ViDa: Visualizing DNA reactions with biophysics-informed deep graph embeddings

December 1st, 2023

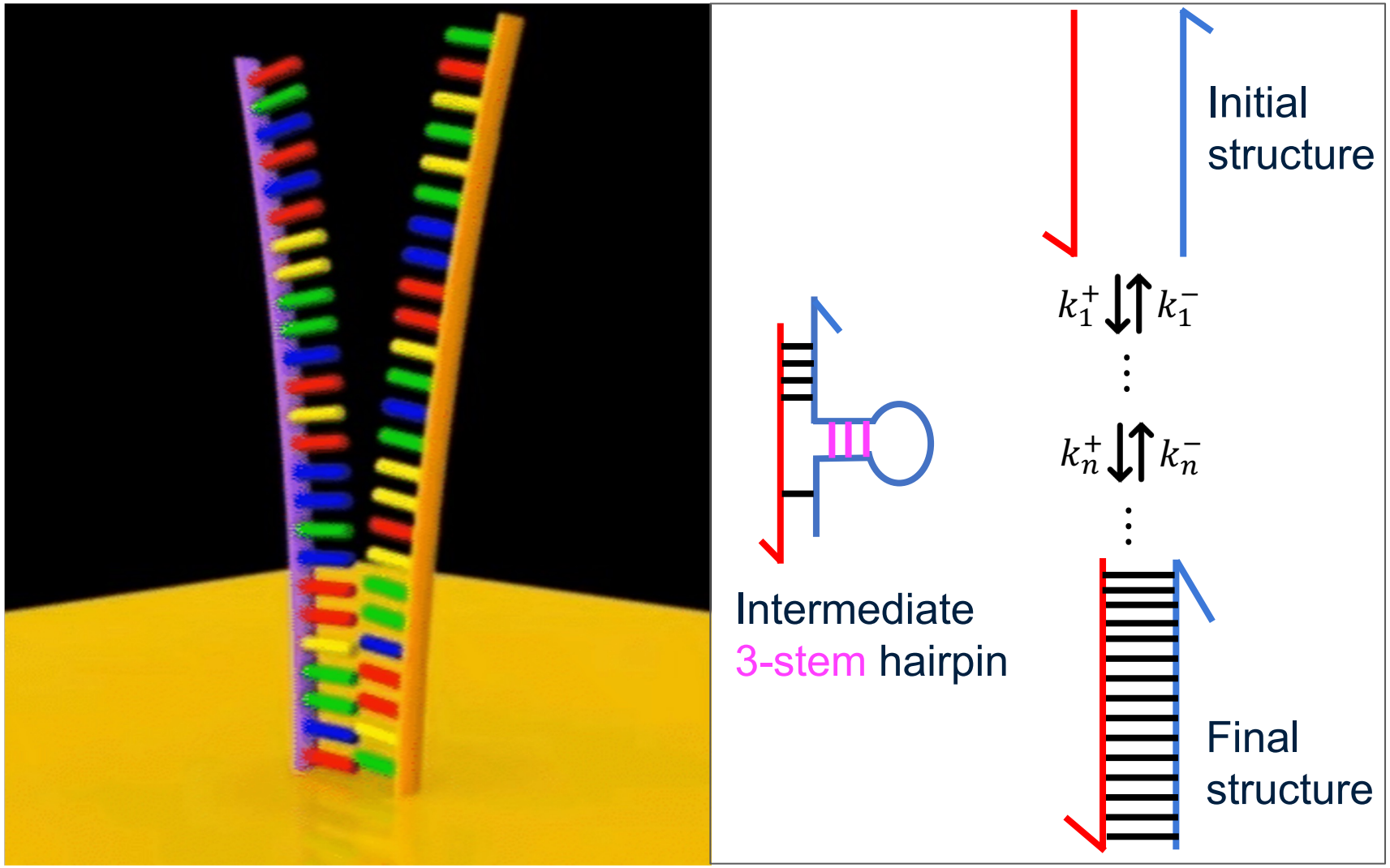
Chenwei Zhang (cwzhang@cs.ubc.ca)[1], Jordan Lovrod[1], Boyan Beronov[1], Khanh Dao Duc[1,2], Anne Condon[1]

THE UNIVERSITY OF BRITISH COLUMBIA

[1] Department of Computer Science, UBC
[2] Department of Mathematics, UBC

MOTIVATION

Substantial number of intermediate states, making the visualization problem challenging!



