# *In-situ* Data Analysis of Protein Folding Trajectories

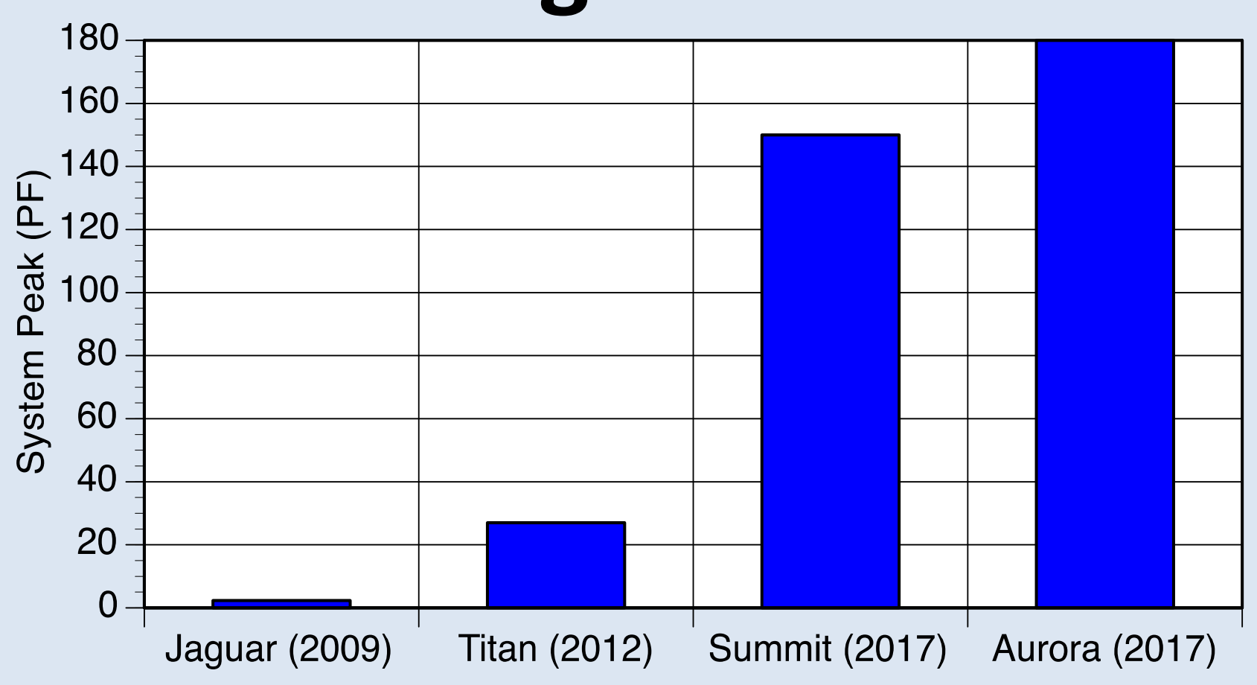
Travis Johnston[1], Boyu Zhang[1], Adam Liwo[2], Silvia Crivelli[3], Michela Taufer[1]

[1]U. Delaware Global Computing Lab, [2]U. Gdansk, Poland, [3]Lawrence Berkeley National Lab

