ViDa: Visualization of DNA Reaction Trajectories Using Deep Graph Embeddings

Chenwei Zhang (*cwzhang@cs.ubc.ca*), Jordan Lovrod, Boyan Beronov, Khanh Dao Duc, Anne Condon

Motivation & contributions

- Existing visualization tools for DNA kinetics are often restricted to a class of reactions, limited in their resolution of individual secondary structures, and unable to capture unexpected patterns.
- Our framework ***ViDa*** creates low-dimensional embeddings of secondary structure state spaces using the output of kinetics simulators, e.g., *Multistrand* [1].
- ViDa's state embeddings are smoother (with less "jumpy" trajectory embeddings) than GSAE [2], PHATE [3], PCA, and MDS, and can provide insight into:
  i. individual trajectories, and
  ii. overall reaction mechanisms.

Background

Kinetics of (multiple) DNA strands:

- Strands stochastically (un-)bind and (un-)fold from initial towards final secondary structures.
- Usually modelled by a continuous time Markov chain (CTMC)

Reaction trajectory:

- The sequence of states (secondary structures), from the reactants to the products of a single instance of a DNA reaction, along with the time to transition from one state to the next

Deep graph embeddings:

- Nodes/edges of a graph are mapped onto a vector space [2].
- The mapping is defined by a neural network trained to capture structural and semantic relationships in the graph.

PHATE: Non-linear dimensionality reduction approach [3]

Model architecture

Secondary structures are embedded into a real vector space by the scatter transform [4] and passed into a variational autoencoder (VAE) [5]. The VAE loss is augmented with three domain-specific regression terms: first passage time, graph edit distance, and free energy. The dimensionality of the samples is then reduced with PCA/PHATE for visualization.

- Well-preserved global metric structure: high-to-low energy trend
- Well-preserved local metric structure: similar neighbors
- Encoder and decoder: 2 FC layers with RELU & batch norm
- Loss function:

$$L_{tot} = \alpha L_{recon} + \beta L_{D_{KL}} + \gamma L_{pred} + \delta L_{D_t} + \varepsilon L_{D_e}$$

$z_i \equiv learned\ embedding\ of\ state\ i$

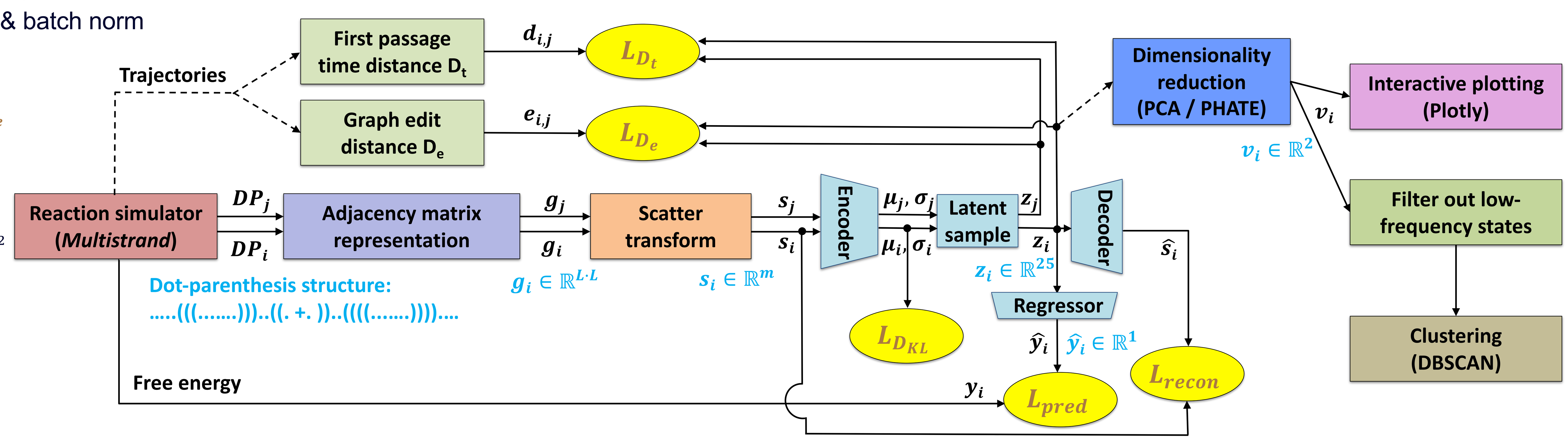
$e_{i,j} \equiv graph\ edit\ distance\ i \rightarrow j$