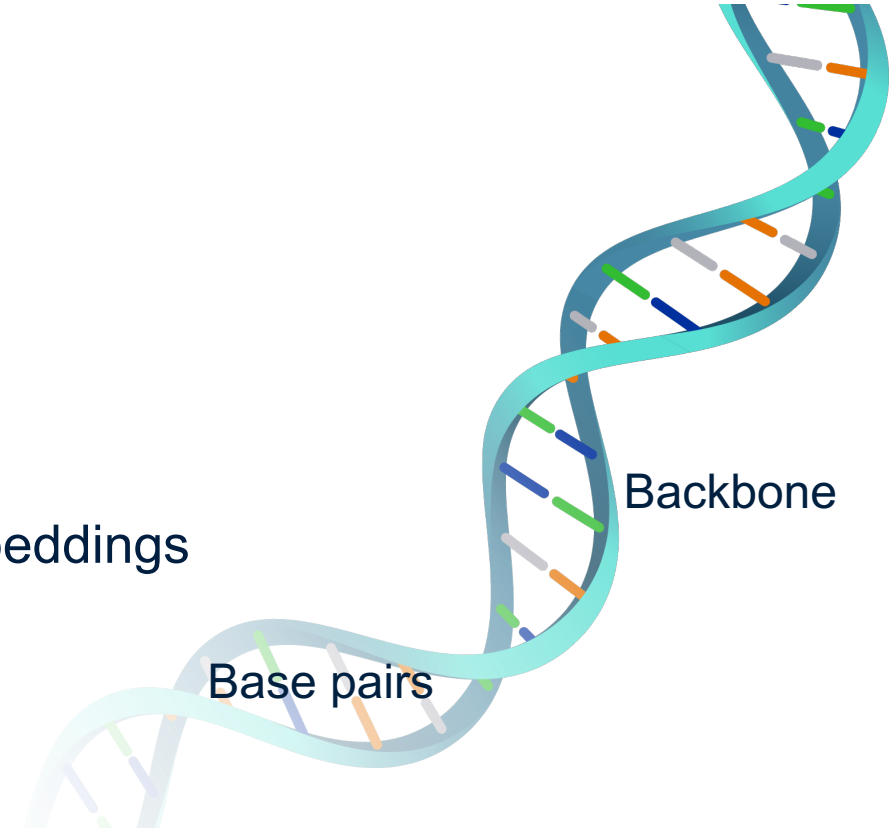
DNA hybridization

https://www.youtube.com/watch?v=0qoqzErrae4

OVERVIEW

1. **Background**
   - Reaction trajectory
   - Multistrand simulator
   - Coarse-grained visualization
2. **Method: ViDa**
   - Model architecture
   - Loss functions
3. **Results**
   - Secondary structure and trajectory embeddings
   - Comparison with other methods
4. **Future work**

Deoxyribonucleic acid (DNA)



https://www.genome.gov/genetics-glossary/Double-Helix

OVERVIEW

1. Background

- Reaction trajectory
- Multistrand simulator
- Coarse-grained visualization

2. Method: ViDa

- Model architecture
- Loss functions

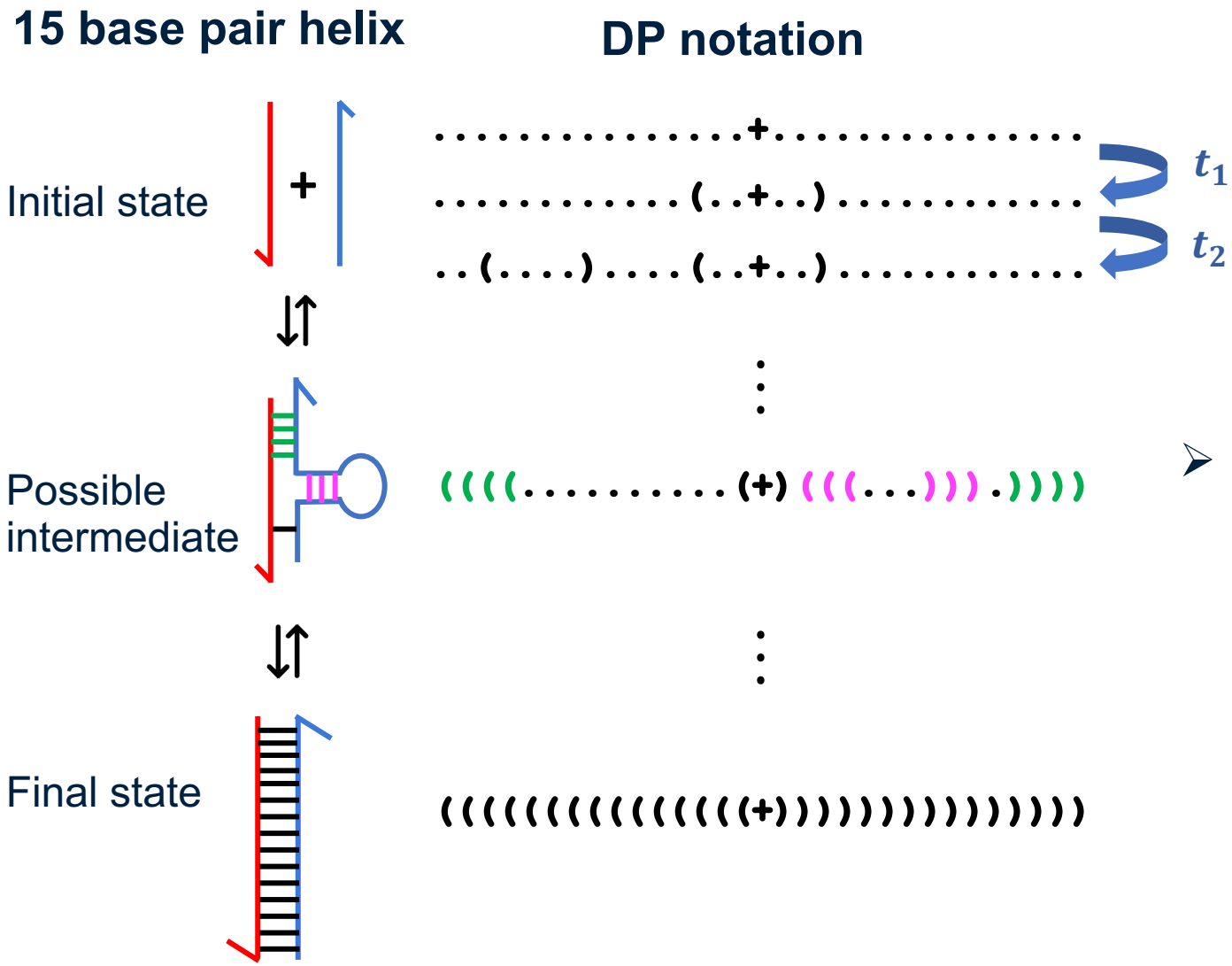
3. Results

- Secondary structure and trajectory embeddings
- Comparison with other methods

4. Future work

Deoxyribonucleic acid (DNA)

