The minimum acheivable distance through warping the two curves	$\mathrm{DTW}(X,Y) = \min_{\pi \in A(X,Y)} \sum_{(i,j)\in\pi} \|x_i - y_j\|$	$O(n^2)^*$
Pearson Correlation	The linear correlation of two vectors	$\frac{\mathbb{E}[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}$	O(n)
Spearman Correlation	The linear correlation of two ranked vectors	$\frac{\mathbb{E}[(X_r - \mu_{x_r})(Y_r - \mu_{y_r}))]}{\sigma_{x_r}\sigma_{y_r}}$	O(n log n)
Kendall's Tau Coeffiecient	Rank Correlation between two vectors move	$\frac{2}{n(n-1)} \sum_{i<j} \mathrm{sgn}(x_i - x_j)\,\mathrm{sgn}(y_i - y_j)$	O(n log n)