$$L_{D_t} = \sum_{i,j} W_{i,j}\left(\|z_i - z_j\| - d_{i,j}\right)^2 \qquad L_{D_e} = \sum_{i,j}\left(\|z_i - z_j\| - e_{i,j}\right)^2$$

$d_{i,j} = min\{estimated\ first\ passage\ times\ \ i \rightarrow j, j \rightarrow i\}$

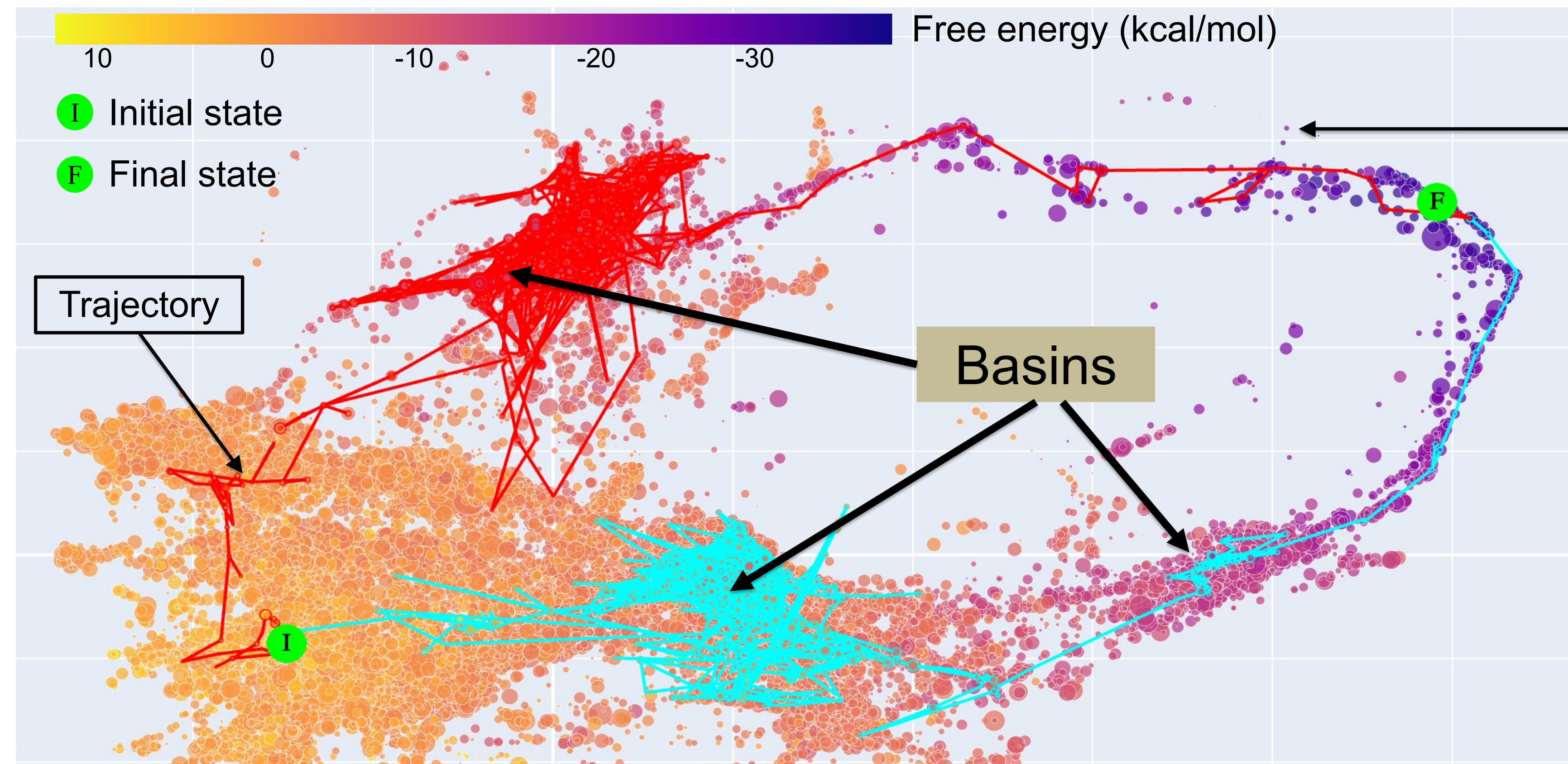
$W_{i,j} = p_i p_j \cdot \mathbb{I}[d_{i,j} \leq \bar{d}]$

