Algorithms for In Situ Data Analytics of Next Generation Molecular Dynamics Workflows

Michela Taufer

Department of Electrical Engineering and Computer Science

The University of Tennessee Knoxville

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

Acknowledgements

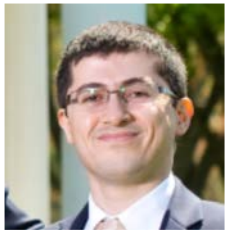

T. Johnston


B. Zhang


T. Estrada


A Liwo


T. Do


A. Razavi


S. Crivelli


R. da Silva


S. Thomas


B. Mulligan

A. Plante


H. Weinstein


E. Deelman

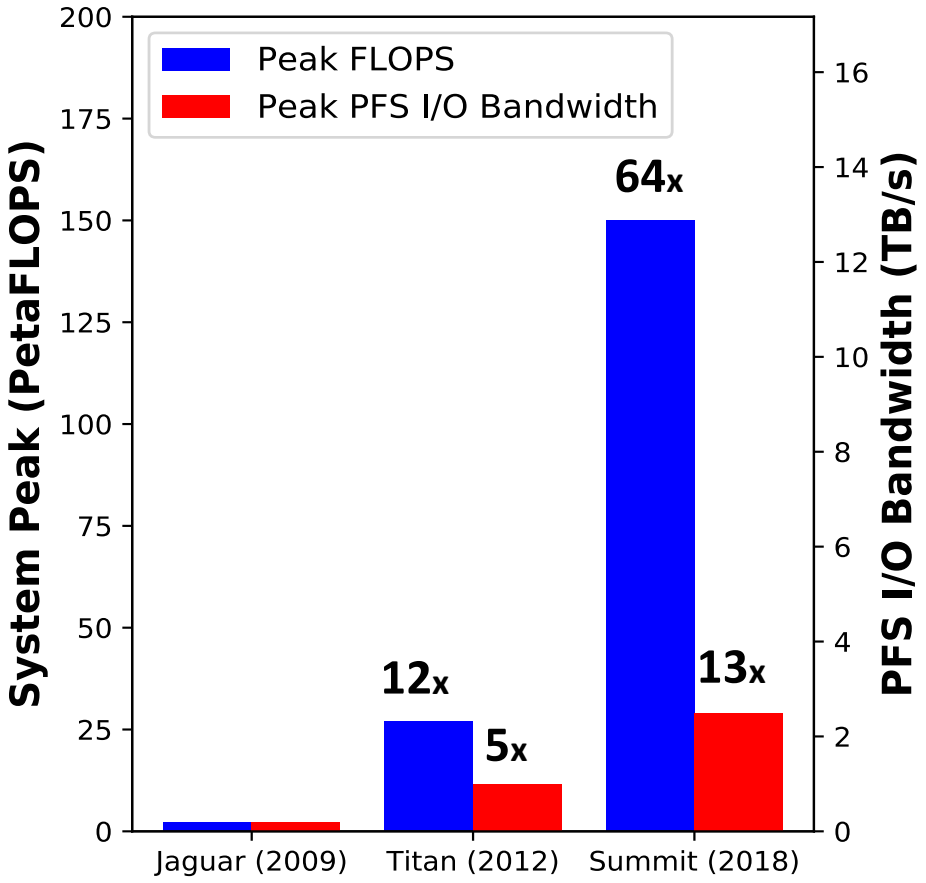

M. Cuendet

Sponsors:

BIG ORANGE BIG IDEAS

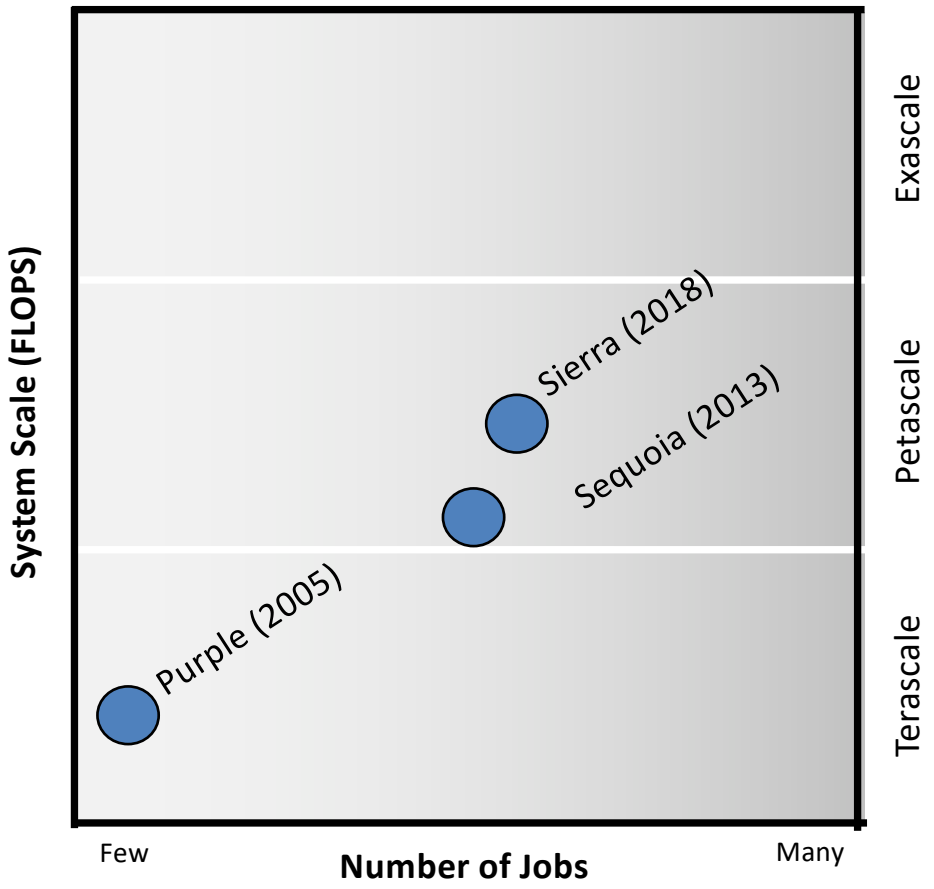
Trends in Next-Generation Systems

Widening I/O Gap



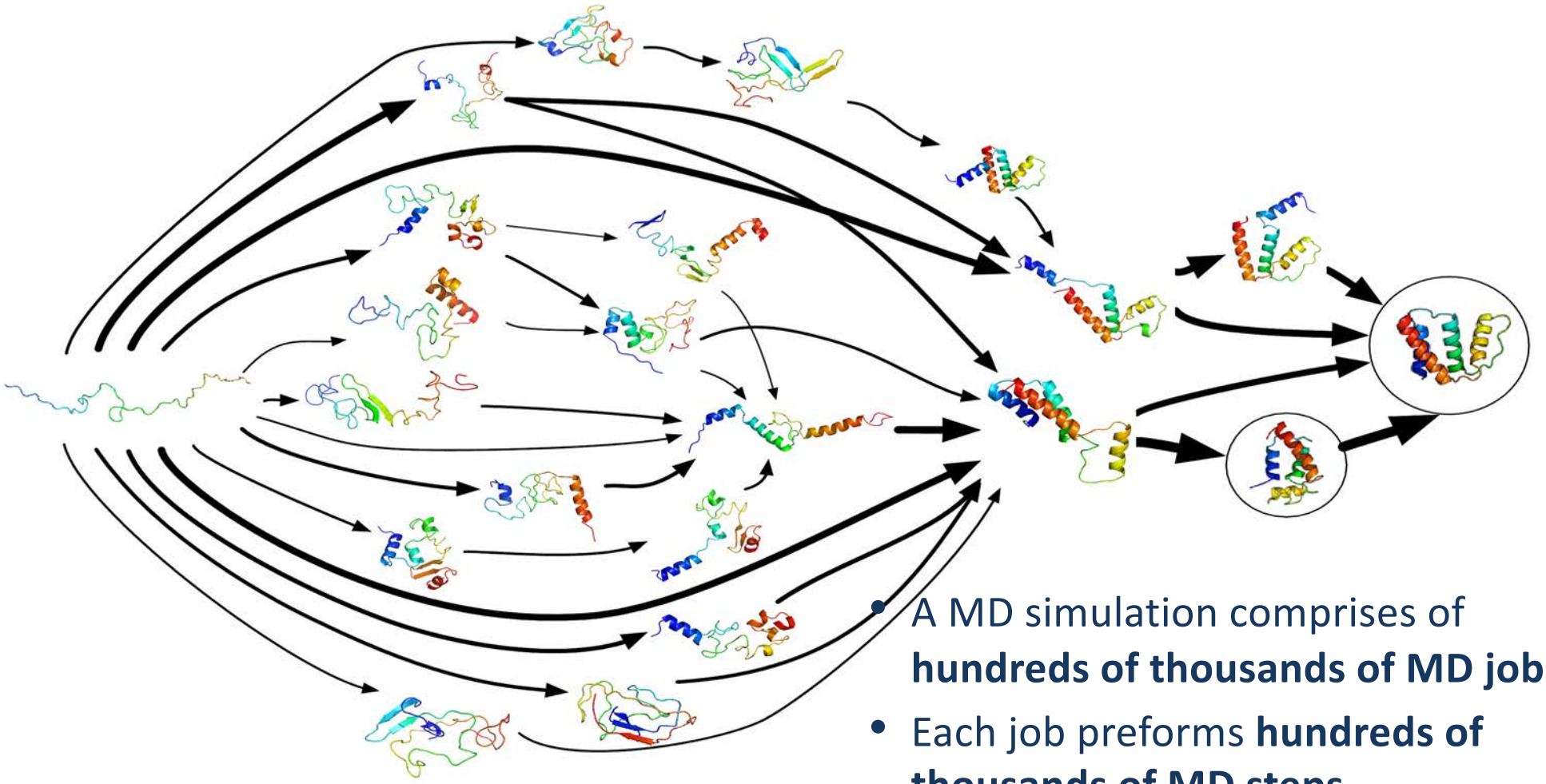
Source: Lucy Nowell (DOE)

Rising Importance of Ensembles



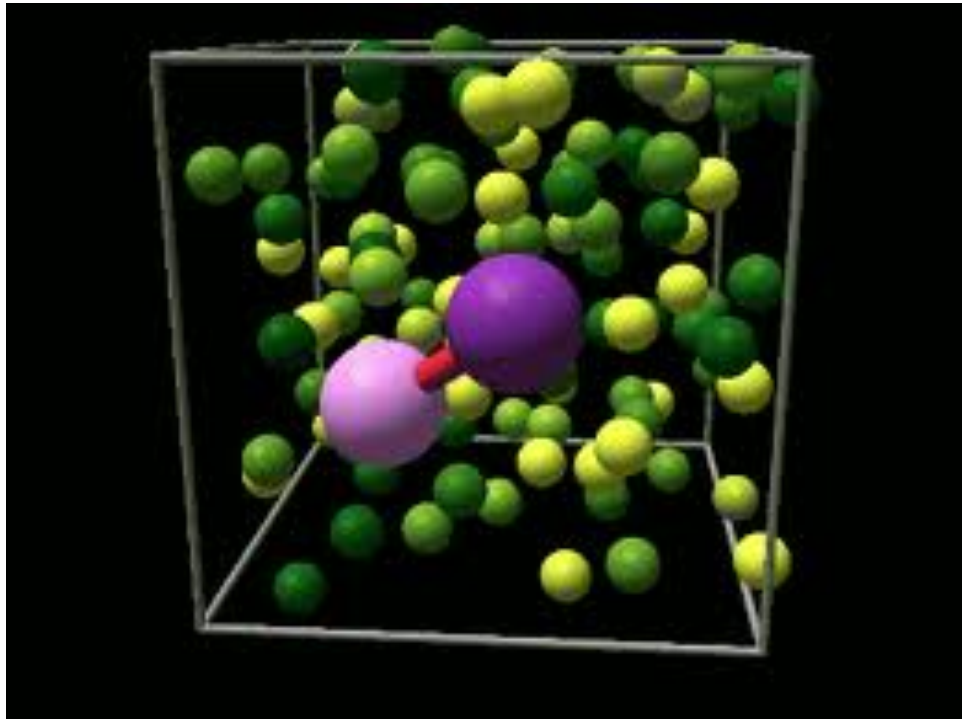
Source: https://wci.llnl.gov/simulation/computer-codes/uncertainty-quantification

Classical Molecular Dynamics Simulations



- A MD simulation comprises of **hundreds of thousands of MD job**
- Each job preforms **hundreds of thousands of MD steps**

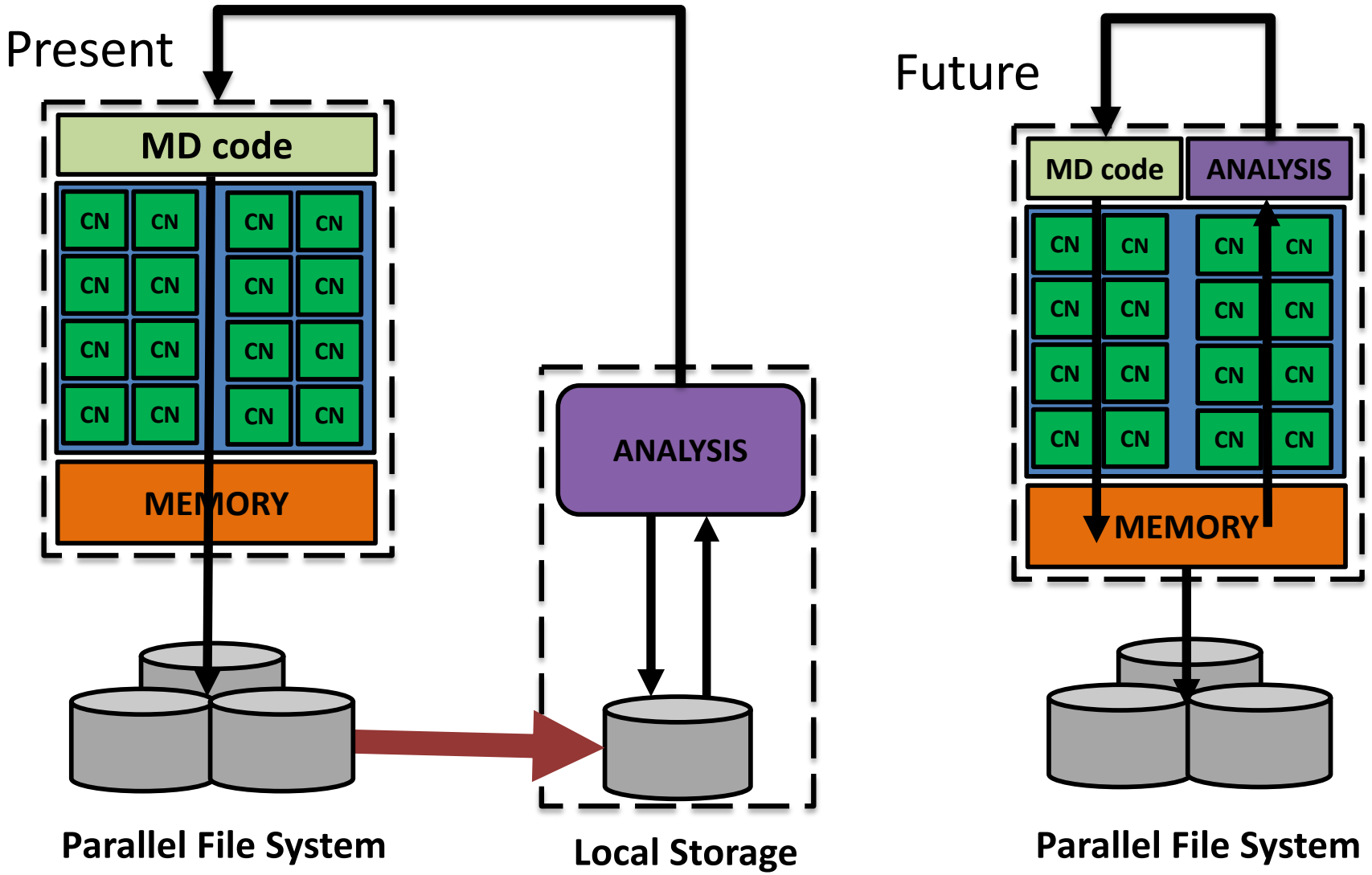
Classical Molecular Dynamics Simulations