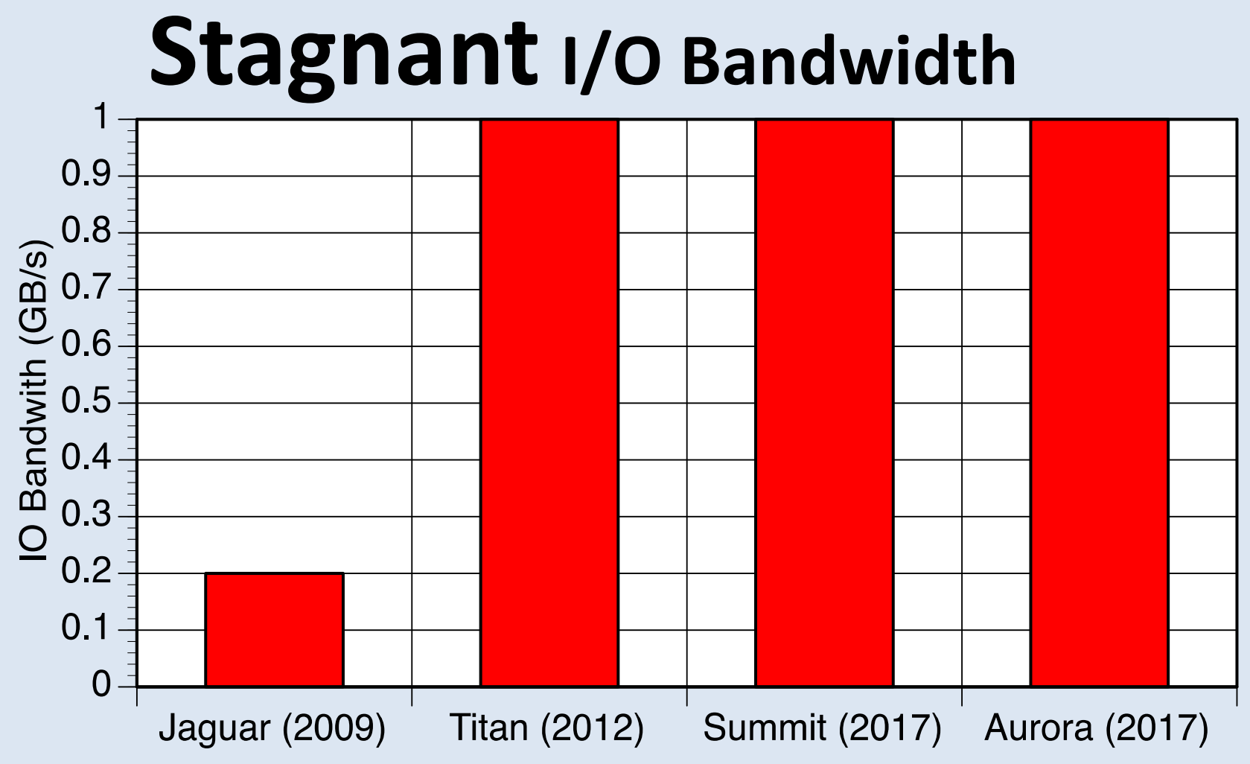
## Motivation

The transition from petascale to exascale computing will be accompanied by substantial changes in computer architectures and technologies. The research community relying on computational simulations is being forced to revisit the algorithms for data generation and analysis due to various concerns, such as higher degrees of concurrency, deeper memory hierarchies, substantial I/O and communication constraints. Simulations today typically save all data for post simulation analysis. Simulations at the exascale will require us to analyze data as it is generated and save only the results that enhance our scientific understanding. The analysis of this data will need to primarily be accomplished *in-situ*, i.e. executed sufficiently fast locally, using very limited amounts of memory and disk space, and avoiding large data movement.