BACKGROUND – REACTION TRAJECTORY

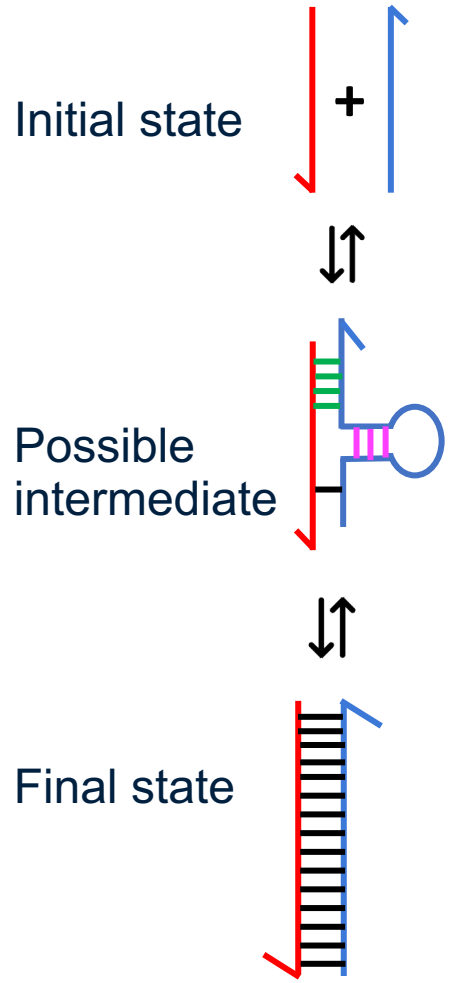


- Dot-parenthesis (DP) notation:
  - Dots: unpaired bases;
  - Parentheses: paired bases;
  - "+" sign: separates two strands

- Reaction trajectory:
  - The sequence of secondary structures, from the reactants to the products of a DNA reaction, along with the time to transition from one state to the next.

BACKGROUND – MULTISTRAND SIMULATOR

- **Multistrand:** DNA elementary step kinetics simulator[1]



Gao-P4T4 (25 bases per strand)[2]

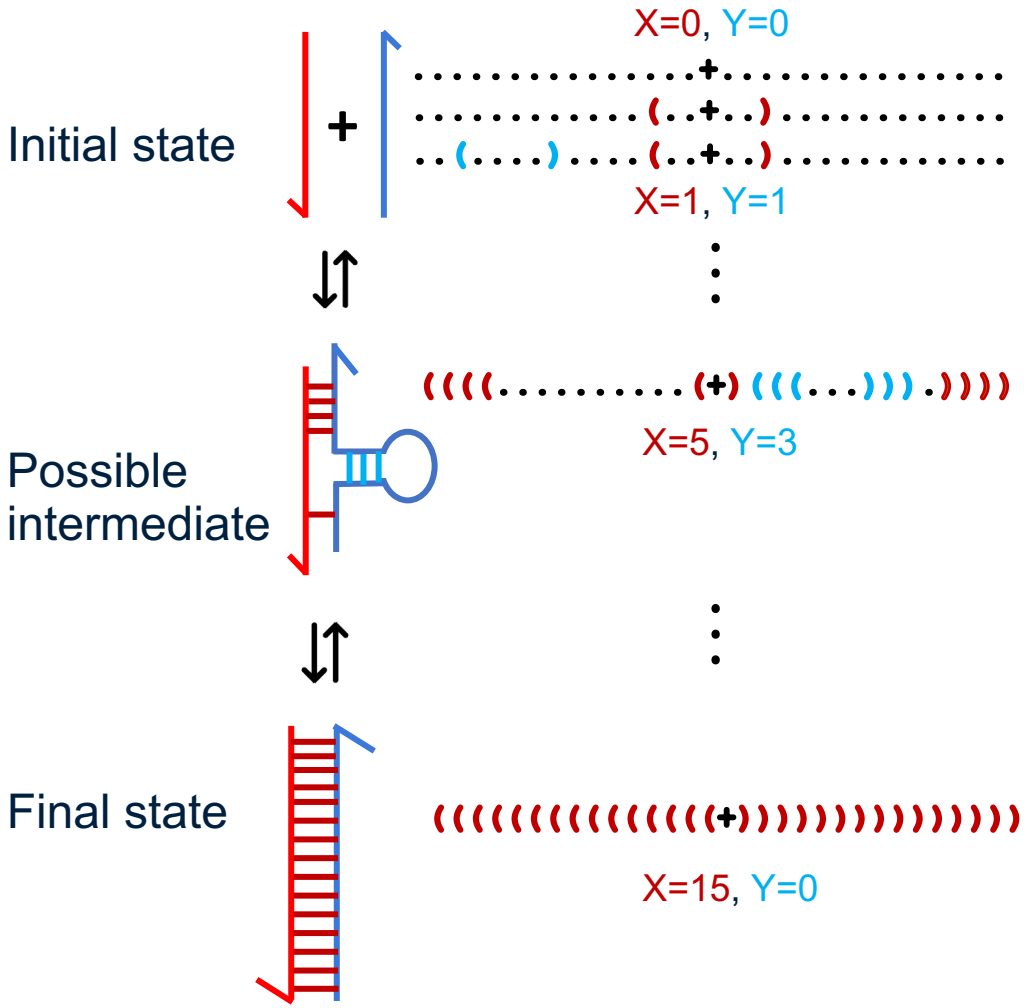
sequences:	3′-ACACGATCATGTCTGCGTGACTAGA-5′ + 3′-TCTAGTCACGCAGACATGATCGTGT-5′
possible hairpins (size 3):	3′-.....(((.....)))....-5′ + 3′-....(((.....)))..... -5′
possible hairpins (size 4):	3′-.((((.....))))......-5′ + 3′-.....((((.....)))).-5′