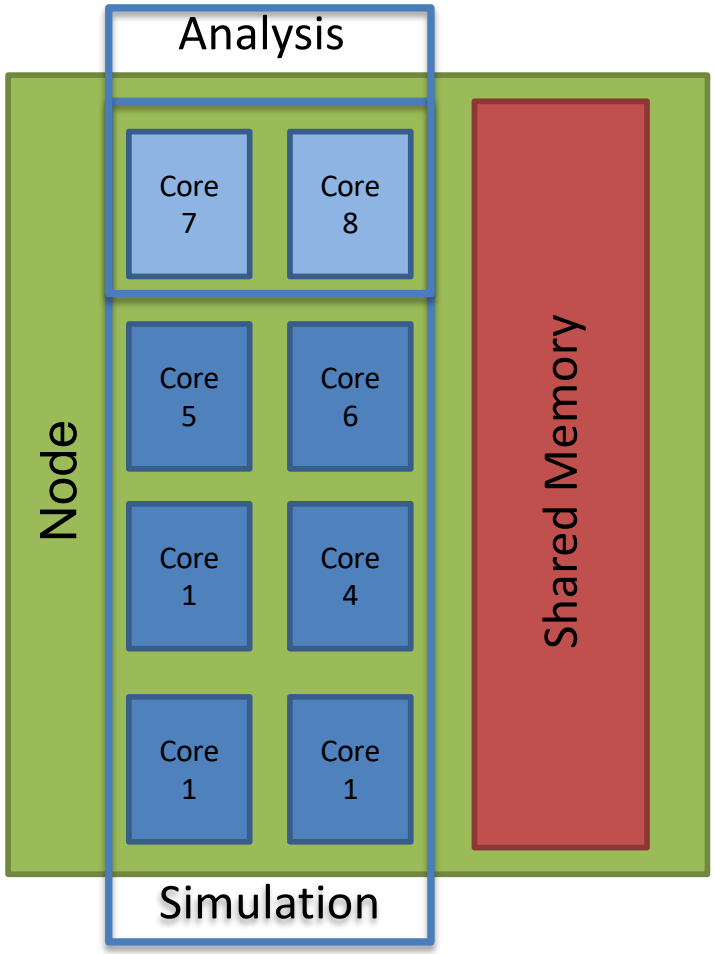
Forces on single atoms
→ Acceleration
→ Velocity
→ Position

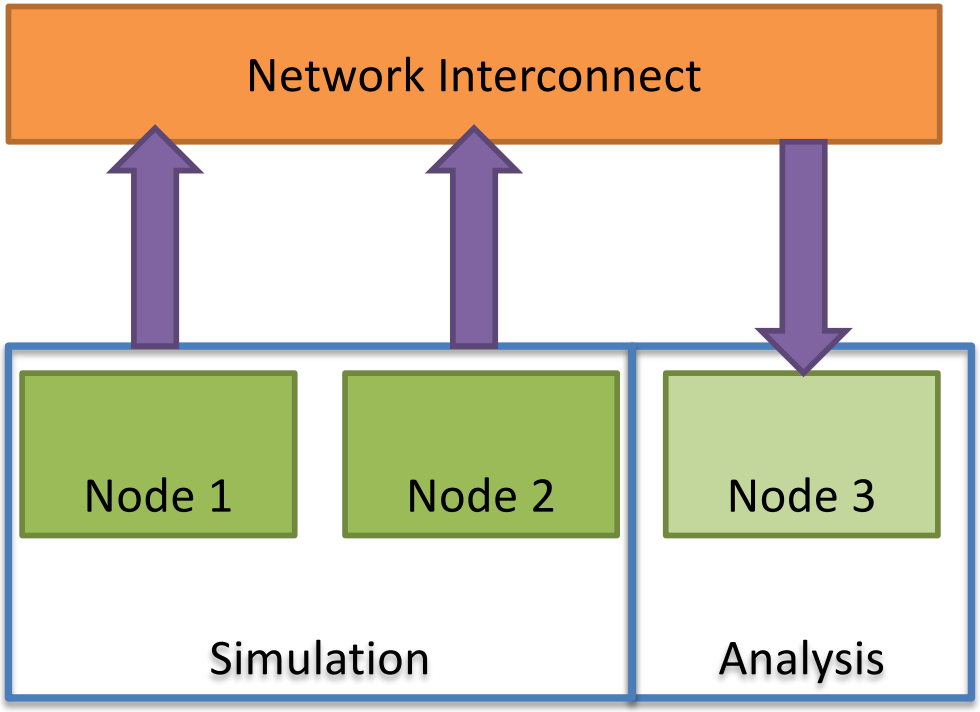
- A MD step computes forces on single atoms (e.g., bond, angle, dihedrals, nonbond)
- Forces are added to compute acceleration
- Acceleration is used to update velocities
- Velocities are used to update the atom positions
- Every *n* steps, all atom positions are stored
  → **3D snapshot or frame**

Analyzing MD Frames: Present and Future



In Situ and *In Transit* Analysis



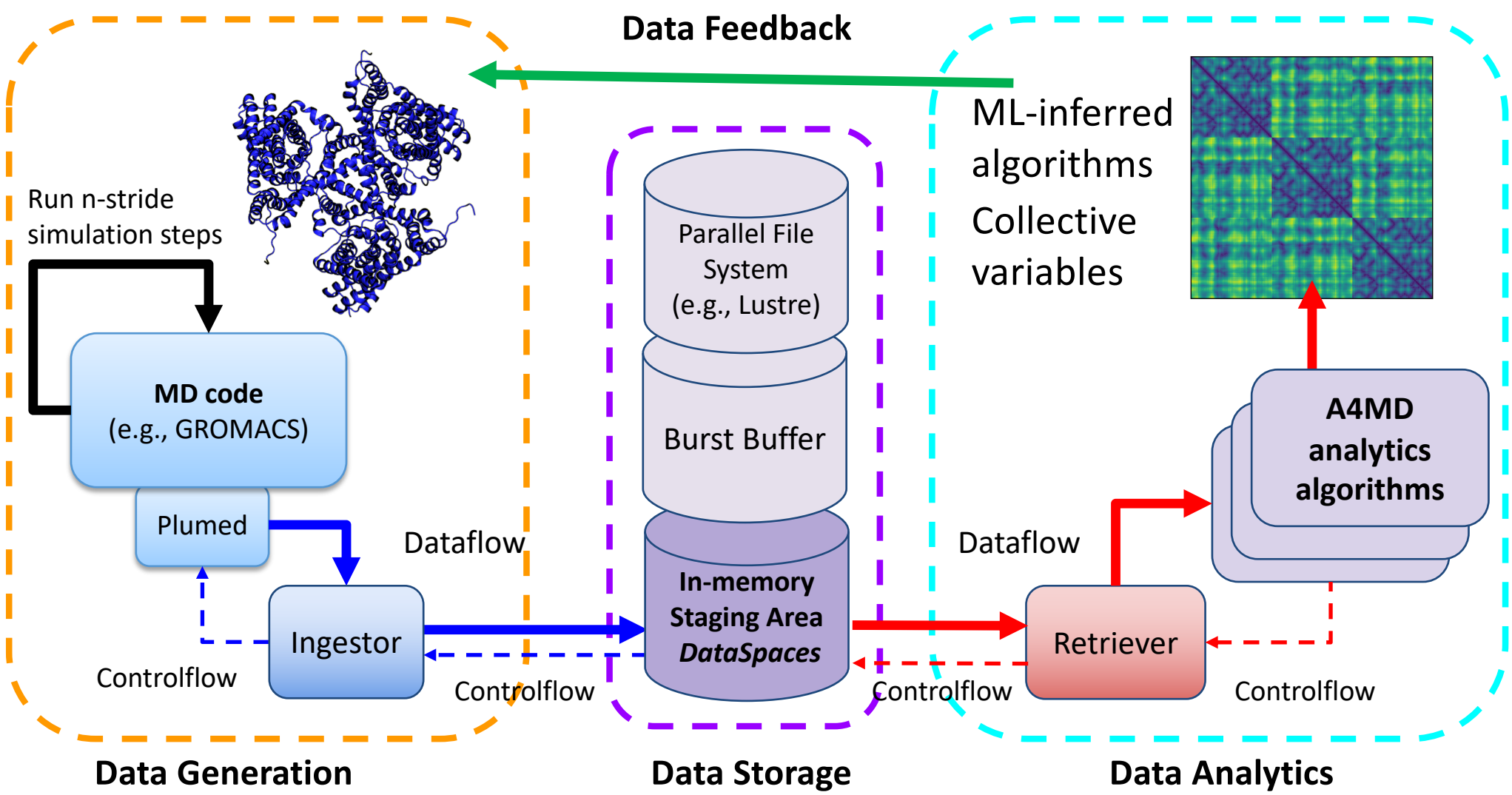


Example of tools:

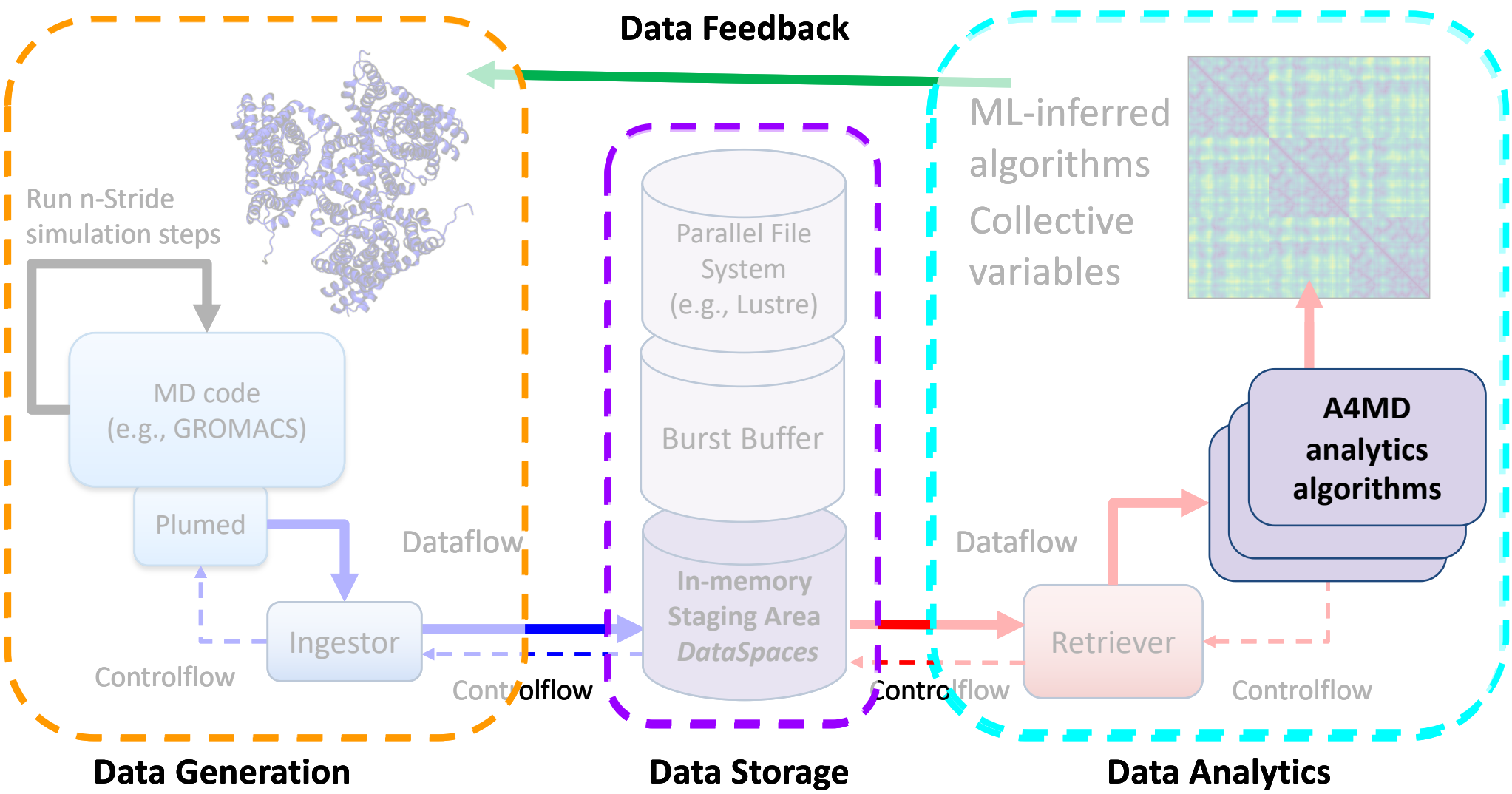
- DataSpaces (Rutgers U.)
- DataStager (GeorgiaTech)