**Increasing** Peak Performance



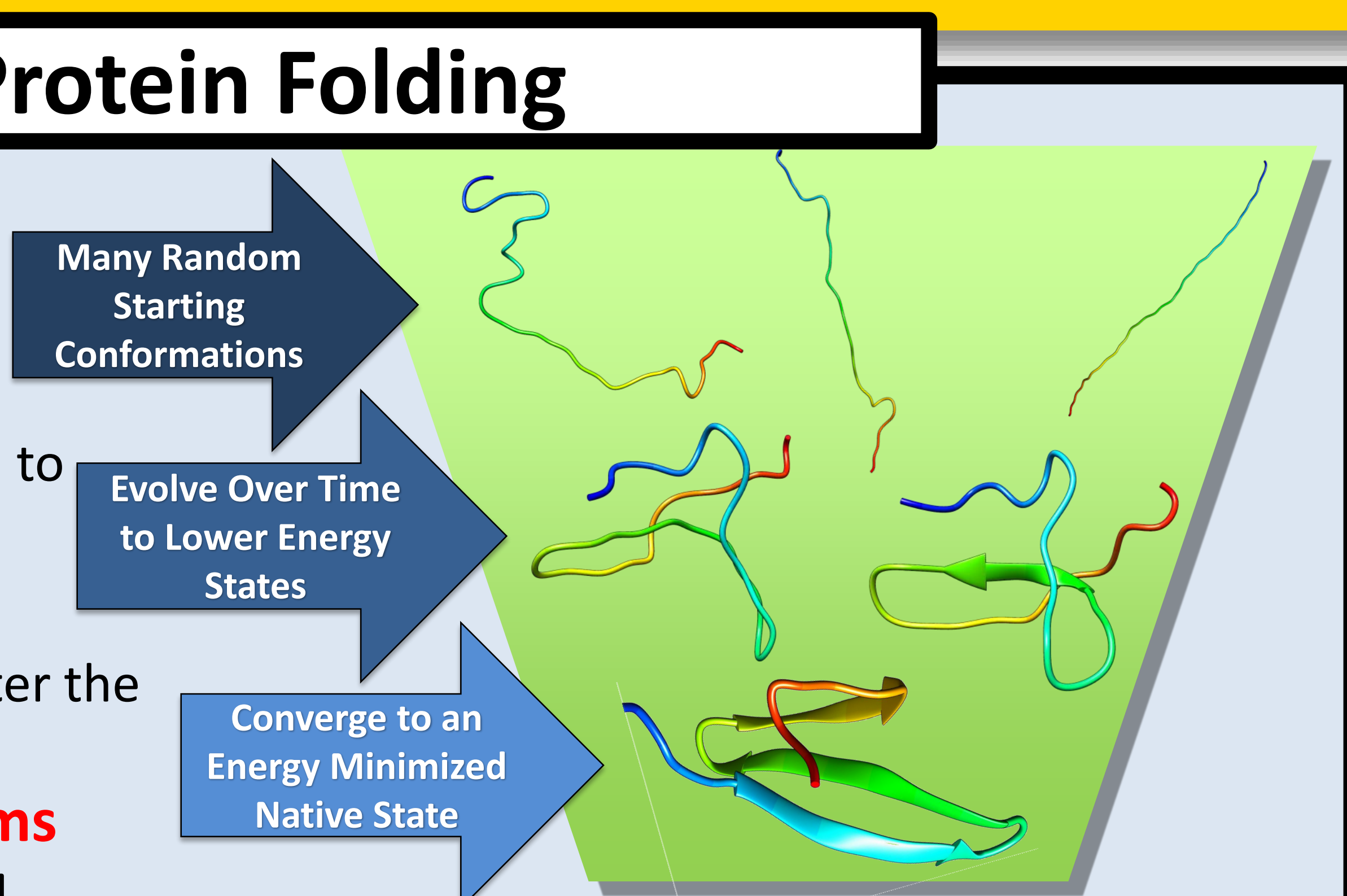
**Stagnant** I/O Bandwidth



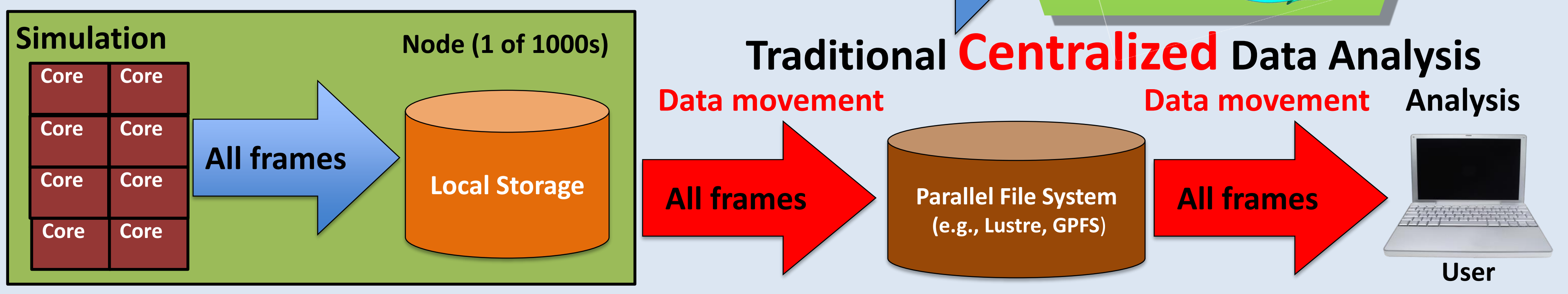
## Protein Folding

- Start from many unfolded conformations of a protein with correct chemical bonds but random torsion angles
- Run simulations on supercomputers to generate conformations (frames) of multiple folding trajectories
- Store all frame and analysis them after the simulation completes

**Does NOT scale to exascale systems**


Many Random Starting Conformations
Evolve Over Time to Lower Energy States
Converge to an Energy Minimized Native State

### Traditional Centralized Data Analysis

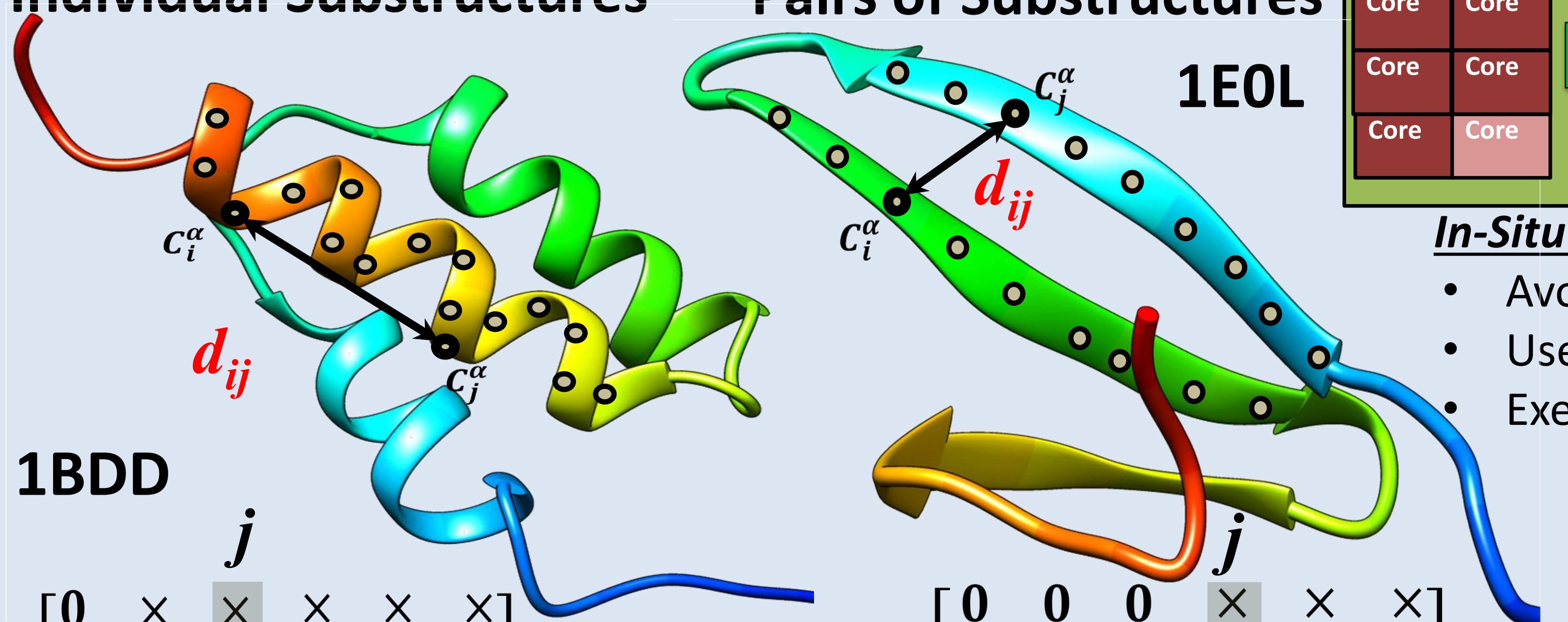


## Method for *In-situ* Data Analysis

### Algorithm

1. Discretize the protein by extracting the positions of α-Carbons and β-Carbons
2. Build a **Euclidean Distance Matrix**, D, (single structure) or a *Bipartite* distance matrix (two structures)
3. Associate the **largest eigenvalue** of *D* with the protein conformation as **metadata**
4. Use the metadata to identify stable and transition states during the simulation