$p_i \equiv empirical\ probability\ of\ state\ i$



Case study: Helix association reaction P4+T4 from Gao *et al.* [6]

- Analysis by Schreck *et al.* [7]: 4-stem hairpins impede zipping
- Simulated using *Multistrand* (trajectory mode, default parameters)

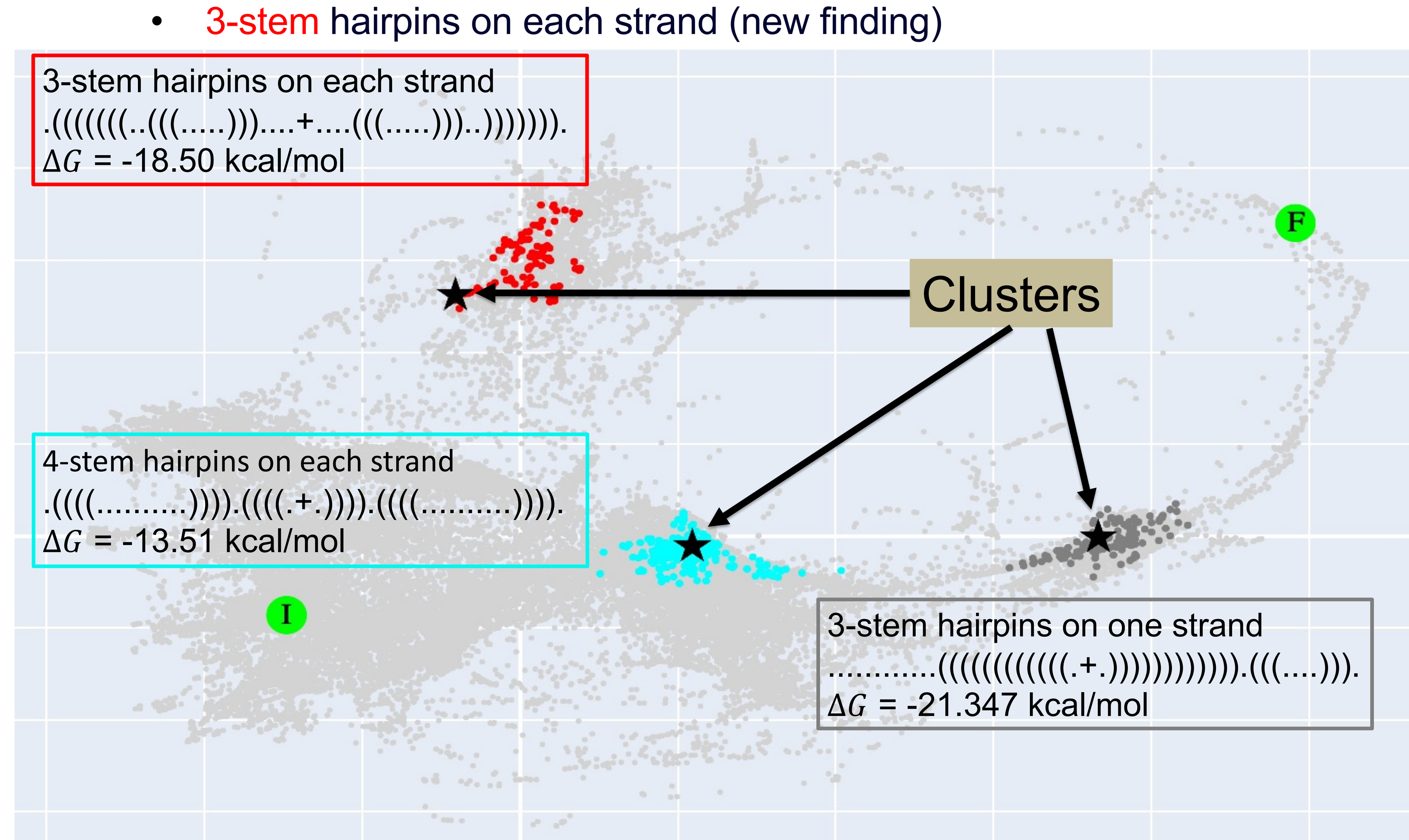


Seq: 3'-ACACGATCATGTCTGCGTGACTAGA-5'+3'-TCTAGTCACGCAGACATGATCGTGT-5'

Hairpin: 3'-.((((..........)))).....-5'+3'-.....((((...........)))).-5'

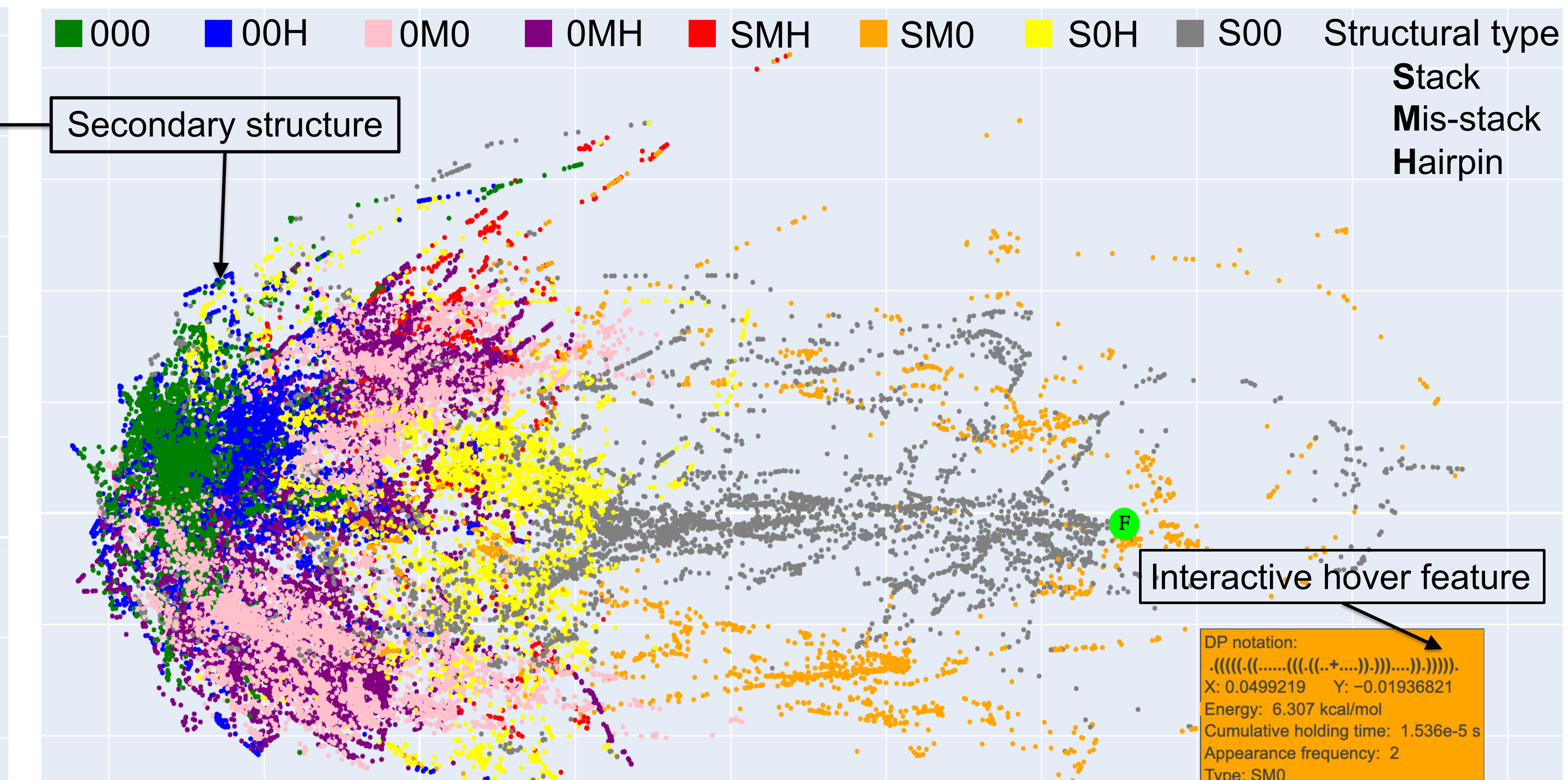
Hybridization rate: $k = 2.0 \pm 0.1 \times 10^5\ M^{-1}s^{-1}$