***In situ* and *in transit* analysis requires rethinking data algorithms**

Building a Closed-loop Workflow



Building a Closed-loop Workflow



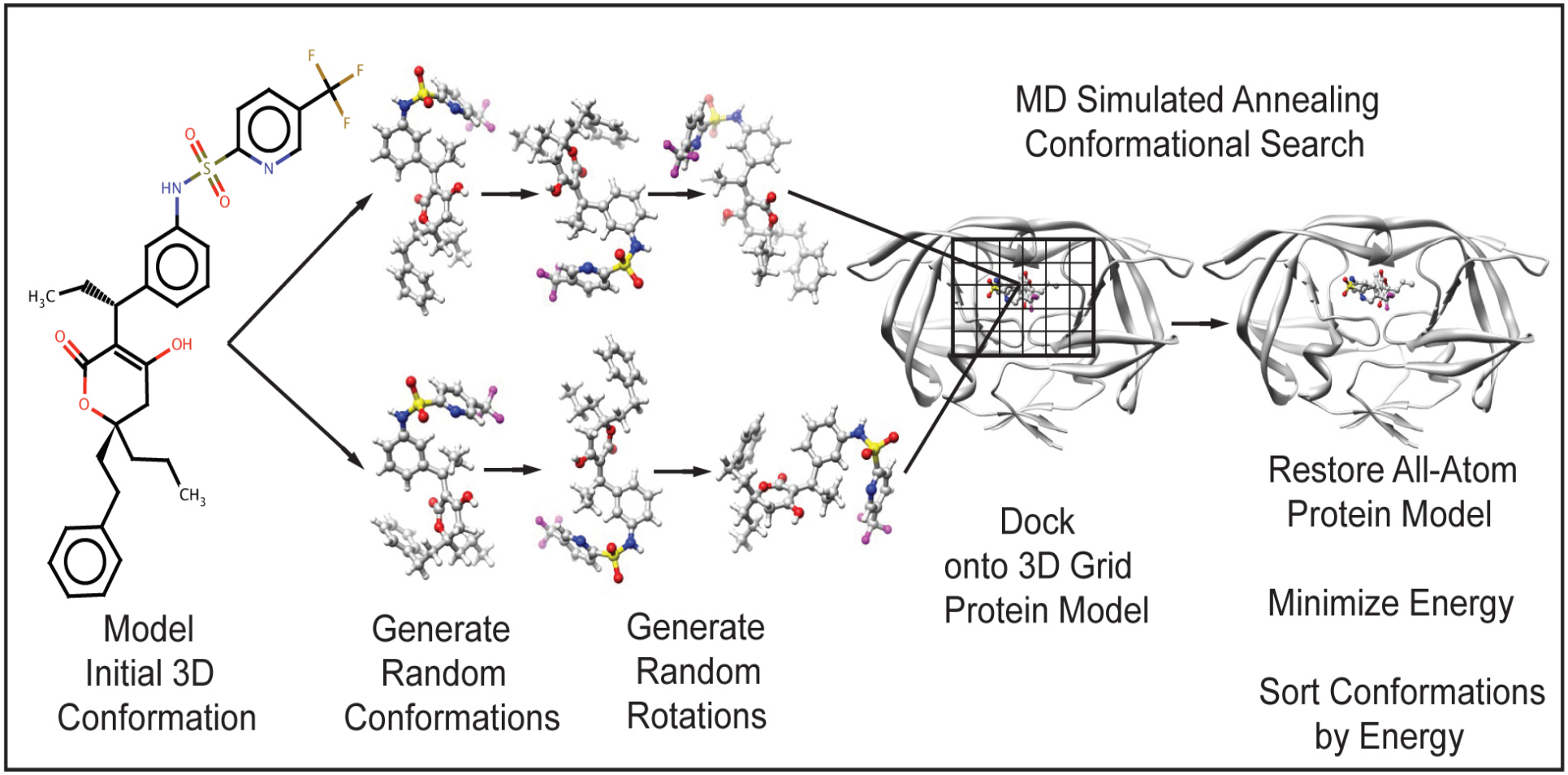
Analytics for Molecular Dynamics

- Drug design and protein-ligand docking
- Protein folding and rare events
- Protein variants expressed from yeast or bacteria and protein engineering

Analytics for Molecular Dynamics

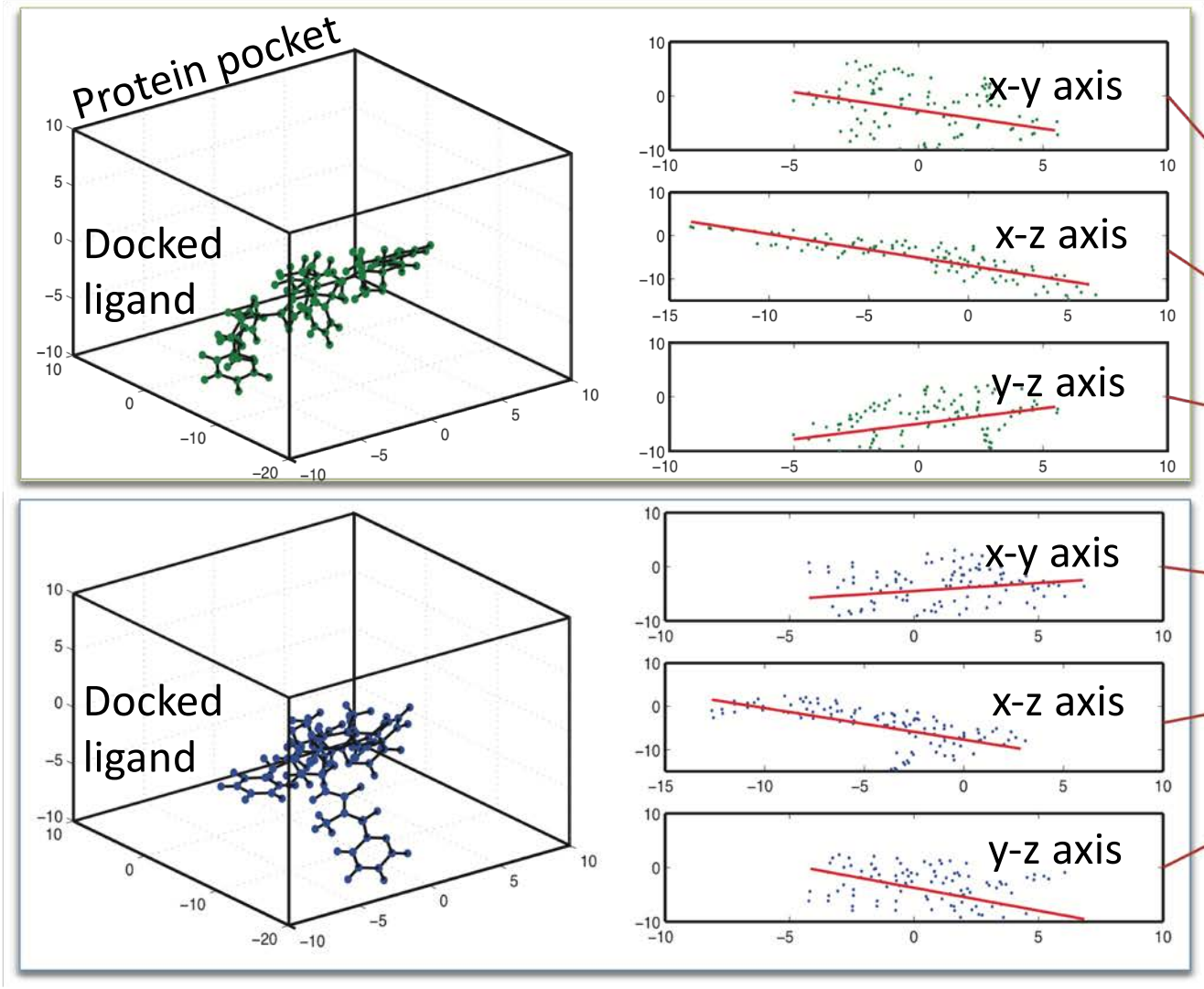
- **Drug design and protein-ligand docking**
- Protein folding and rare events
- Protein variants expressed from yeast or bacteria and protein engineering

A4MD: Protein-Ligand Docking

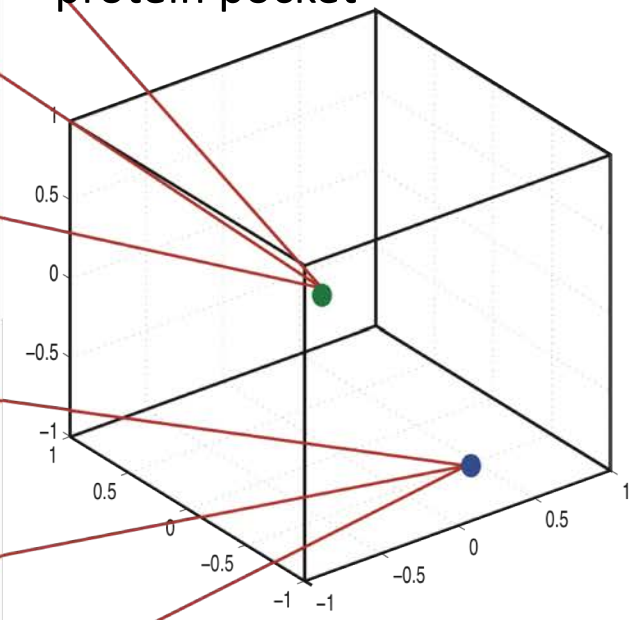


T. Estrada, B. Zhang, P. Cicotti, R. S. Armen, M. Taufer: ***A scalable and accurate method for classifying protein-ligand binding geometries using a MapReduce approach.*** Comp. in Bio. and Med. 42(7): 758-771 (2012)