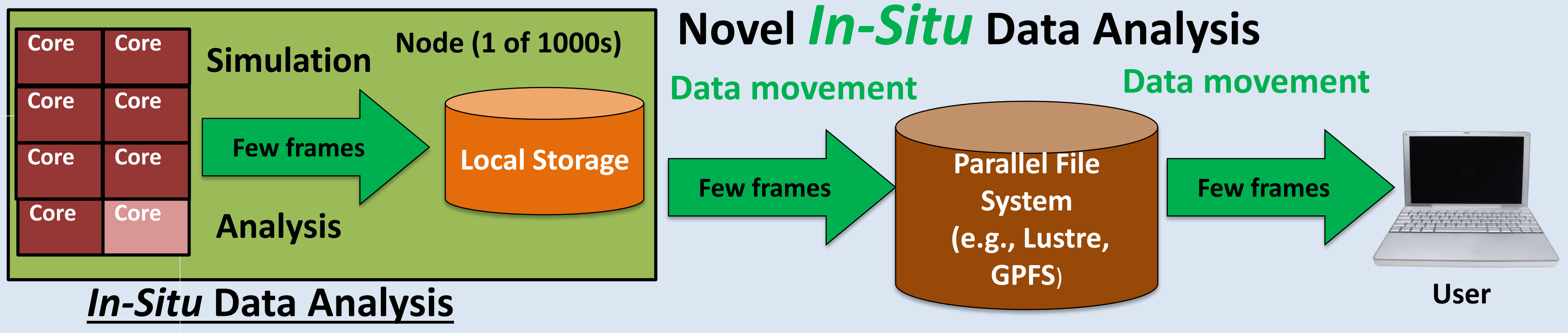
**Individual Substructures** **Pairs of Substructures**



$$D = \begin{array}{ccccccc} & & & j & & & \\ & 0 & \times & \times & \times & \times & \times \\ i & \times & 0 & d & \times & \times & \times \\ & \times & d & 0 & \times & \times & \times \\ & \times & \times & \times & 0 & \times & \times \\ & \times & \times & \times & \times & 0 & \times \\ & \times & \times & \times & \times & \times & 0 \end{array} \Rightarrow \lambda_1$$

$$D = \begin{array}{ccccccc} & & & & j & & \\ & 0 & 0 & 0 & \times & \times & \times \\ i & 0 & 0 & 0 & d & \times & \times \\ & 0 & 0 & 0 & \times & \times & \times \\ & \times & d & \times & 0 & 0 & 0 \\ & \times & \times & \times & 0 & 0 & 0 \\ & \times & \times & \times & 0 & 0 & 0 \end{array} \Rightarrow \lambda_1$$

### Novel *In-Situ* Data Analysis



***In-Situ* Data Analysis**

- Avoid data movement among nodes
- Use a small amount of memory
- Execute sufficiently fast