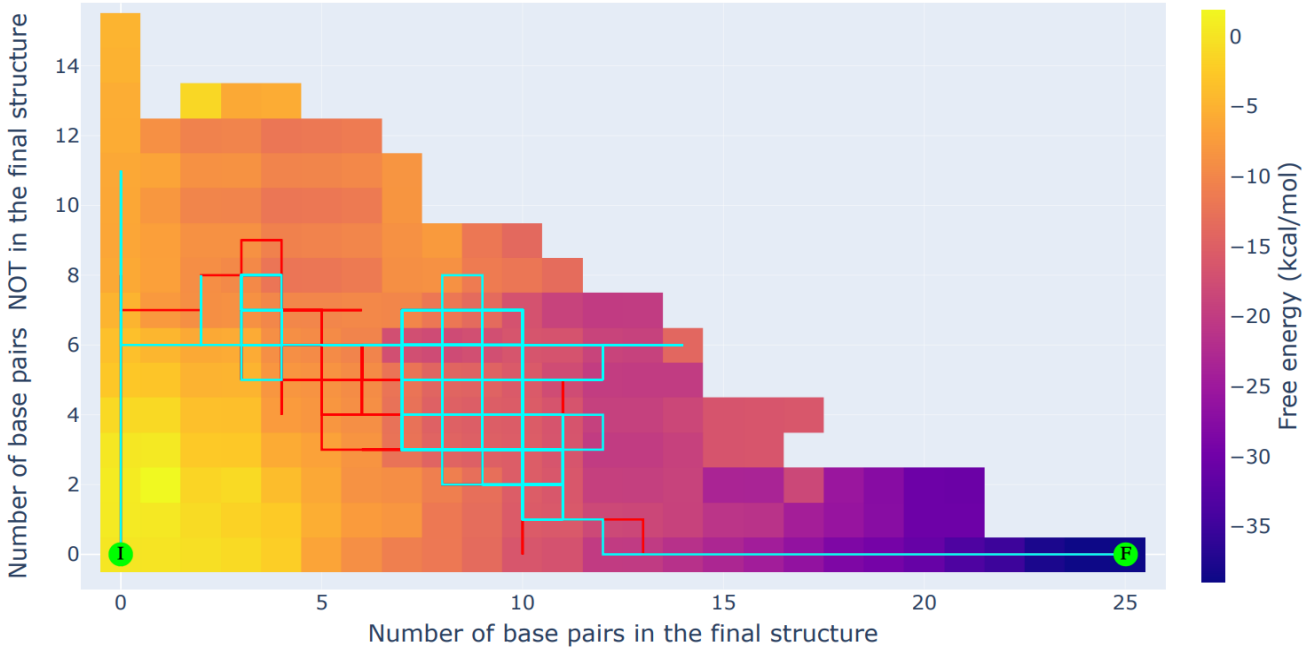
```
.........................+.........................  t=0.000000000 seconds, dG=  0.00 kcal/mol
.(.....................+......).....................  t=0.000000001 seconds, dG=  1.03 kcal/mol
......(.................+.)..............(.....)..  t=0.000000002 seconds, dG=  1.79 kcal/mol
......(........(.......).+.)..............(.....)..  t=0.000000002 seconds, dG=  3.78 kcal/mol
......(........(.......).+.).....................  t=0.000000002 seconds, dG=  3.02 kcal/mol
......(........(.......).+.)..............(.....)..  t=0.000000003 seconds, dG=  3.78 kcal/mol
....(.(........(.......).+.)...)..........(.....)..  t=0.000000003 seconds, dG=  4.25 kcal/mol
....(((........(.......).+.)..))..........(.....)..  t=0.000000004 seconds, dG=  3.47 kcal/mol
....(((........(.......).+.)..))..................  t=0.000000005 seconds, dG=  2.71 kcal/mol
....(((........(.......).+.)..))......(....).......  t=0.000000007 seconds, dG=  3.44 kcal/mol
....(((........(.......).+.)..))..................  t=0.000000008 seconds, dG=  2.71 kcal/mol
....(((........(.......).+.)..))..(...........)....  t=0.000000009 seconds, dG=  4.26 kcal/mol
....((((......(.......).+))..))..(...........)....  t=0.000000010 seconds, dG=  2.95 kcal/mol
....((((..................+))..))..(...........)....  t=0.000000011 seconds, dG=  0.96 kcal/mol
....((((.....(.....).....+))..))..(...........)....  t=0.000000014 seconds, dG=  1.49 kcal/mol
.
.
.
((((((((((((.....).....+...........))))))))))))  t=0.000012852 seconds, dG=-18.85 kcal/mol
((((((((((((.....+...........))))))))))))  t=0.000012855 seconds, dG=-20.10 kcal/mol
((((((((((((......(....)+...........))))))))))))  t=0.000012855 seconds, dG=-20.19 kcal/mol
((((((((((((.....(....)+...........))))))))))))  t=0.000012856 seconds, dG=-23.92 kcal/mol
((((((((((((....(....)+...........))))))))))))  t=0.000012856 seconds, dG=-25.29 kcal/mol
((((((((((((.........+.........))))))))))))  t=0.000012857 seconds, dG=-25.20 kcal/mol
((((((((((((.(.......+.......).))))))))))))  t=0.000012859 seconds, dG=-24.64 kcal/mol
((((((((((((.........+.........))))))))))))  t=0.000012860 seconds, dG=-25.20 kcal/mol
((((((((((((.........+.........))))))))))))  t=0.000012862 seconds, dG=-25.08 kcal/mol
((((((((((((.........+.........))))))))))))  t=0.000012862 seconds, dG=-28.16 kcal/mol
((((((((((((.......+.......))))))))))))  t=0.000012865 seconds, dG=-28.17 kcal/mol
((((((((((((......+......))))))))))))  t=0.000012865 seconds, dG=-28.55 kcal/mol
((((((((((((.(....+....).))))))))))))  t=0.000012866 seconds, dG=-29.28 kcal/mol
((((((((((((.((...+...)).))))))))))))  t=0.000012872 seconds, dG=-32.18 kcal/mol
((((((((((((.(((..+..))).))))))))))))  t=0.000012879 seconds, dG=-32.60 kcal/mol
((((((((((((((((((..+..))))))))))))))))))  t=0.000012884 seconds, dG=-37.00 kcal/mol
((((((((((((((((((.(+).))))))))))))))))))  t=0.000012884 seconds, dG=-35.64 kcal/mol
((((((((((((((((((((+))))))))))))))))))))  t=0.000012888 seconds, dG=-39.80 kcal/mol
```

[1]J.M. Schaeffer. PhD thesis, California Institute of Technology, 2013.