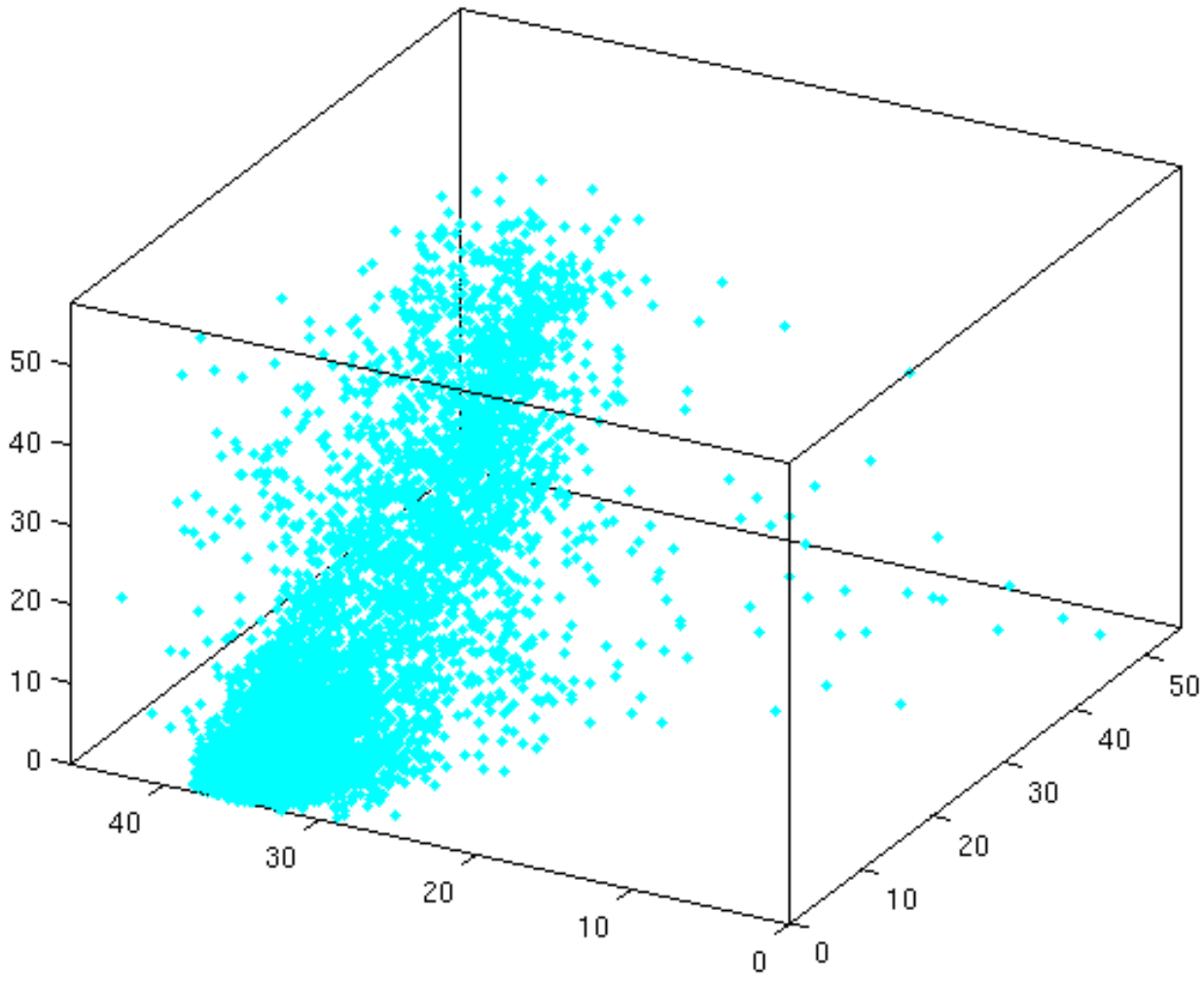
From 3D Atomic Structures to 3D Points



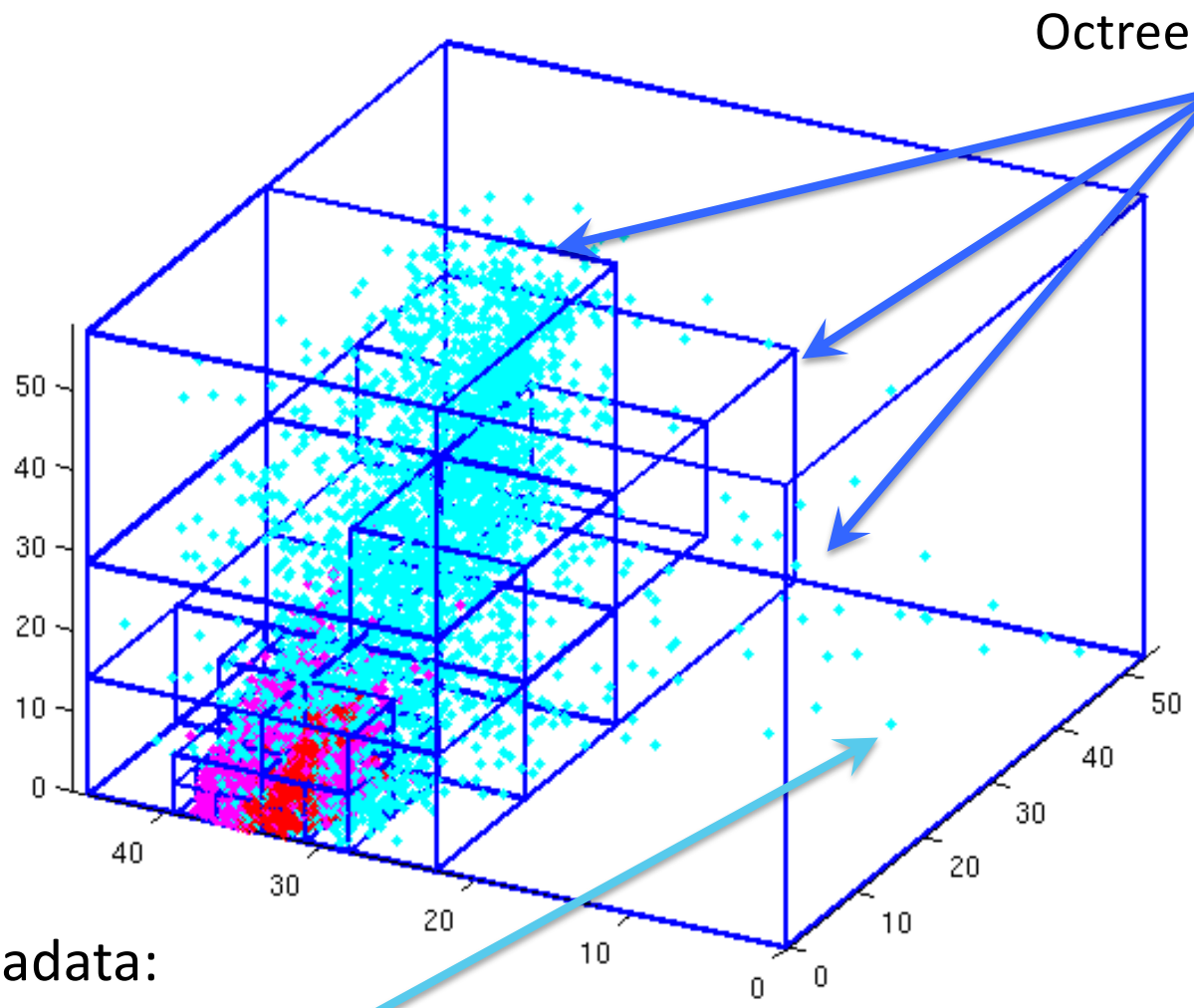
Metadata: each 3D point represents the position of one docked ligand in the protein pocket



Search for Dense Spaces: Octree Clustering

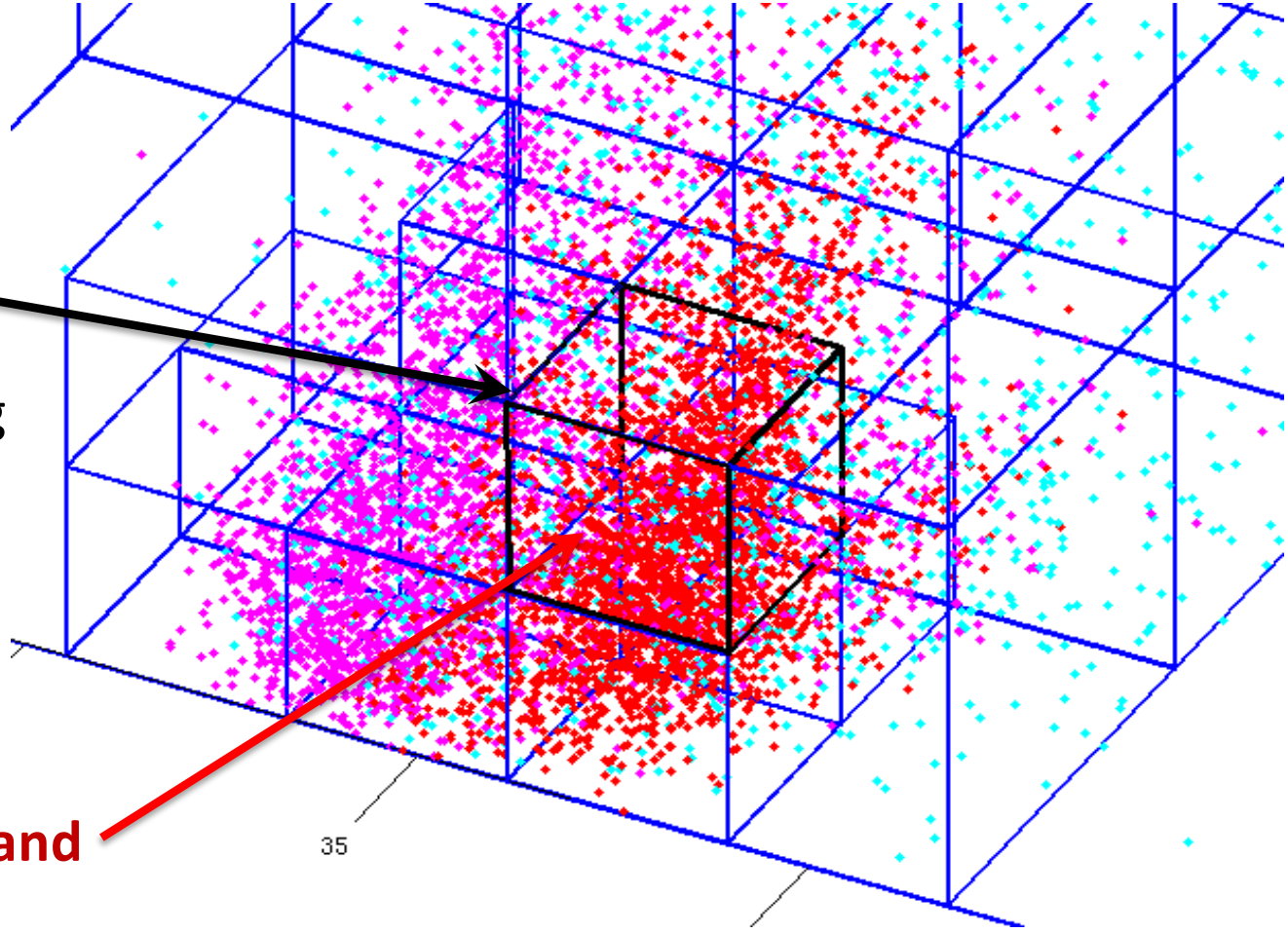


Search for Dense Spaces: Octree Clustering



Search for Dense Spaces: Octree Clustering

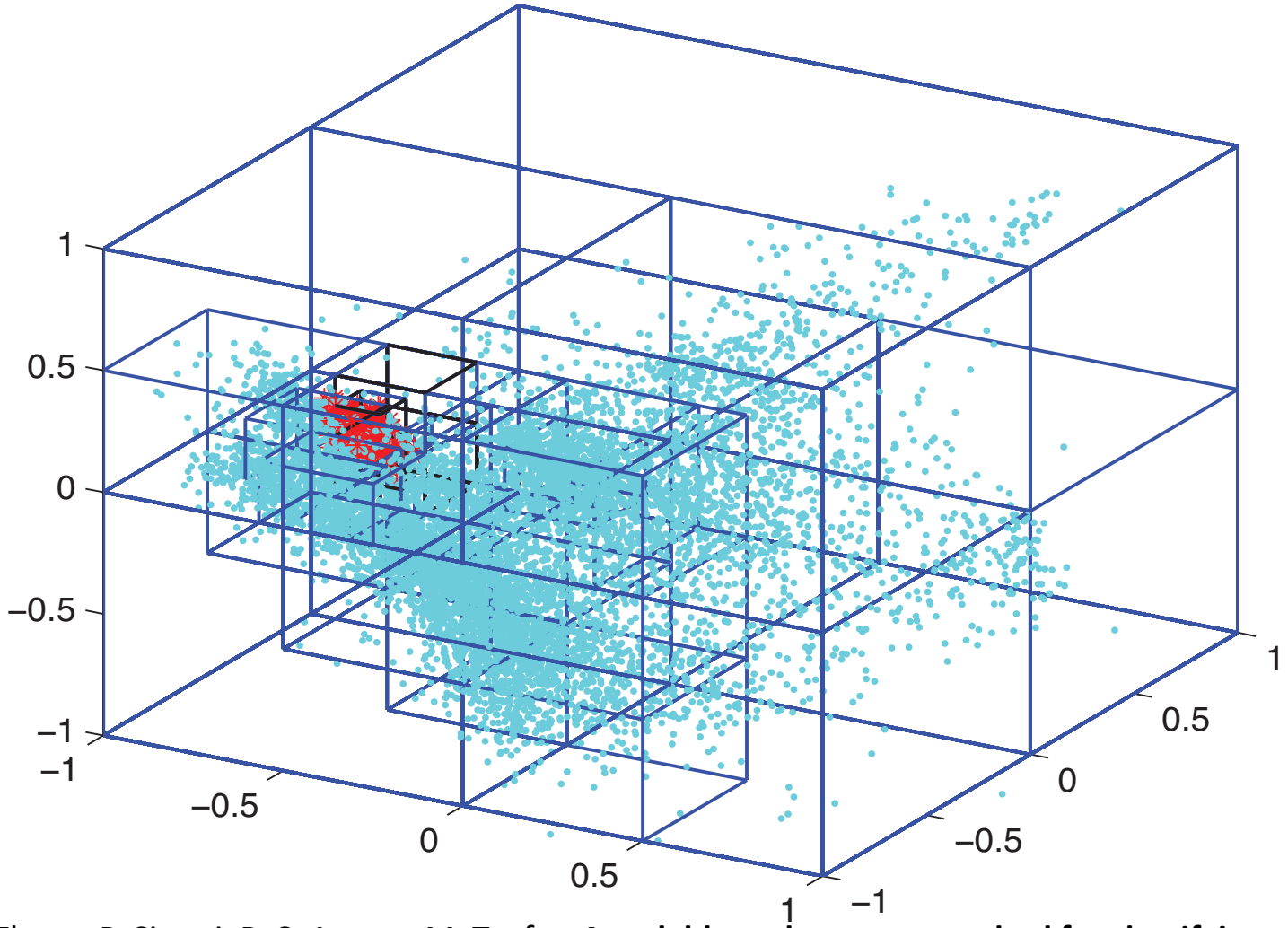
Deepest, more dense octant found by octree clustering



Near-native ligand structures (RMSD <= 2A)

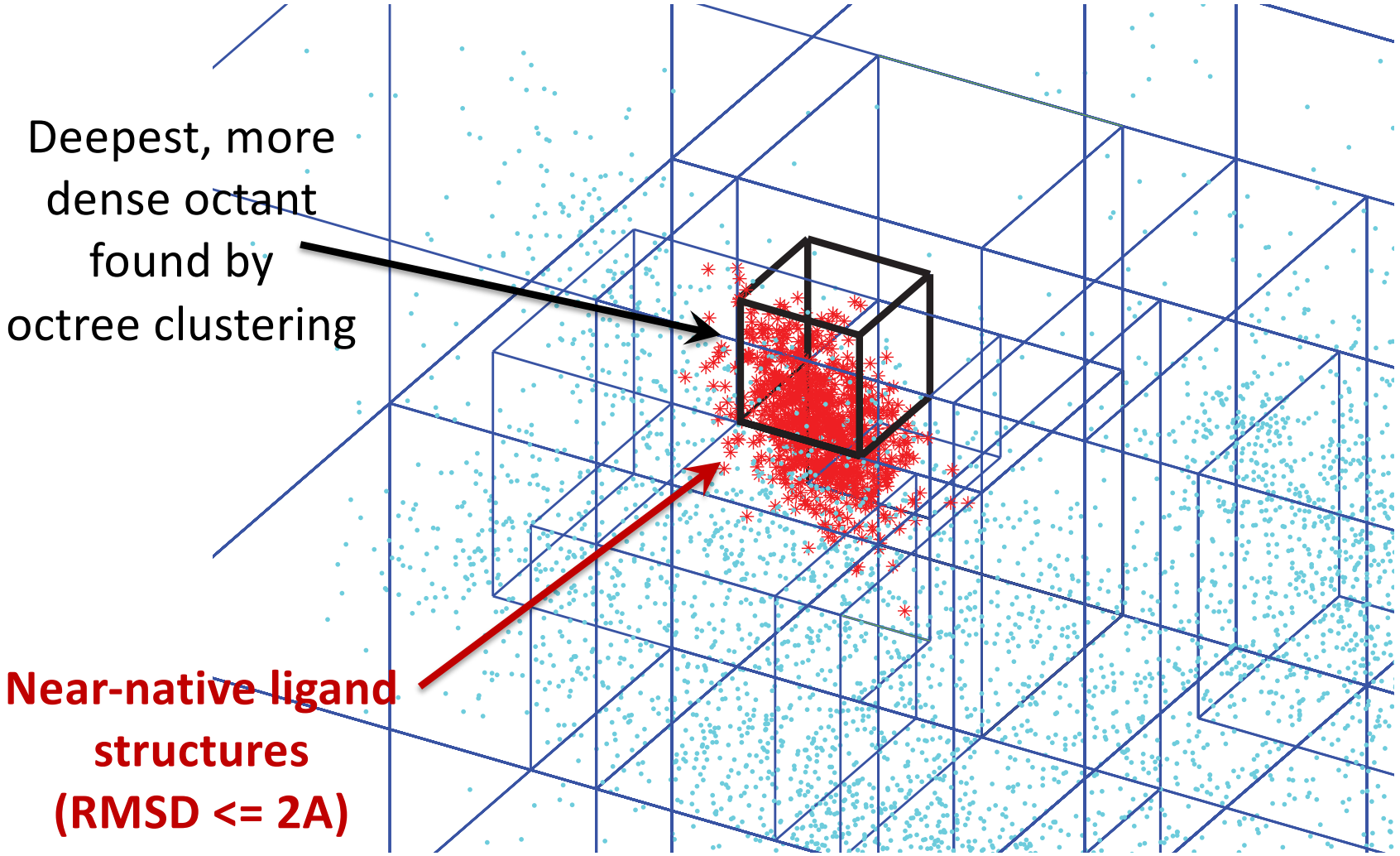
Search: Linear in complexity using Mimir - a MapReduce over MPI framework

Case Study: Sampled Conformations - Ligand 1k1l



T. Estrada, B. Zhang, P. Cicotti, R. S. Armen, M. Taufer: **A scalable and accurate method for classifying protein-ligand binding geometries using a MapReduce approach.** Comp. in Bio. and Med. 42(7): 758-771 (2012)