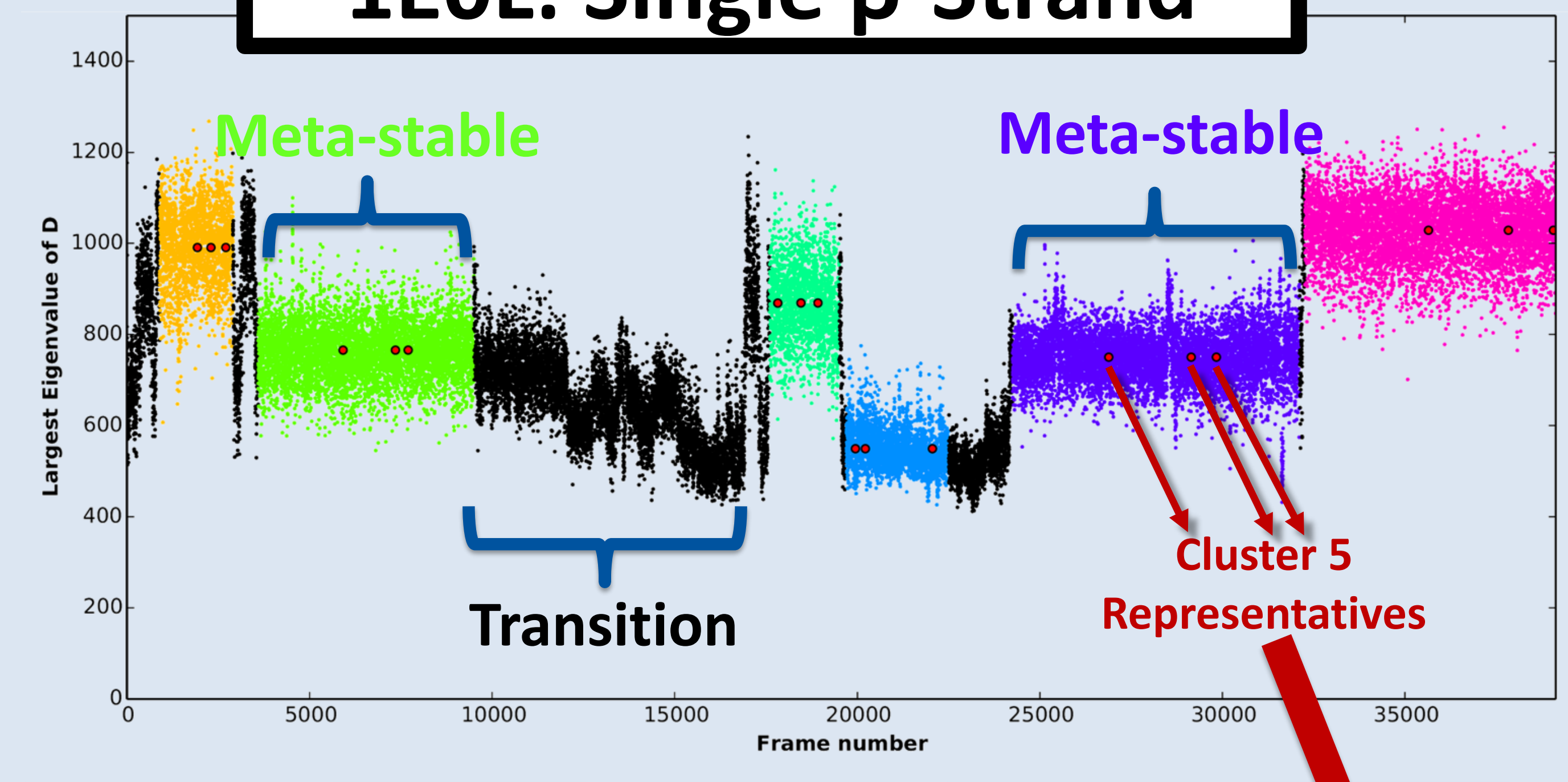
**Does scale to exascale systems**

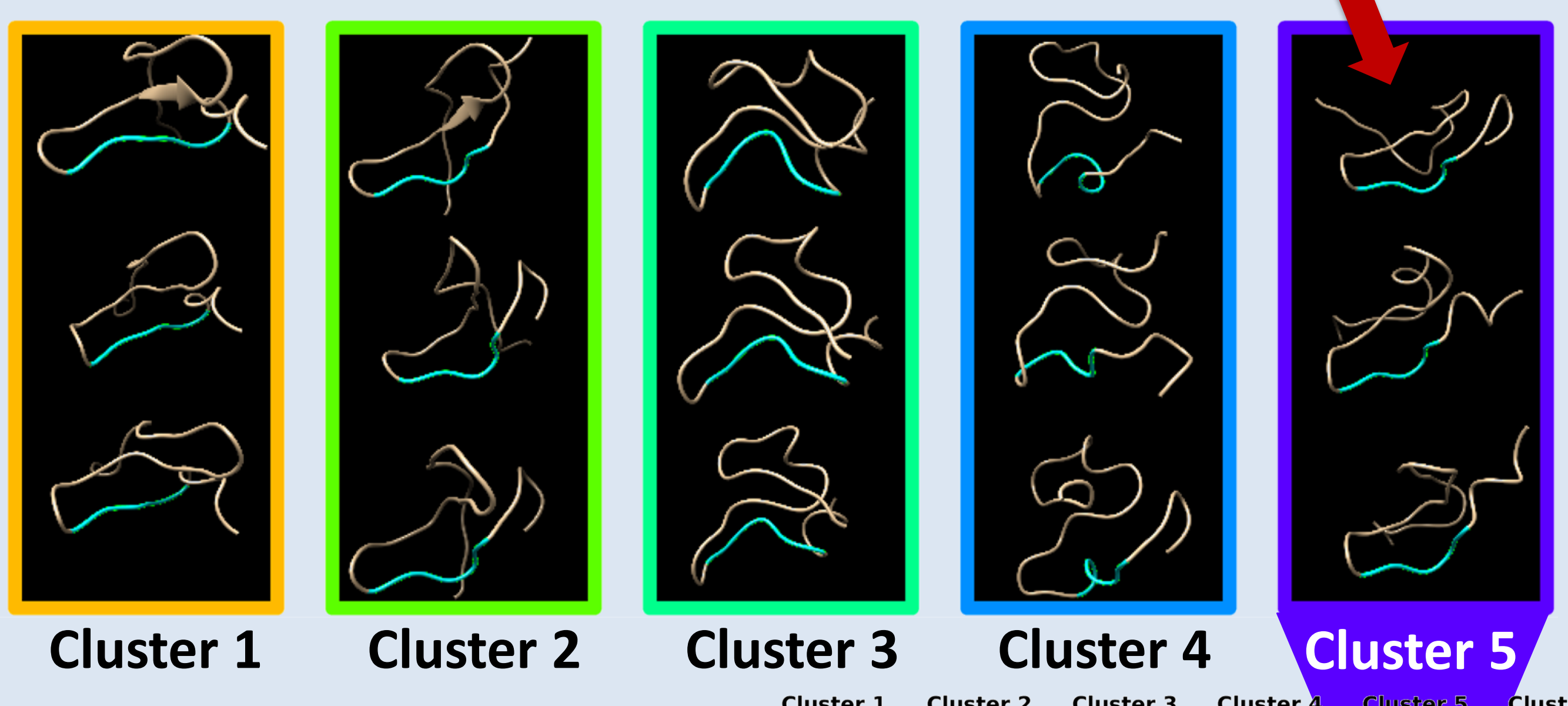
## Our Contribution

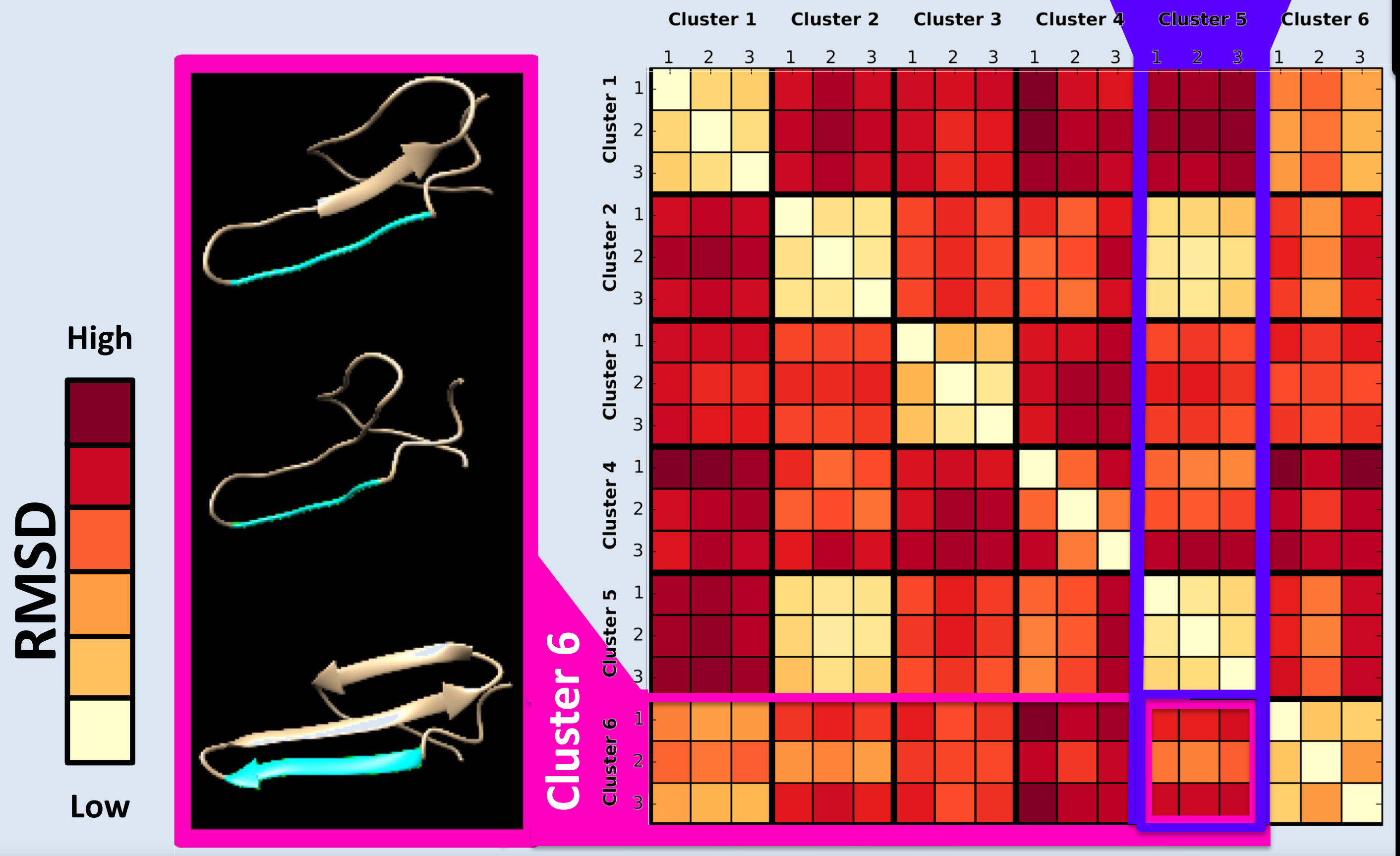
To satisfy I/O bandwidth constraints at exascale, we propose a method for ***in-situ*** data analysis:

- Accurately captures features of protein trajectory
- Performs light computation and does not interfere with ongoing folding simulation
- Significantly reduces the amount of data written to disk (from about 40k frames to a few 10s of frames)