[2]Gao Y., *et al*. Nucleic Acids Res. 2006.

BACKGROUND – COARSE-GRAINED VISUALIZATION





Limited visual interpretation of physical process:

i. A single "cell" (macrostate) may contain with very different structures;

ii. Overlapping trajectories may represent very distinct pathways.

CHALLENGES

What is a good state embedding for domain experts?

- High resolution:
  - Distinguishability of individual secondary structure embeddings
- Well-preserved global structure (energy landscape)
  - State embedding follows the trend of the course of a reaction
- Well-preserved local structure (state topology)
  - Closeness of embeddings for structurally adjacent states
  - Separation of embeddings for structurally distinct states
- Distinguishable and smooth trajectories laid over the embedded states
  - Separated trajectories and no large jumps along the trajectories

CONTRIBUTIONS

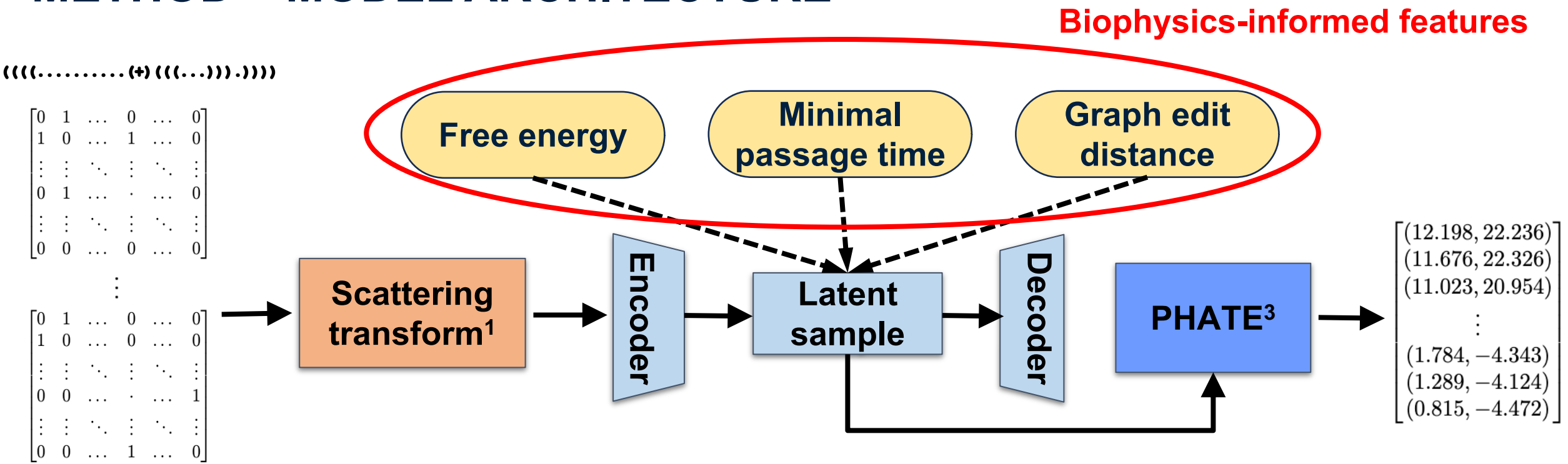
Criterion	Coarse-grained	ViDa (ours)
High resolution	✗	✓
Well-preserved global structure	✓	✓
Well-preserved local structure	✗	✓
Distinguishable and smooth trajectories	✗	✓

OVERVIEW

1. Background
   - Reaction trajectory
   - Multistrand simulator
   - Coarse-grained visualization
2. **Method: ViDa**
   - Model architecture
   - Loss functions
3. Results
   - Secondary structure and trajectory embeddings
   - Comparison with other methods
4. Future work

Deoxyribonucleic acid (DNA)

METHOD – MODEL ARCHITECTURE



Nodes: bases

Edges: base pairs & backbones

Map graph into a vector space (scattering coefficients)

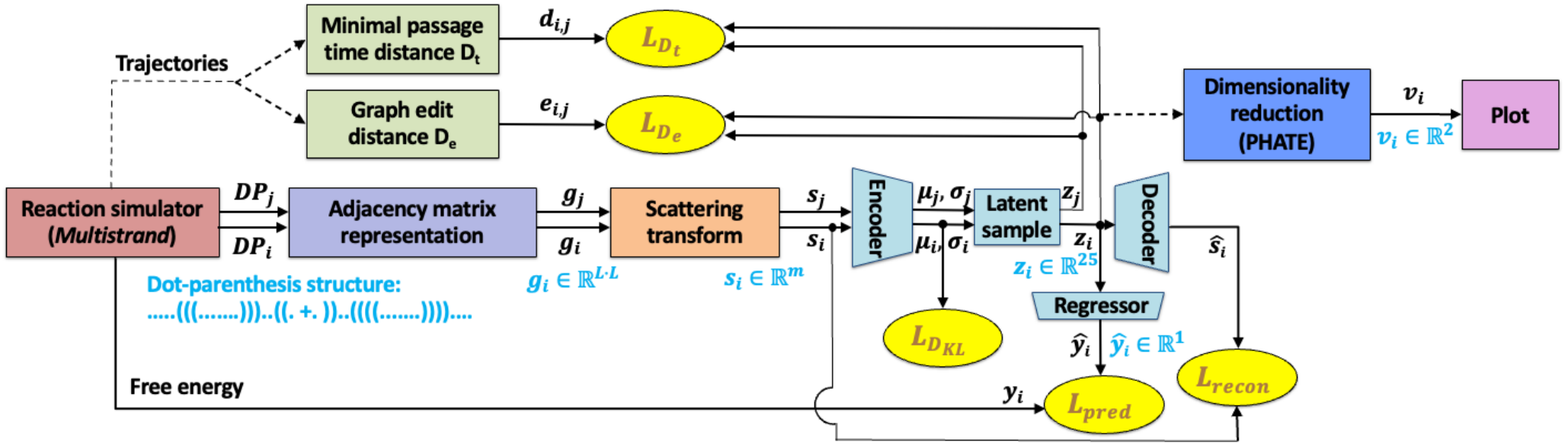
Embed input vectors with a VAE[2] trained by incorporating biophysics-informed features into the loss function

Further reduce dimension to 2D

Plot

1. Gao, F. *et al*. ICML, pp. 2122-2131, 2019.
2. Kingma, D. P. *et al*. Foundations and Trends® in Machine Learning, 12 (4): 307-392, 2019.
3. Moon, K.R., *et al*. Nat. Biotechnol., 37 (12): 1482-92, 2019.