Case Study: Sampled Conformations - Ligand 1k1l

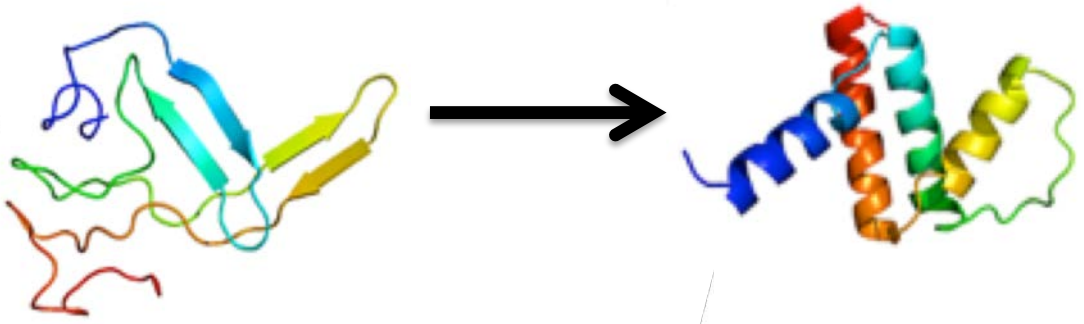


Analytics for Molecular Dynamics

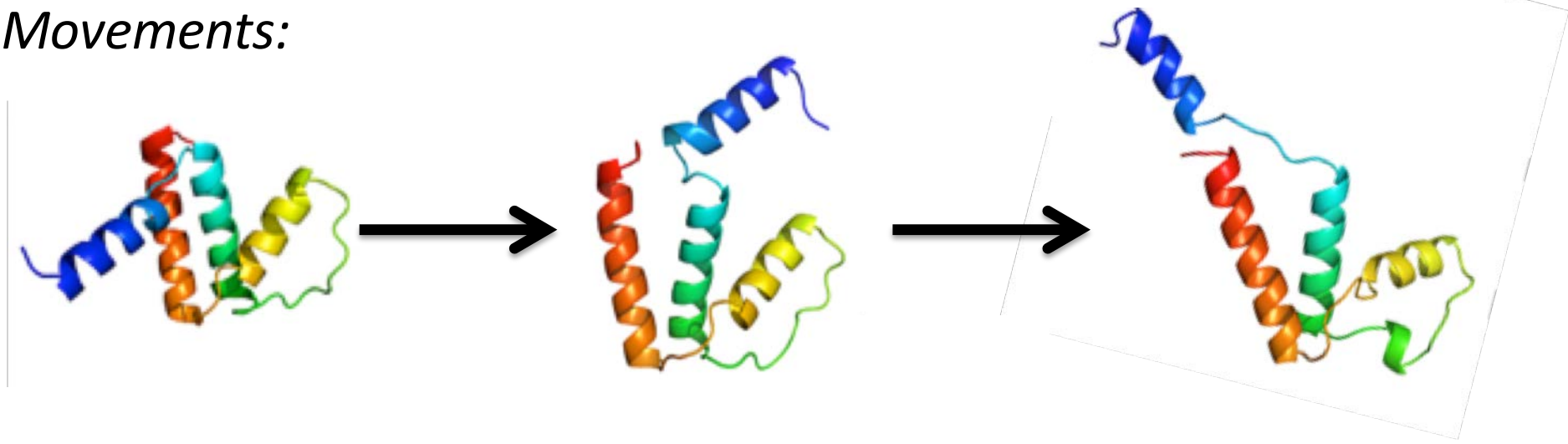
- Drug design and protein-ligand docking
- **Protein folding and rare events**
- Protein variants expressed from yeast or bacteria and protein engineering

A4MD: Rare Events in MD Simulations

Transformations:

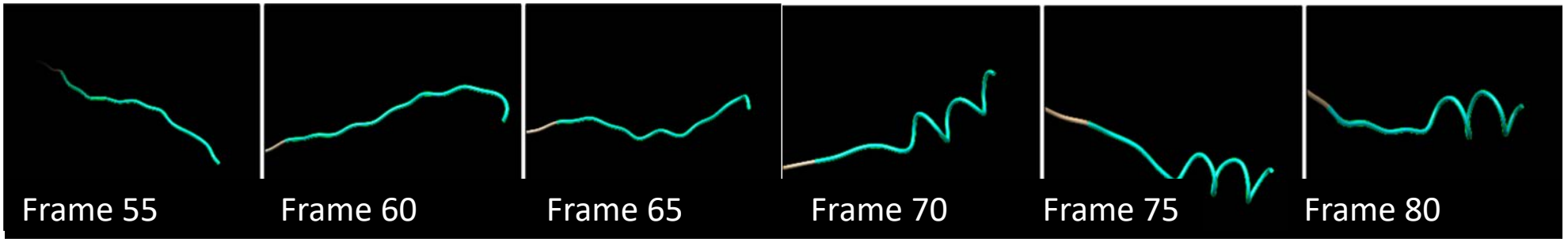


Movements:



A4MD: Rare Events in MD Simulations

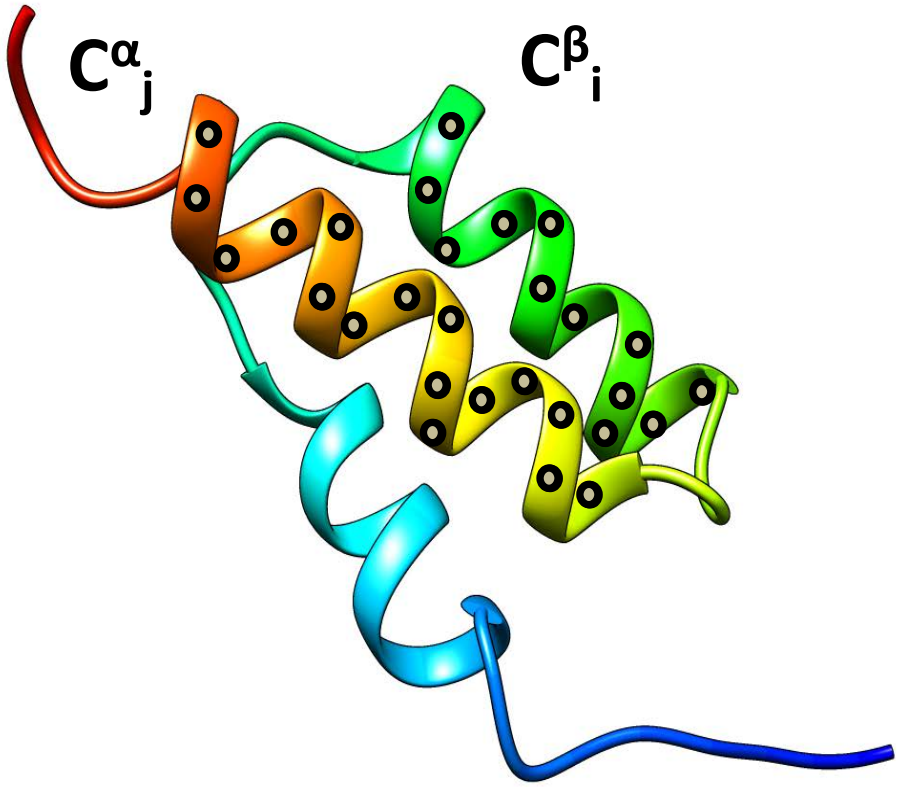
Frames (or snapshots) of an MD trajectory:



- We want to capture what is going on in each frame **without**:
  - Disrupting the simulation (e.g., stealing CPU and memory on the node)
  - Moving all the frames to a central file system and analyzing them once the simulation is over
  - Comparing each frame with past frames of the same job
  - Comparing each frame with frames of other jobs

From 3D Atomic Structure to a Single Eigenvalue

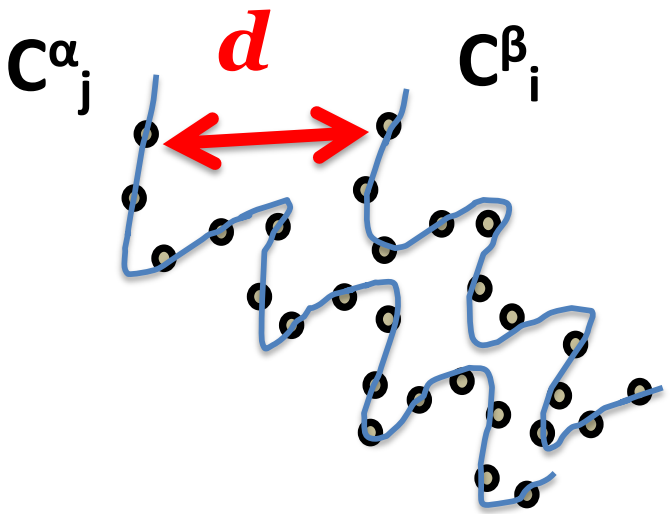
Drop all but not the backbone atoms (C^{α} atoms)



T. Johnston, B. Zhang, A. Liwo, S. Crivelli, and M. Taufer. In-Situ Data Analytics and Indexing of Protein Trajectories. *Journal of Computational Chemistry (JCC),* 38(16):1419-1430, 2017.

From 3D Atomic Structure to a Single Eigenvalue

Measure the distance between C^{α}_{j} and C^{β}_{i}



Build a **bipartite distance matrix** by comparing two substructures

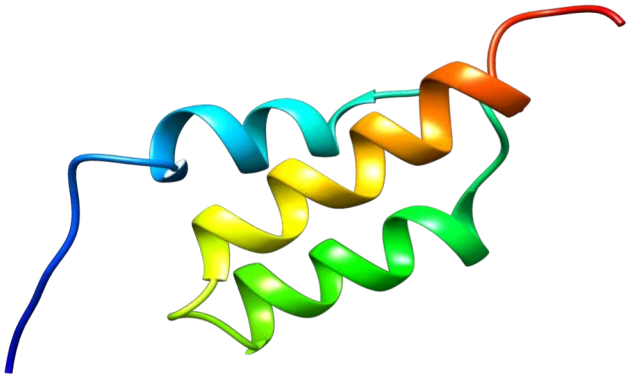
$$D = \begin{array}{c} \\ j \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccc} & & & i & & \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & X & X & X \\ 0 & 0 & 0 & d & X & X \\ 0 & 0 & 0 & X & X & X \\ X & d & X & 0 & 0 & 0 \\ X & X & X & 0 & 0 & 0 \\ X & X & X & 0 & 0 & 0 \end{bmatrix} \end{array}$$

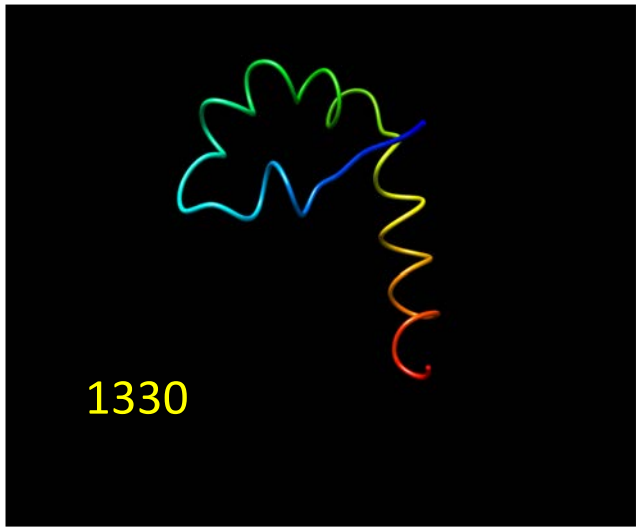
Compute largest eigenvalue → λ_{max}

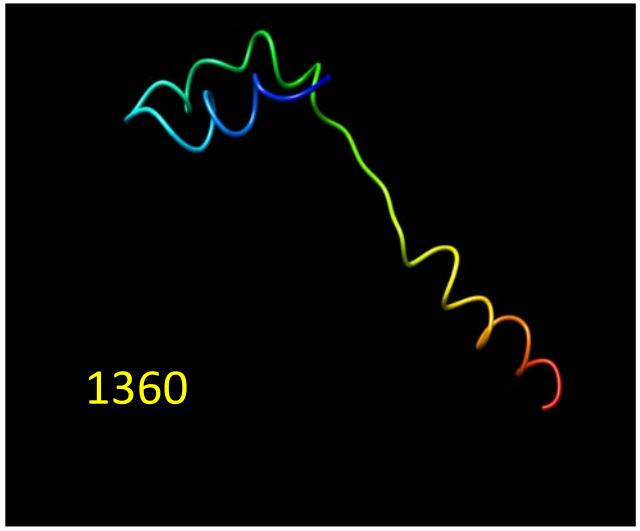
T. Johnston, B. Zhang, A. Liwo, S. Crivelli, and M. Taufer. In-Situ Data Analytics and Indexing of Protein Trajectories. *Journal of Computational Chemistry (JCC),* 38(16):1419-1430, 2017.