- Branches of the state space reflect different reaction pathways.
- Two major energy basins with different hairpins:
  - 4-stem hairpins on each strand (consistent with analysis in [7])
  - 3-stem hairpins on each strand (new finding)



Case study: Helix association reaction 39 from Hata *et al.* [8]

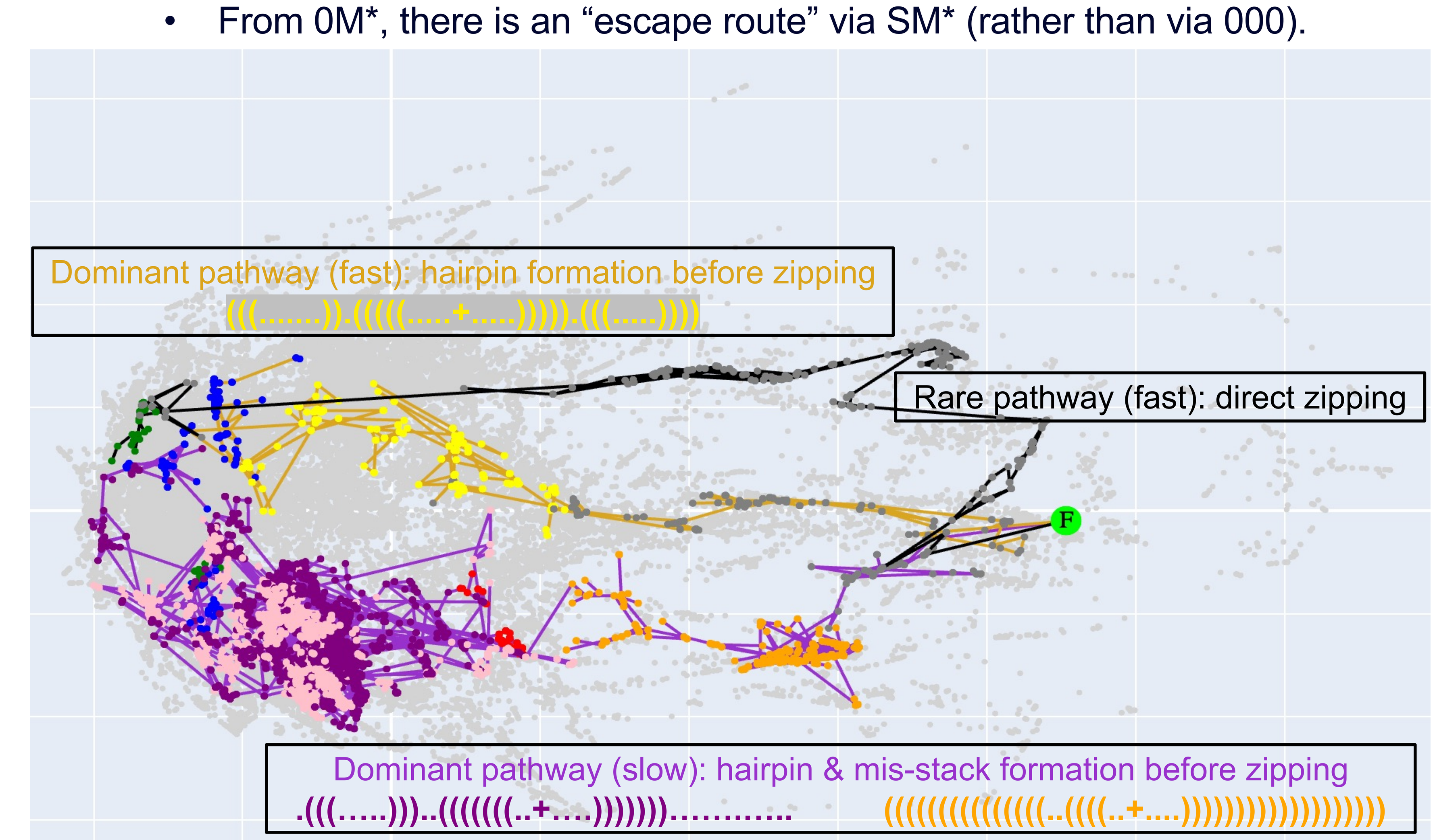
- Analysis by Lovrod *et al.* [9]: mis-stacks and/or hairpins precede zipping
- Simulated using *Multistrand* (first step mode, parameters by Lovrod *et al.* [9])



Seq: 3'-CCATCAGGAATGACACACACAAA-5'+3'-TTTGTGTGTGTCATTCCTGATGG-5'

Hybridization rate: $k = 14.7 \pm 0.4 \times 10^5\ M^{-1}s^{-1}$

- Embeddings distinguish many different reaction mechanisms.
- High probability of forming stable mis-stacks or hairpins (direct zipping is rare)
- Mis-stacking state space regions (0M*) can behave as kinetic traps.
  - From 0M*, there is an "escape route" via SM* (rather than via 000).



References

[1] Schaeffer, J.M., Ph.D. thesis, Caltech, 2013.
[2] Castro, E., *et al.* IEEE Intl. Conf. on Big Data, 2020.
[3] Moon, K.R., *et al.* Nat. Biotechnol., 37 (12): 1482-92, 2019.
[4] Gao, F. *et al.* ICML, pp. 2122-2131, 2019.
[5] Kingma, D.P., *et al.* Found. Trends Mach. Learn, 12(4): 307-392, 2019.
[6] Gao Y., *et al.* Nucleic Acids Res., 34 (11): 3370-77, 2006.
[7] Schreck J.S., *et al.* Nucleic Acids Res., 43(13): 6181-90, 2015.
[8] Hata, H., *et al.* Nucleic Acids Res., 46(2): 782-91, 2018.
[9] Lovrod, J., *et al.* DNA29, 2023.