METHOD – LOSS FUNCTION



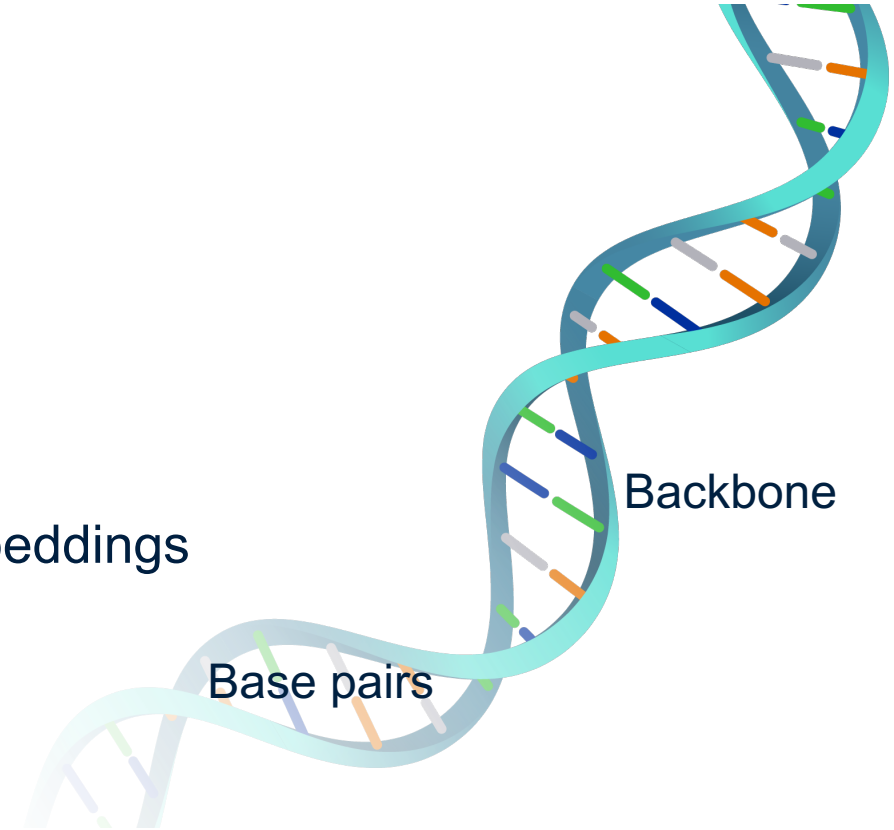
$$L_{tot} = \underbrace{\alpha L_{recon} + \beta L_{D_{KL}}}_{\text{VAE losses (unsupervised)}} + \underbrace{\gamma L_{pred} + \delta L_{D_t} + \varepsilon L_{D_e}}_{\text{Biophysics-informed losses (supervised)}}$$

- Free energy (ΔG) loss: $L_{pred} = \frac{1}{n}\sum_i (\hat{y}_i - y_i)^2$, where $\hat{y}_i$ is the predicted energy; y_i is Multistrand energy.
- Graph edit distance (GED) loss: $L_{D_e} = \sum_{i,j} (\|z_i - z_j\| - e_{i,j})^2$, where $e_{i,j}$ is the graph edit distance between state i and j.
- Minimal passage time (MPT) loss: $L_{D_t} = \sum_{i,j} w_{i,j} (\|z_i - z_j\| - d_{i,j})^2$, where $w_{i,j}$ is an importance weight; $d_{i,j}$ is a normalized estimate of minimal passage time from state i to j or from j to i.