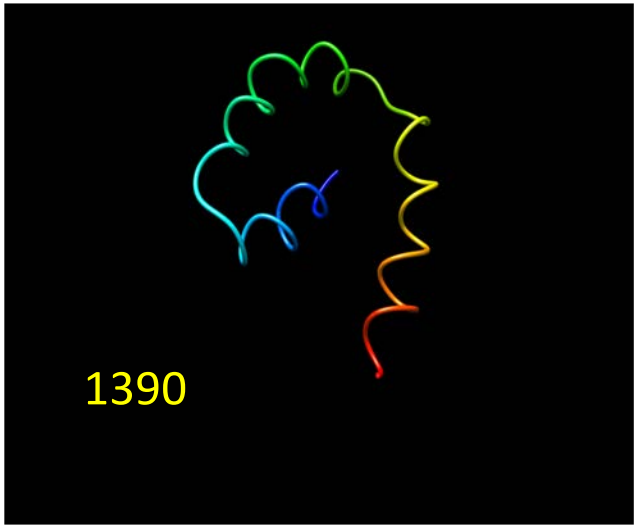
Case Study: Capturing Movement of α-helices

Capture movement of structures (α-helices) with respect to each other





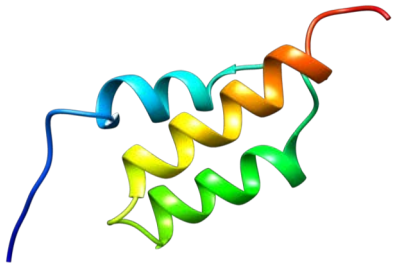


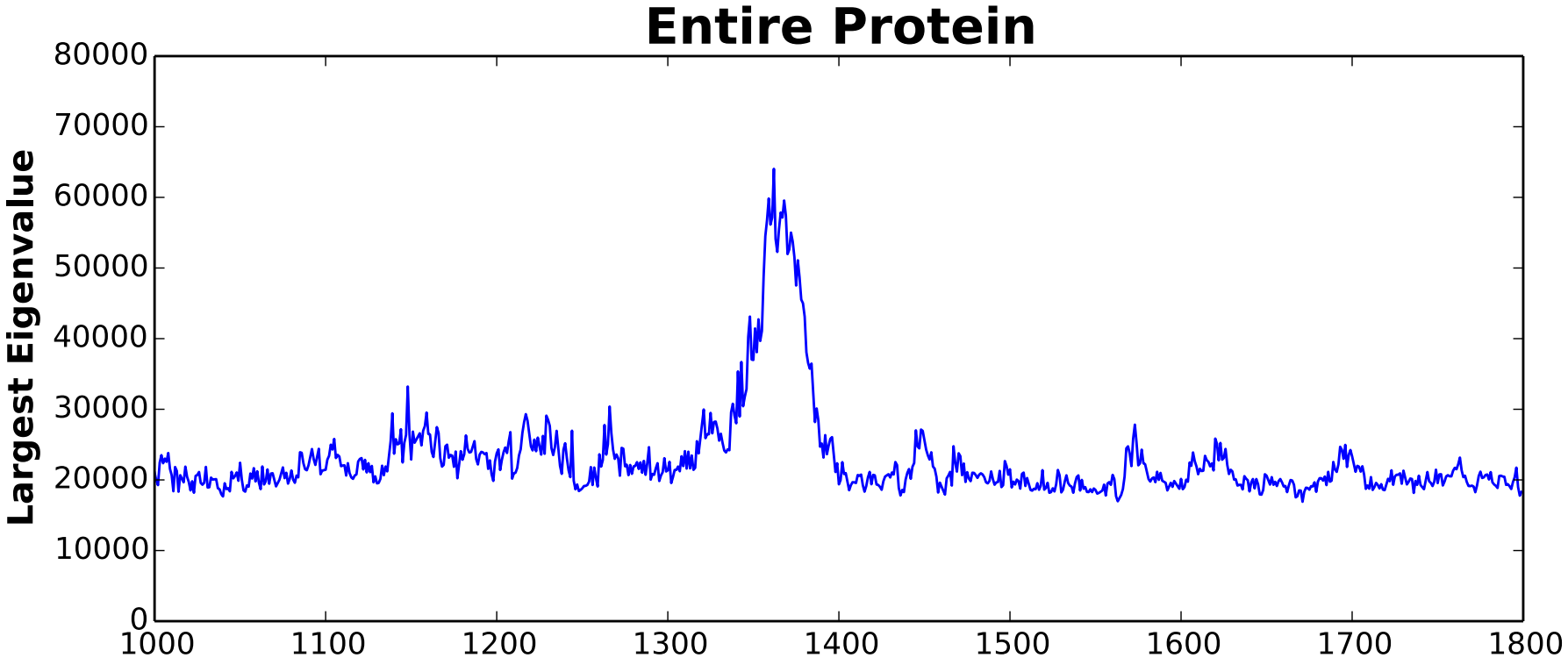


T. Johnston, B. Zhang, A. Liwo, S. Crivelli, and M. Taufer. In-Situ Data Analytics and Indexing of Protein Trajectories. *Journal of Computational Chemistry (JCC),* 38(16):1419-1430, 2017.

Case Study: Capturing Movement of α-helices

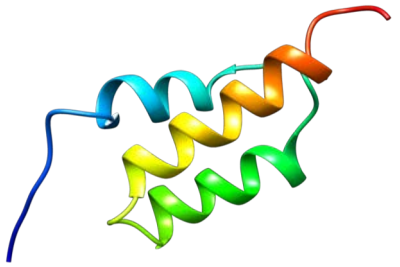
Monitor largest eigenvalue of entire protein

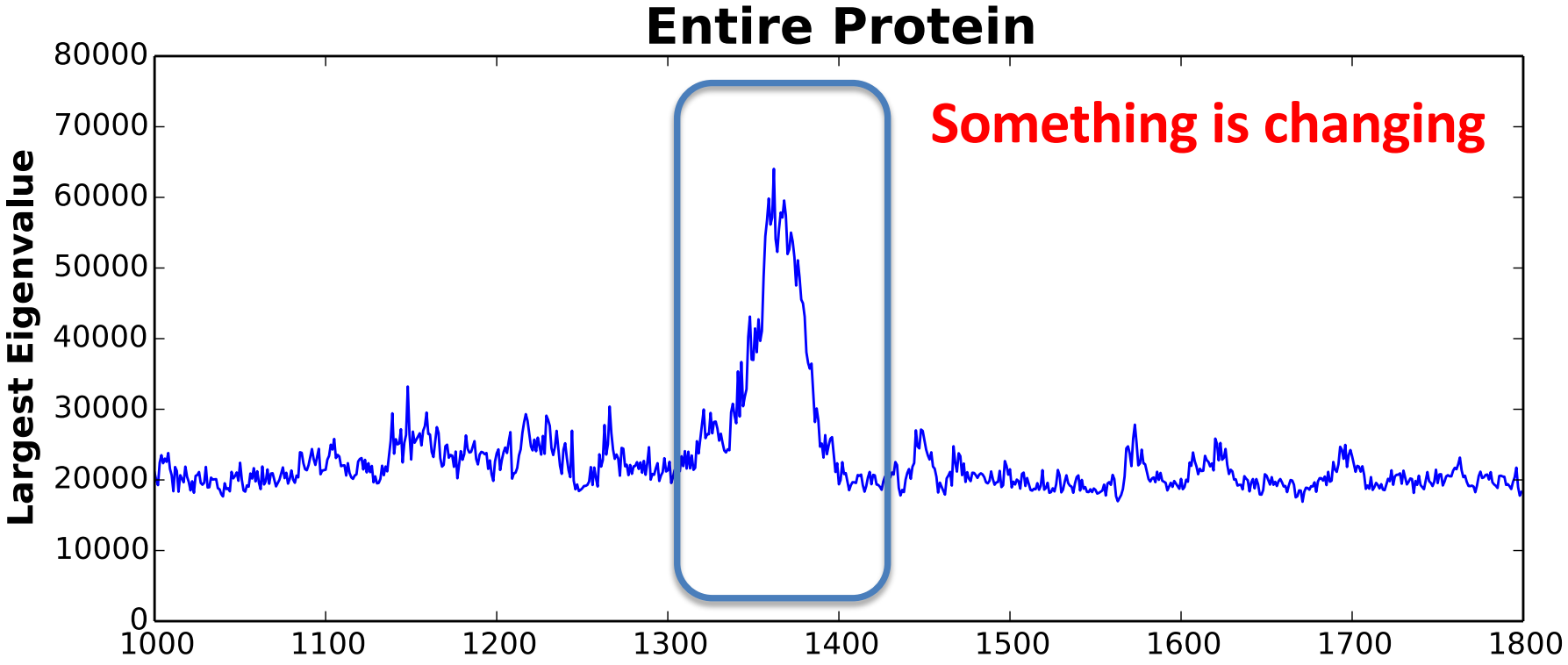




Case Study: Capturing Movement of α-helices

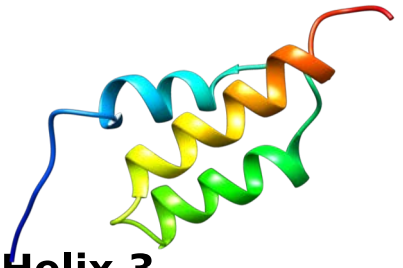
Monitor largest eigenvalue of entire protein