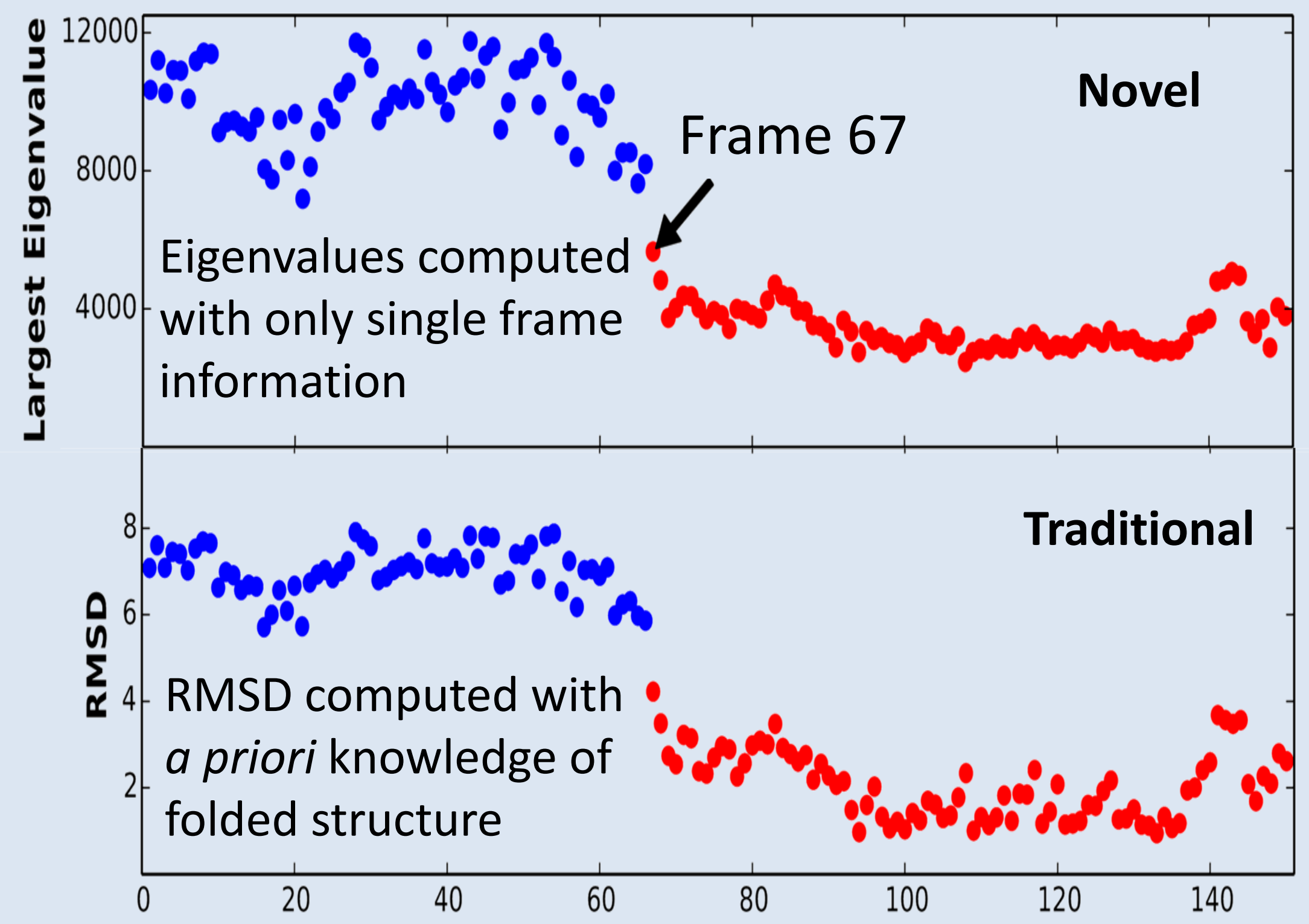
## 1E0L: Single β-Strand

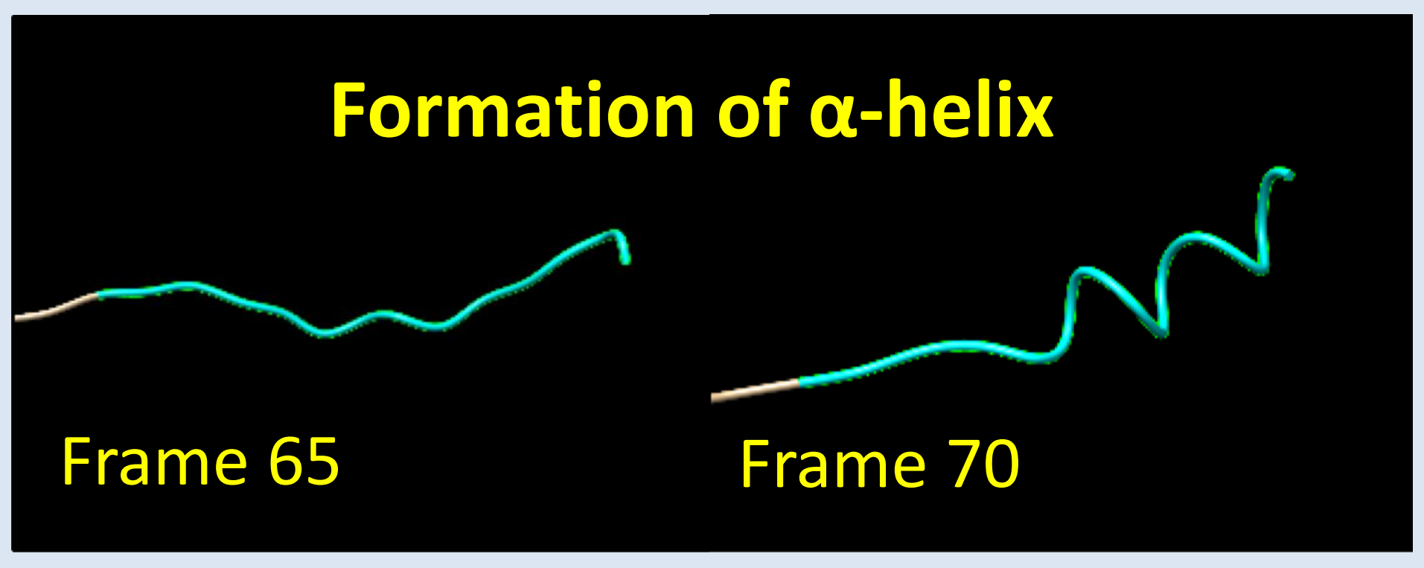






## 1BDD: Single α-Helix

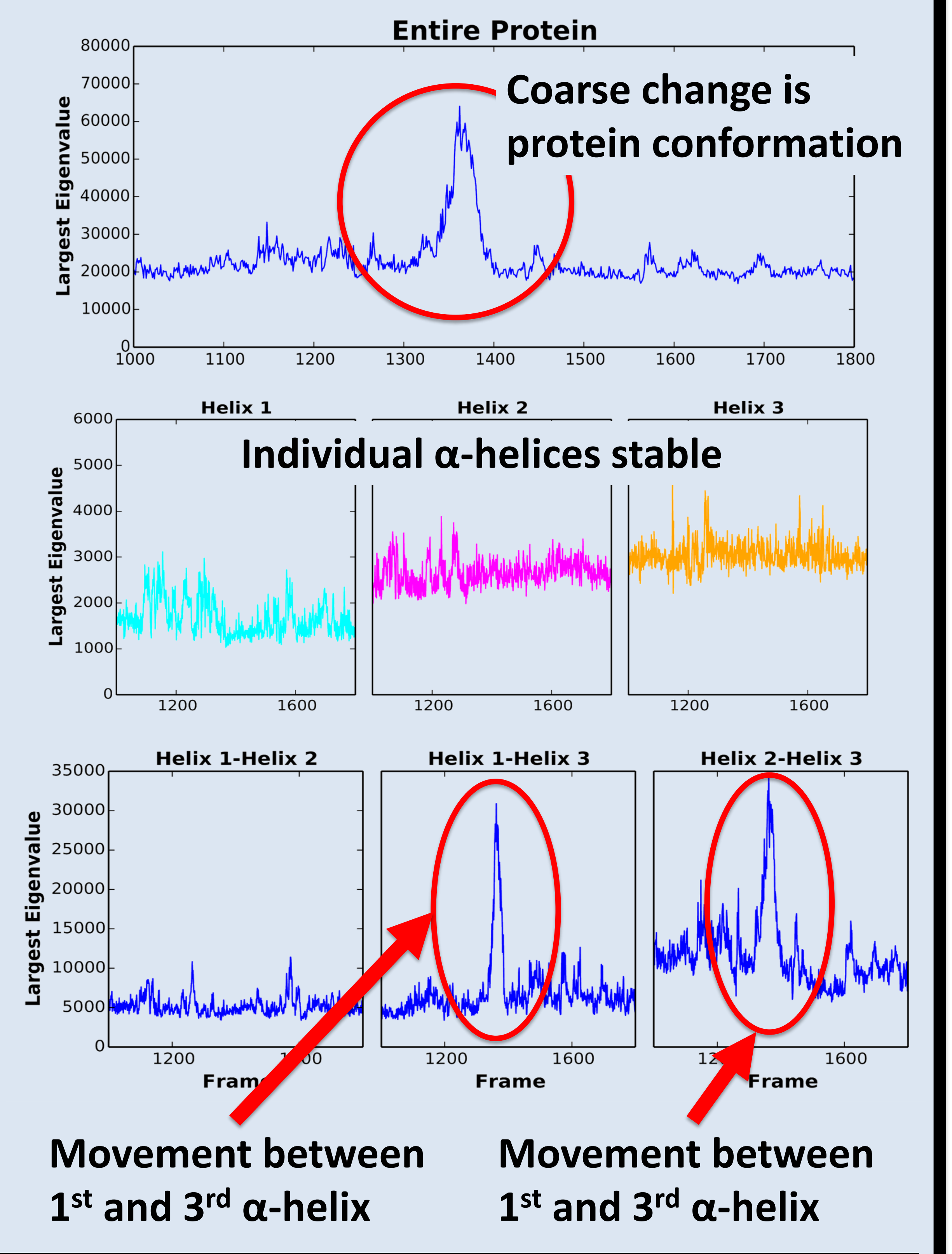


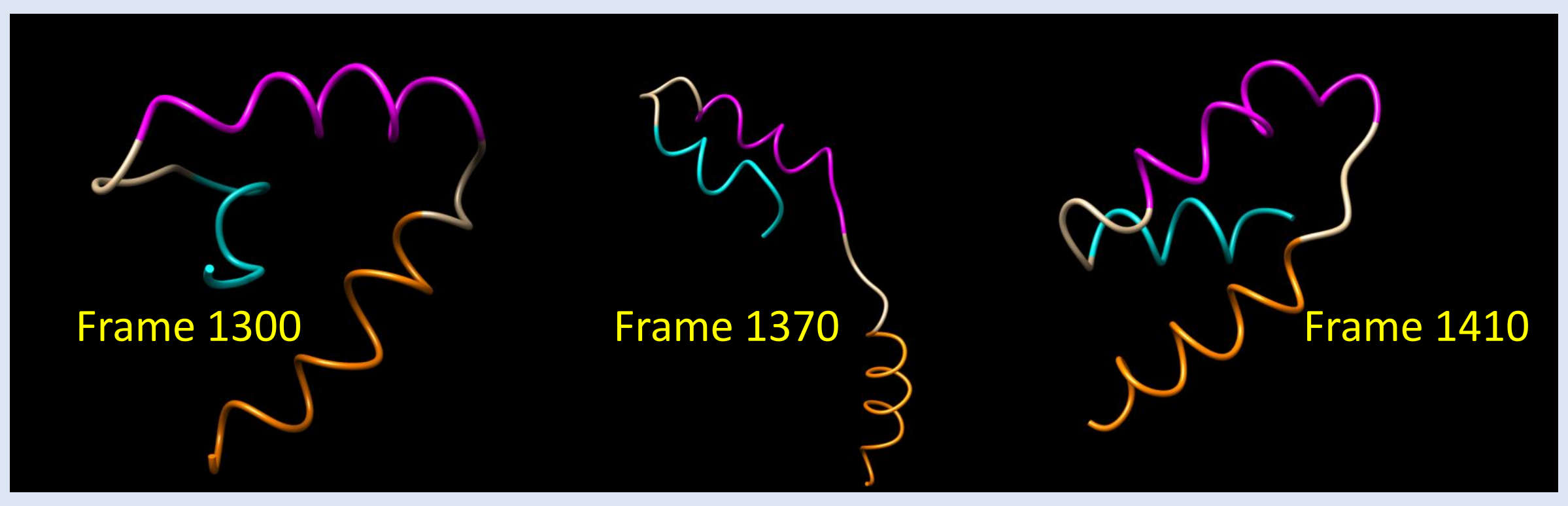


## Conclusions

We propose a novel method for *in-situ* data analysis of protein folding trajectories. We validate our metadata mapping method by applying it to two 40k frame trajectories: one trajectory of 1BDD (containing 3 α-helices), and one trajectory of 1E0L (containing 3 β-strands). Our metadata mapping enabled us to observe metastable states, transition states, the formation of an individual substructure and the repositioning of substructure relative to others.

## 1BDD: Pair of α-Helices





Acknowledgements: NSF Grant #1318445