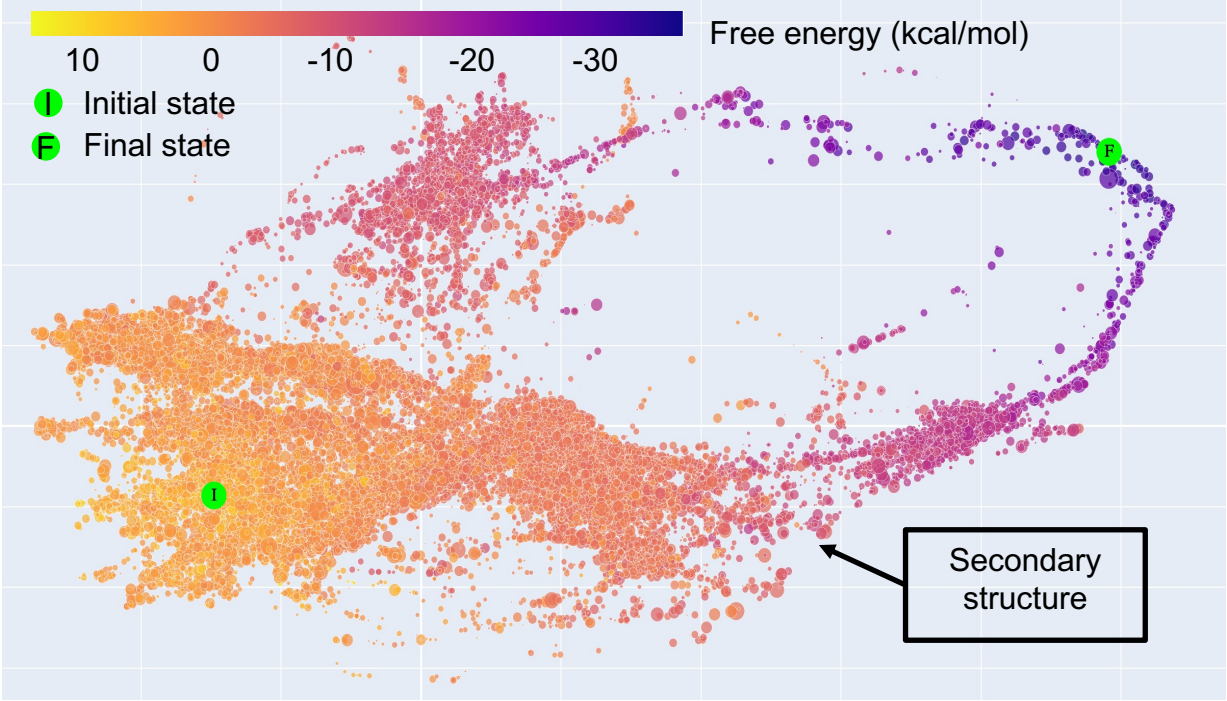
OVERVIEW

1. Background
   - Reaction trajectory
   - Multistrand simulator
   - Coarse-grained visualization
2. Method: ViDa
   - Model architecture
   - Loss functions
3. **Results**
   - Secondary structure and trajectory embeddings
   - Comparison with other methods
4. Future work

Deoxyribonucleic acid (DNA)



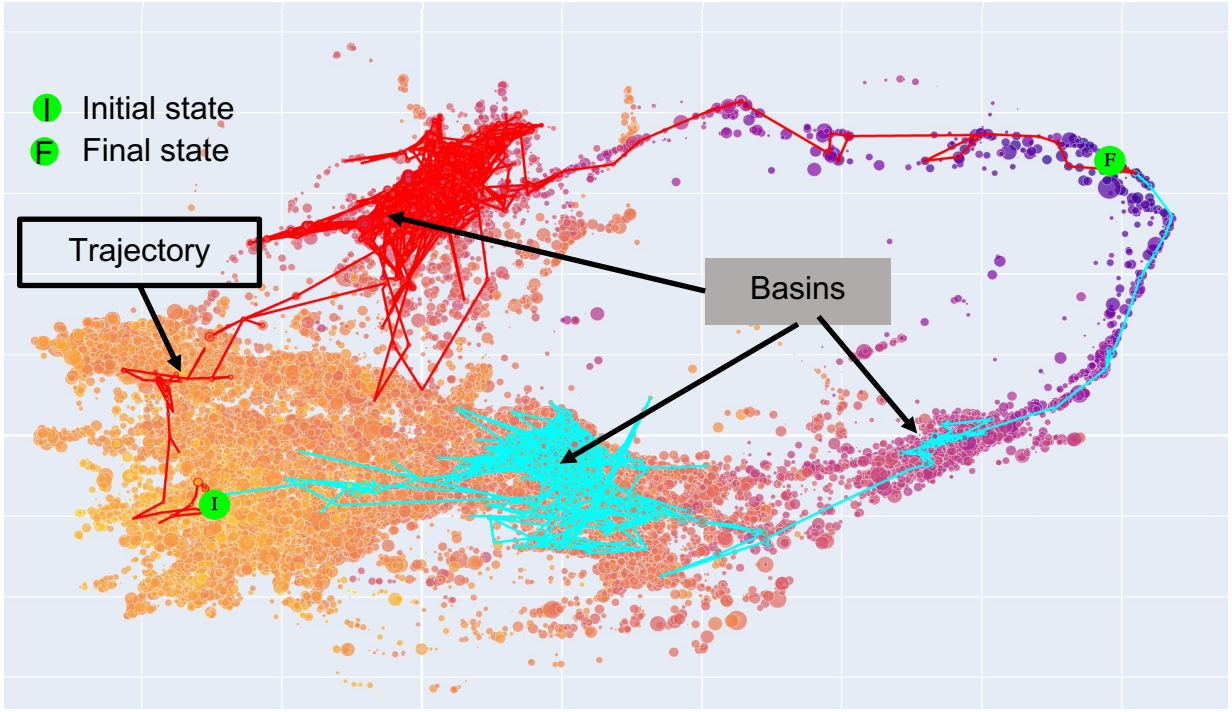
RESULTS – Gao-P4T4



- Well-preserved global and local structure

[1]Schreck J.S., *et al*. Nucleic Acids Res. 2015.