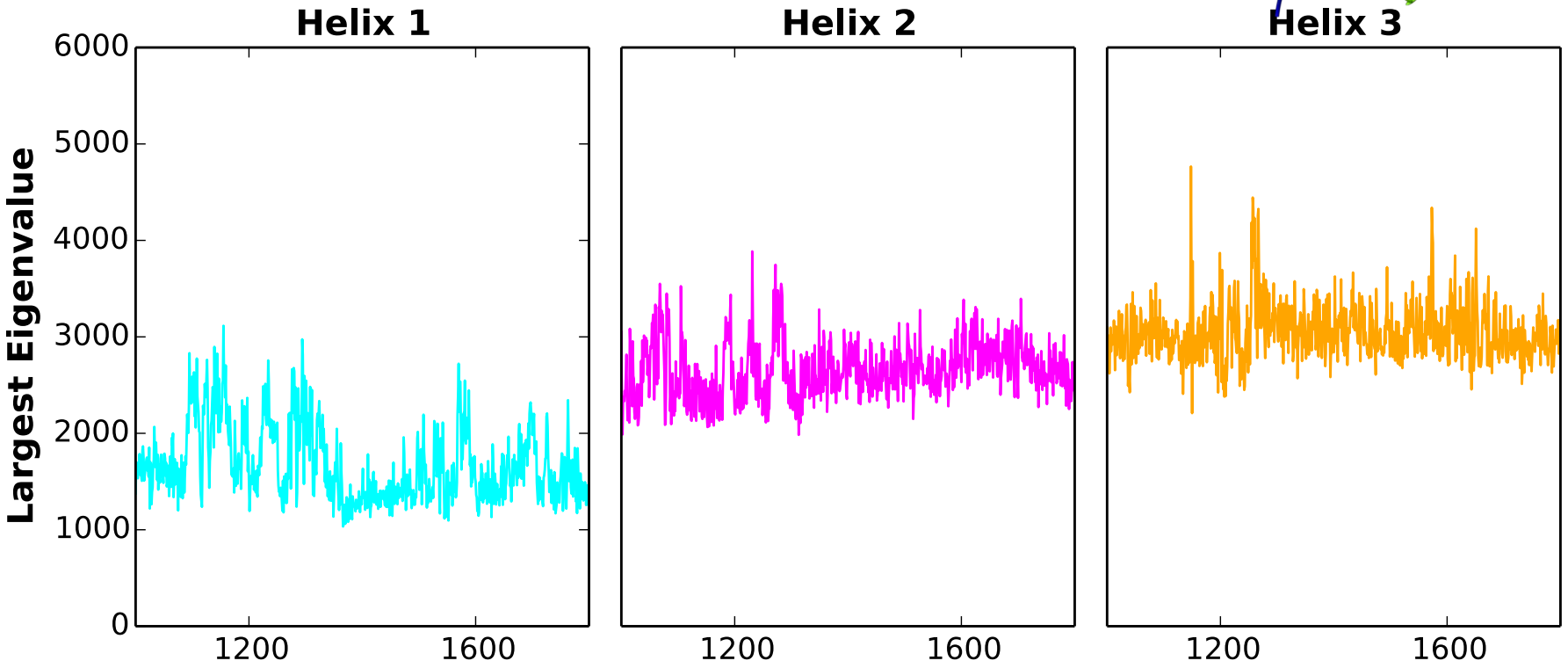




Case Study: Capturing Movement of α-helices

Monitor largest eigenvalue of single α-helices

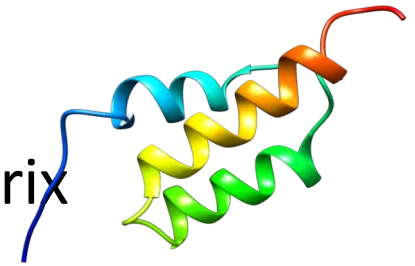


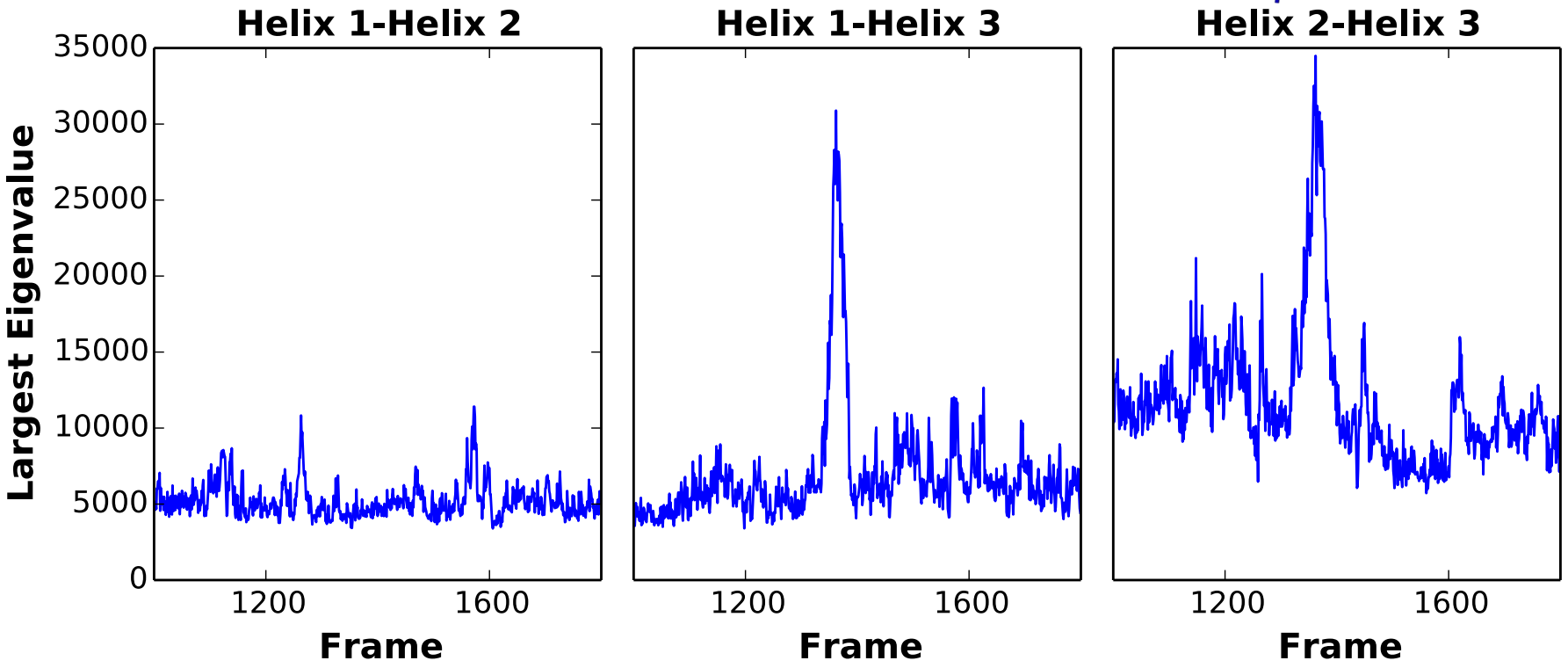


Individual α-helices (Helix 1, Helix 2, and Helix 3) appear stable

Case Study: Capturing Movement of α-helices

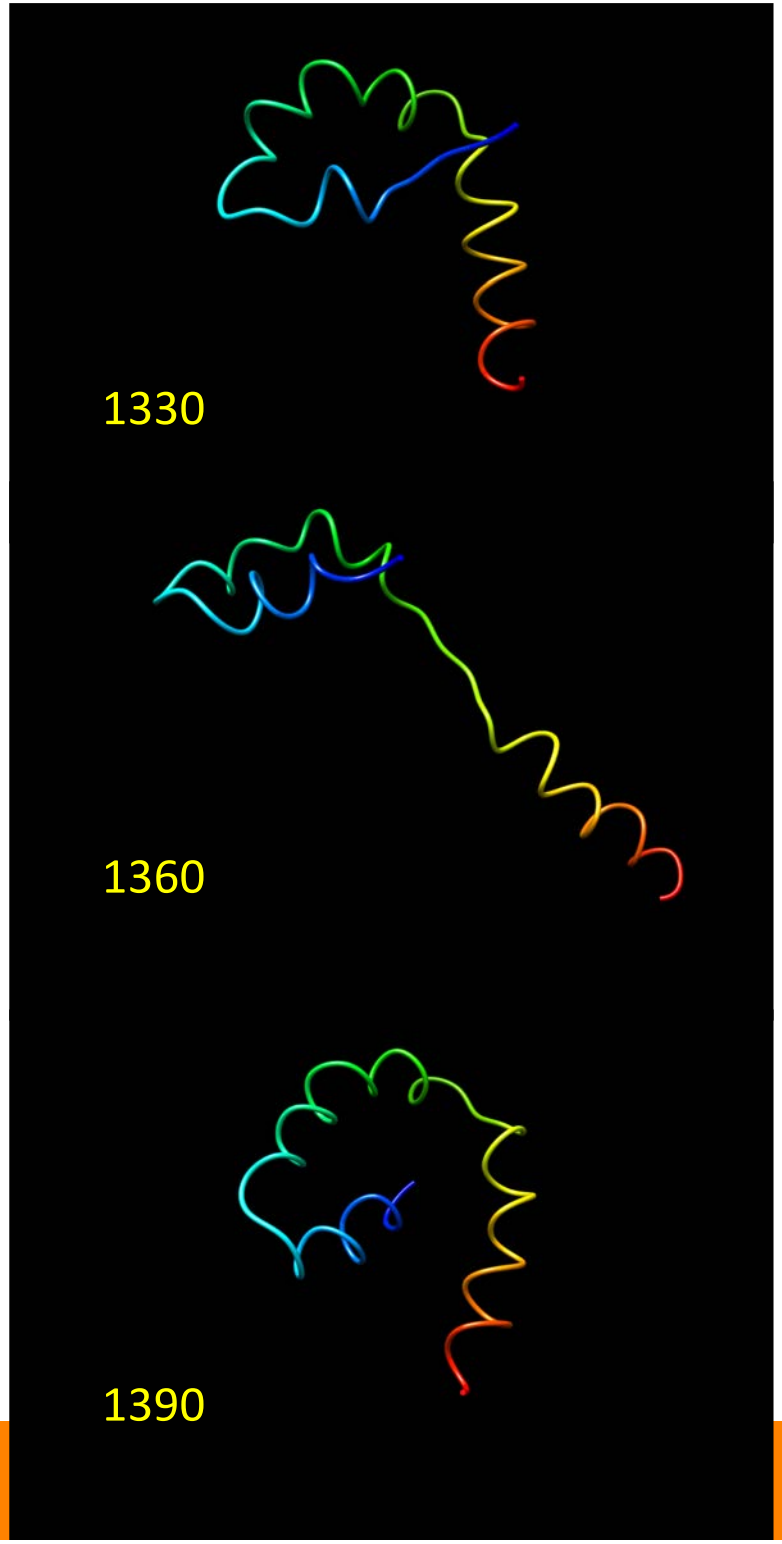
Monitor largest eigenvalue of bipartite distance matrix

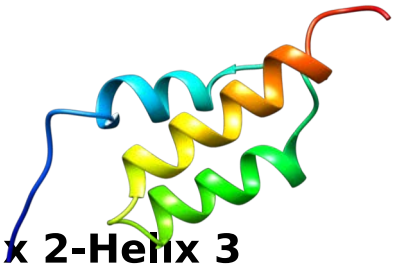


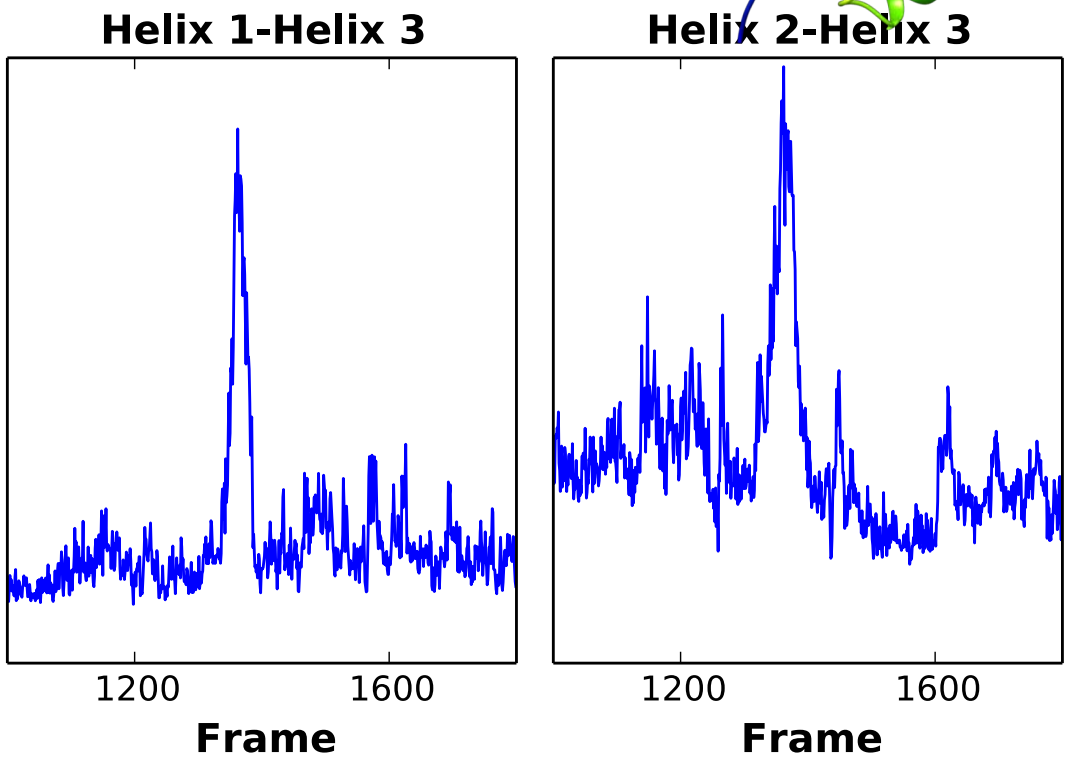


First and second α-helices appear stable; third helix moves

Case Study: Capturing Movement of α-helices







Analysis: Linear in complexity using local metadata (eigenvalues) with DataSpaces

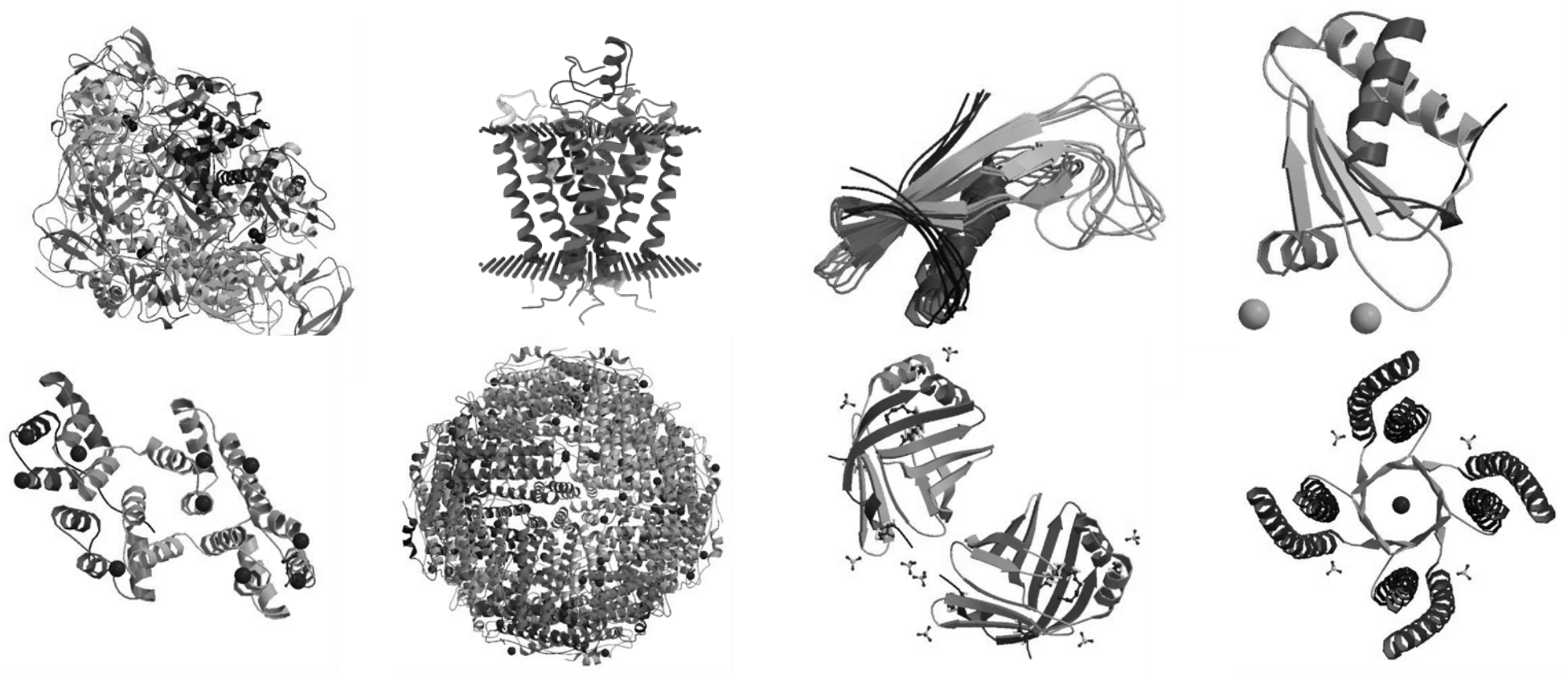
Analytics for Molecular Dynamics

- Drug design and protein-ligand docking
- Protein folding and rare events
- **Protein variants expressed from yeast or bacteria and protein engineering**

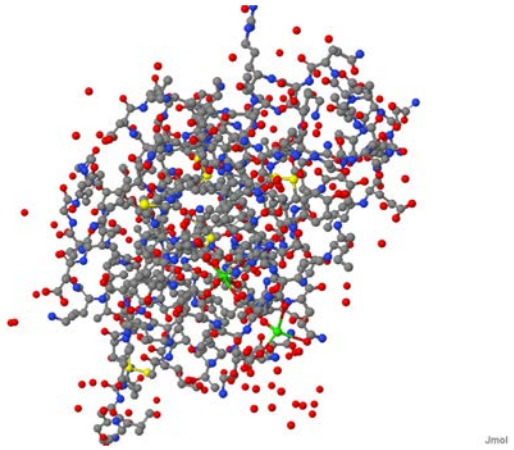
A4MD: Proteins with Similar Functions

Key principle: proteins with similar sequences have similar functions

- Measure millions of protein variants expressed from yeast or bacteria
- Structure proteins to produce desired properties (protein engineering)