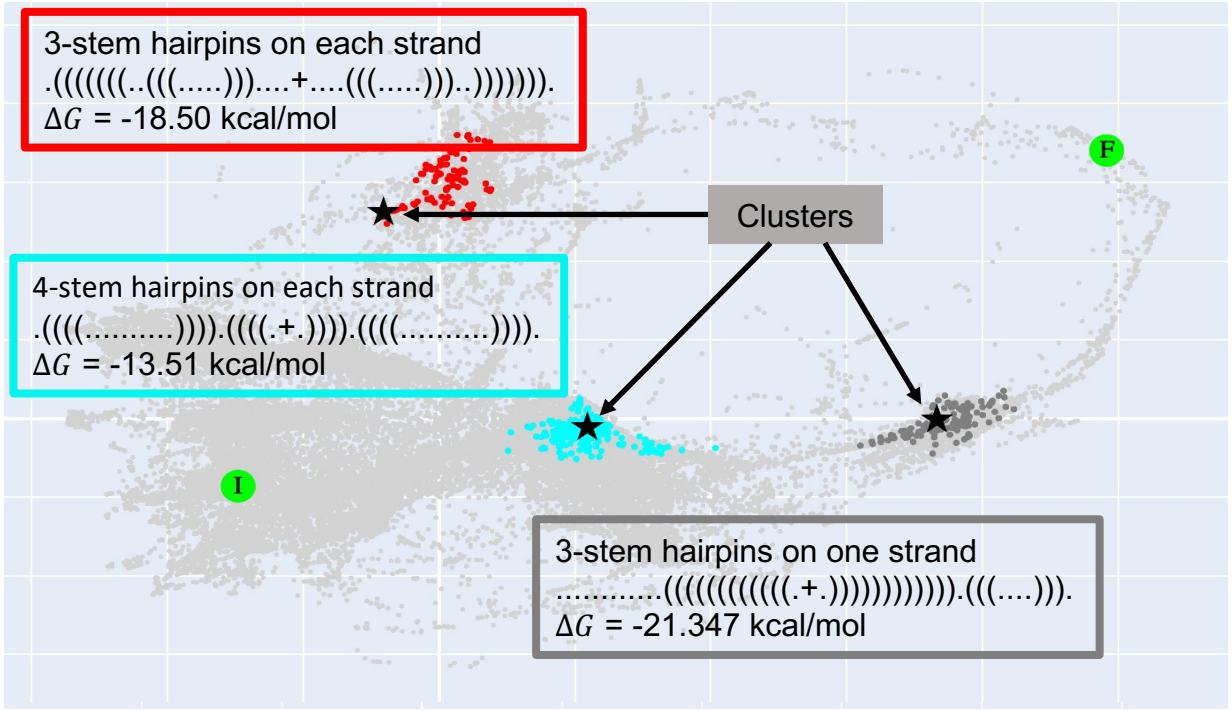
RESULTS – Gao-P4T4



- Well-preserved global and local structure
- Embedded trajectories are distinguishable and smooth

[1]Schreck J.S., *et al*. Nucleic Acids Res. 2015.

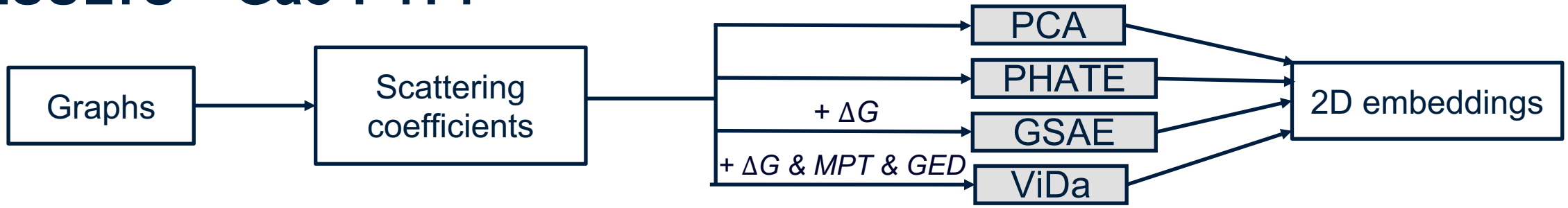
RESULTS – Gao-P4T4



- Well-preserved global and local structure
- Embedded trajectories are distinguishable and smooth
- Two major energy basins with different hairpins
  - 4-stem hairpins on each strand (consistent with analysis by Schreck *et al.*[1])
  - 3-stem hairpins on each strand (**new finding**)

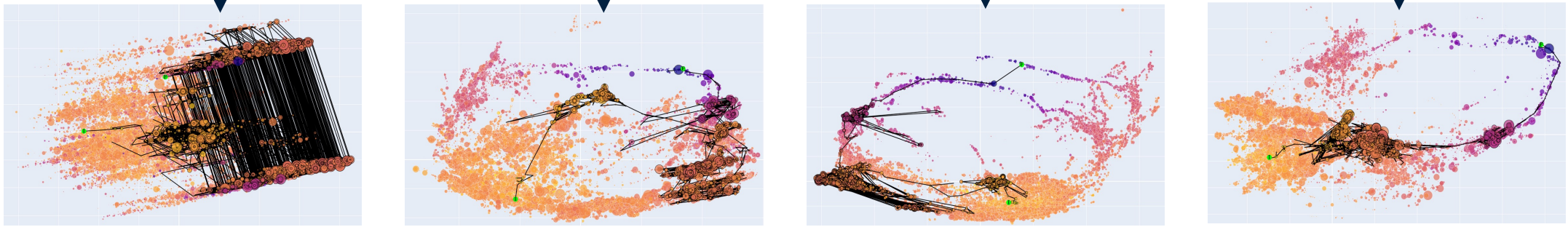
[1]Schreck J.S., *et al*. Nucleic Acids Res. 2015.

RESULTS – Gao-P4T4



- **Average distortion**: the frequency-weighted mean Euclidean distance between consecutive secondary structure pairs in the sampled trajectory dataset

	PCA	PHATE	GSAE+PHATE	**ViDa (ours)**
Avg. distortion	0.159	0.105	0.030	**0.019**



- Integrating biophysics-informed features improves embeddings and reduces trajectory distortion!

OVERVIEW

1. **Background**
   - Reaction trajectory
   - Multistrand simulator
   - Coarse-grained visualization
2. **Method: ViDa**
   - Model architecture
   - Loss functions
3. **Results**
   - Secondary structure and trajectory embeddings
   - Comparison with other methods
4. **Future work**

Deoxyribonucleic acid (DNA)

FUTURE WORK

- Generalize ViDa for more complicated DNA reactions (e.g. three-way strand displacement), and for other types of datasets such as RNA reactions, molecular dynamics trajectories, etc.

- Improve ViDa to enable partitioning of secondary structure microstates into clusters corresponding to different strand-level complexes.

ACKNOWLEDGEMENT

- Co-supervisors
  Anne Condon(UBC)

Khanh Dao Duc (UBC)

- Co-workers
  Jordan Lovrod (UBC)

Boyan Beronov (UBC)

- Collaborators
  Erik Winfree (Caltech)

Thank you for listening!

Q&A

THE UNIVERSITY OF BRITISH COLUMBIA

arXiv

code

https://chwzhang.com

cwzhang@cs.ubc.ca