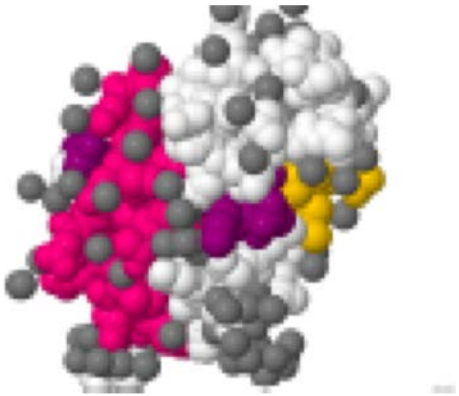


Protein Representations



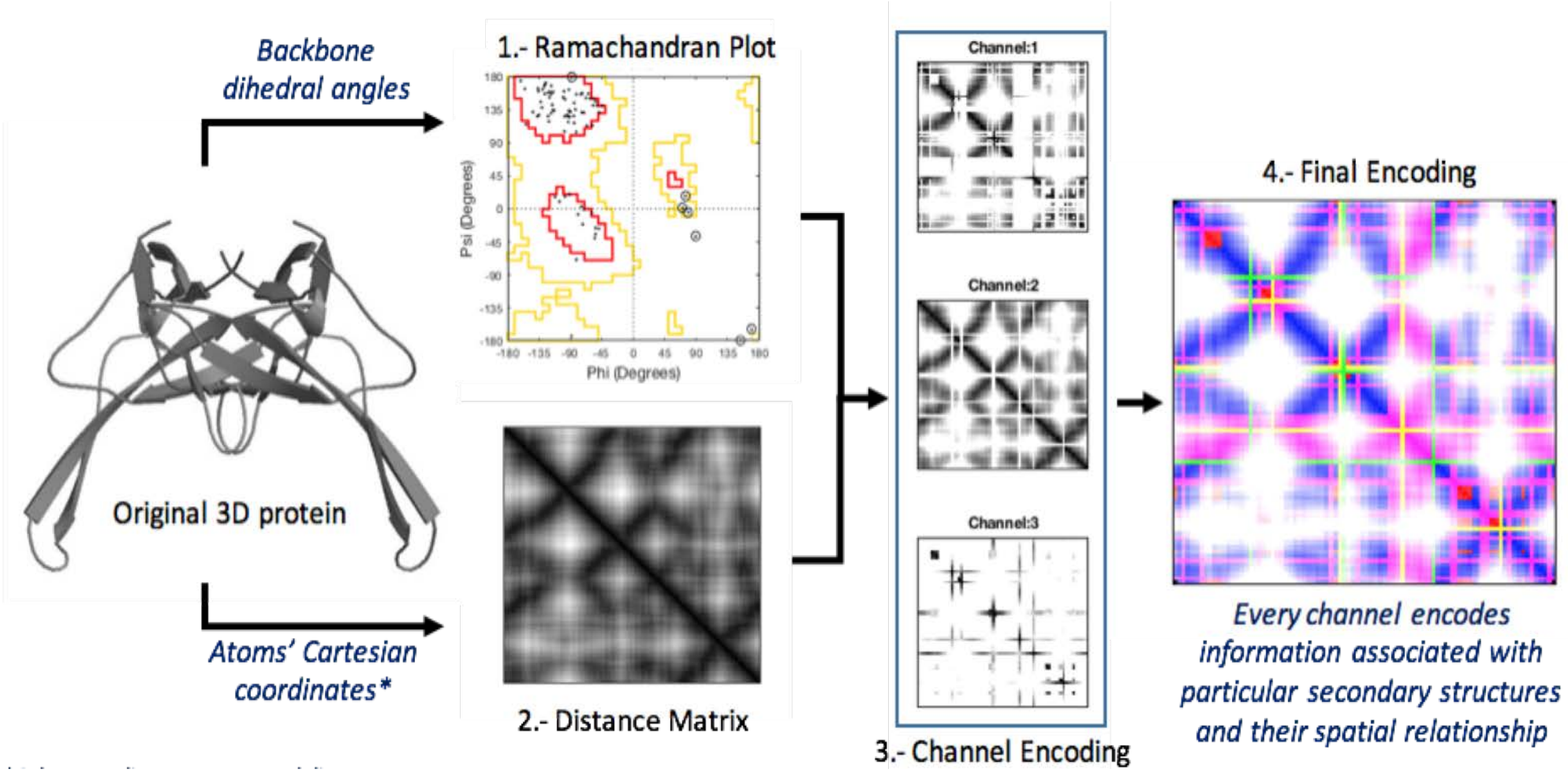
3D Cartesian representation

Multi-fold representation



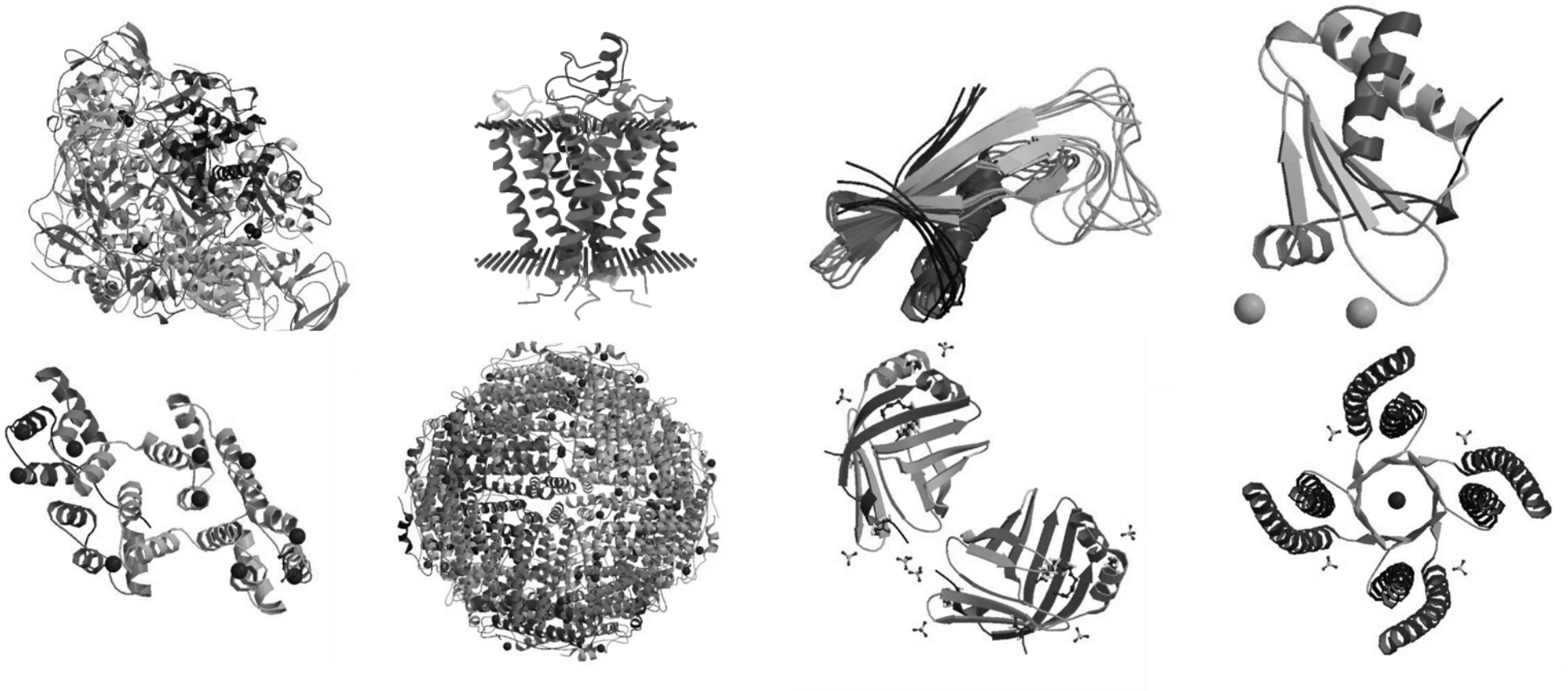
Surface representation

From Multi-fold Representation to Image Encoding



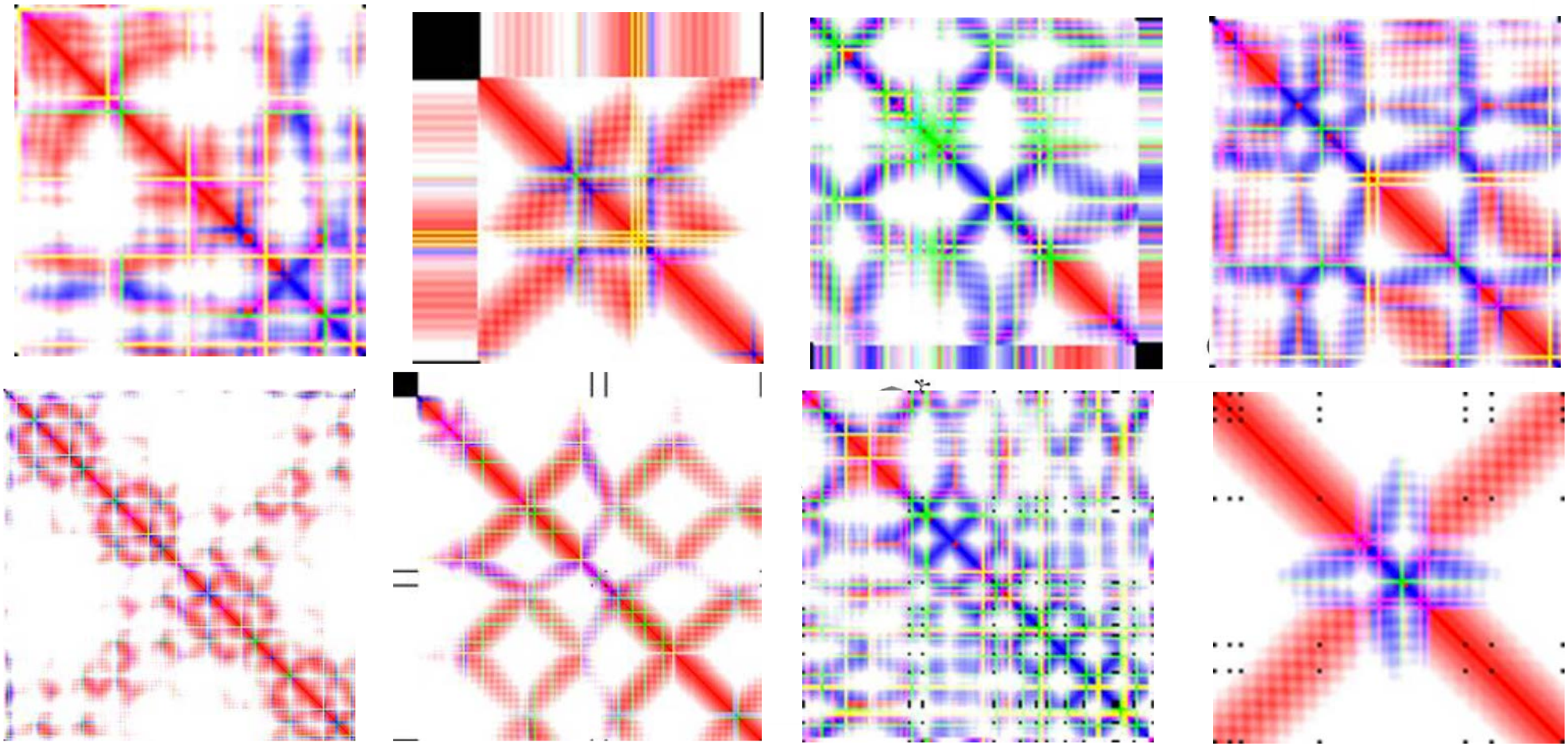
T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. ***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

From Multi-fold Representation to Image Encoding



T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. ***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

From Multi-fold Representation to Image Encoding

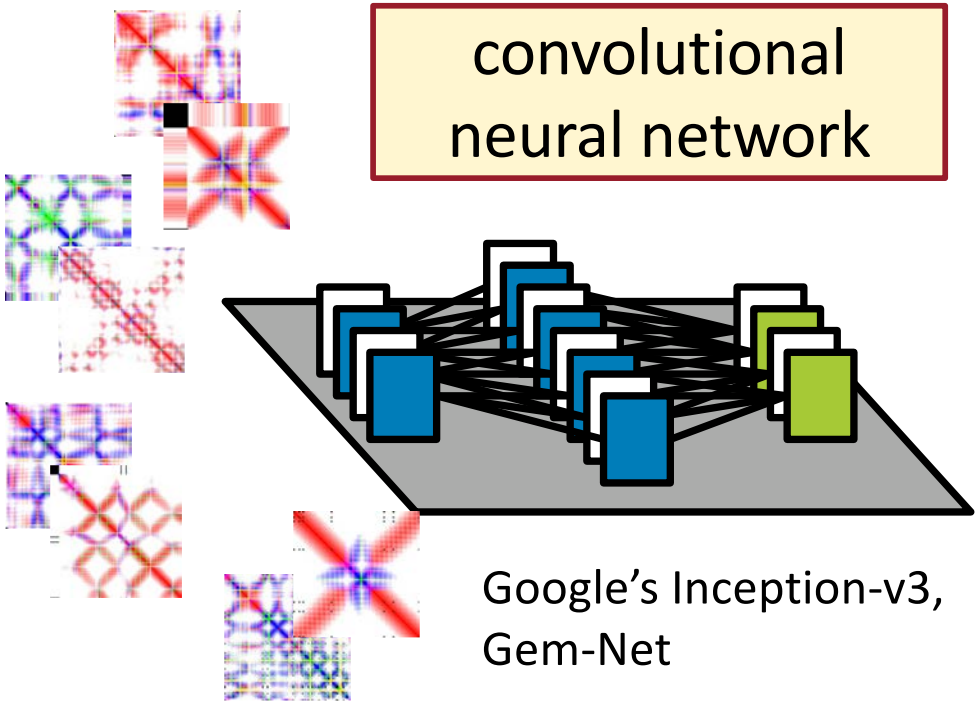


T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. ***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

Case Study: High-Throughput Protein Analysis

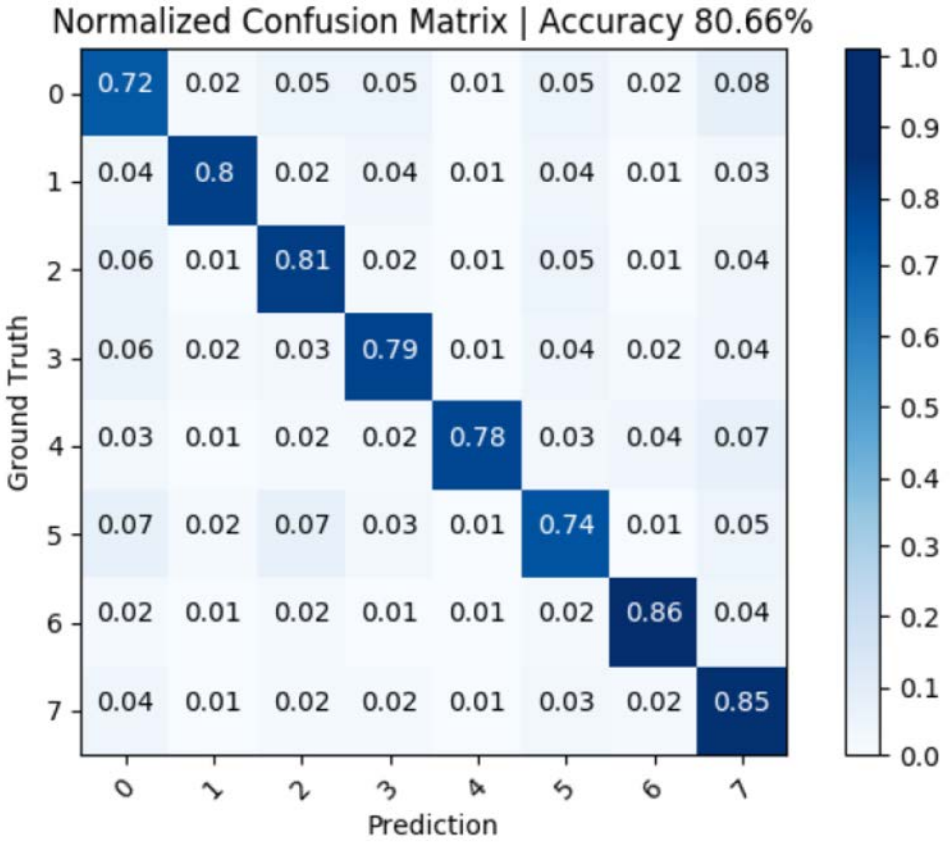
- 62,991 proteins from the Protein Data Bank
- Eight biological processes from biological process taxonomy in RCSB-PDB

Proteins as 3D tens



convolutional neural network

Google's Inception-v3, Gem-Net



T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. ***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

Challenges and Opportunity

A workflow that integrates both simulations and analytics must have these key properties:

- *Efficiency:* Optimize workflows' performance and power usage associated to data movement and analytics
- *Generality:* Build workflows that support different types of analytics across different MD applications
- *Non-invasive:* Capture data from MD simulations without rewriting legacy codes or simulation scripts
- *Portability:* Execute *c*ombined simulations and analytics across different platforms and with heterogenous resources
- *Scalability:* (Re)design ML algorithms for knowledge discovery at scale