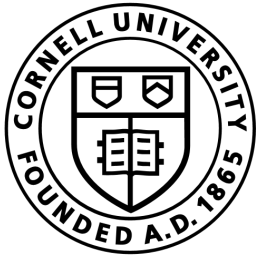
# In Situ Data Analytics for Next Generation Molecular Dynamics Workflows

*Michela Taufer*

*Jack Dongarra Professor in High Performance Computing*
*The University of Tennessee*




THE UNIVERSITY OF TENNESSEE KNOXVILLE

USC University of Southern California

# Acknowledgements


T. Estrada


H. Weinstein


M.A. Cuendet


E. Deelman


R. da Silva


T. Johnston

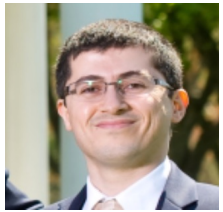

T. Do


M. Wyattc


E. Kots


S. Thomas


H. Carrillo


A. Razavi


S. Caino-Lores


I. Lumsden

## Global Computing Lab@UTK

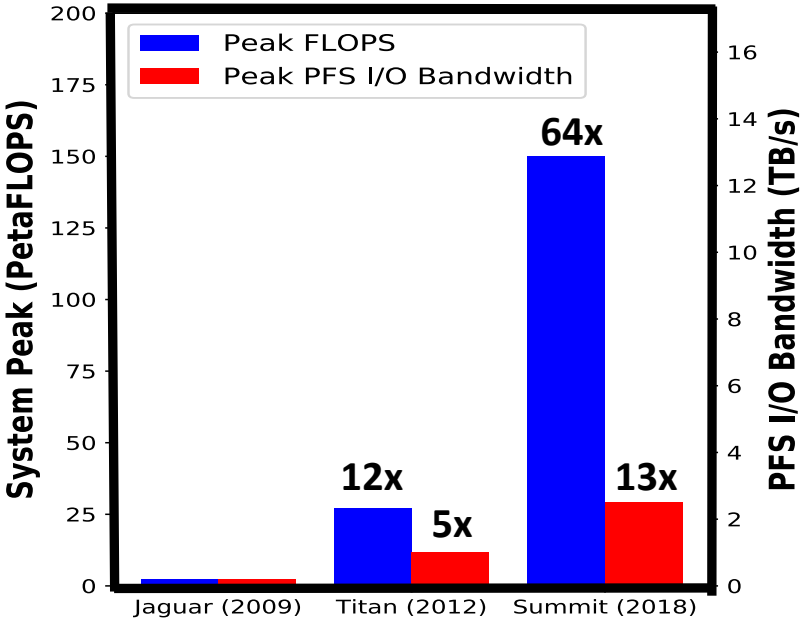
## Sponsors:



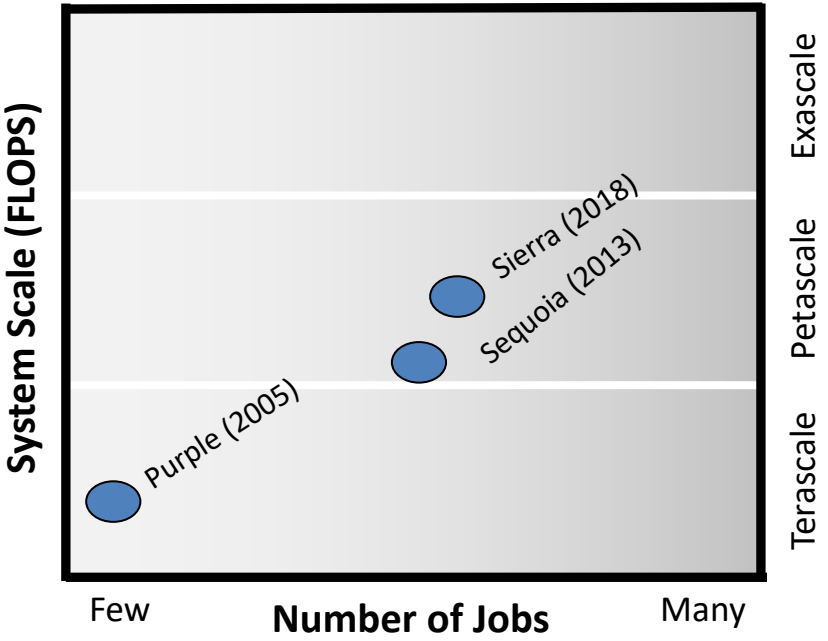
1741057, 1740990,
1741040, 1841758

# Trends in HPC: Computing-IO Gap and Ensembles of Jobs

## Widening IO Gap




## Rising Importance of Ensembles



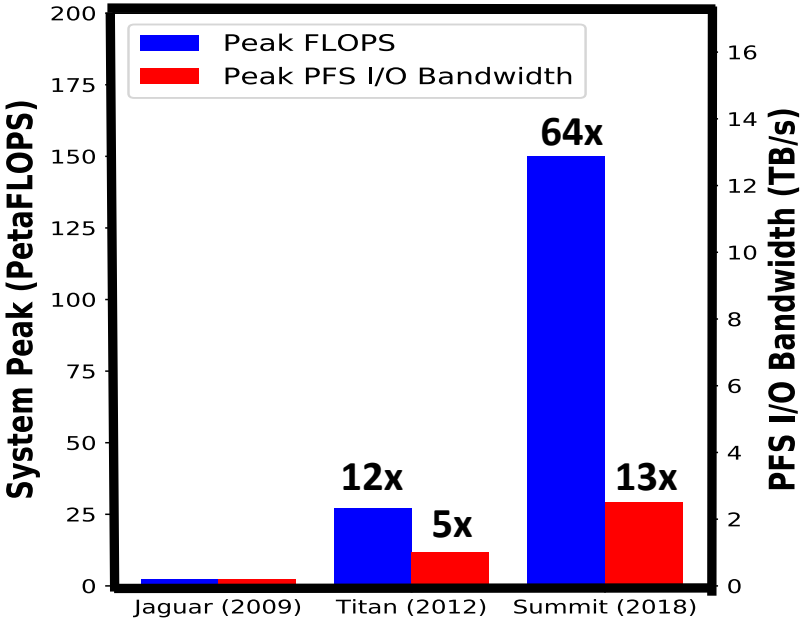



Source: Lucy Nowell (DOE)

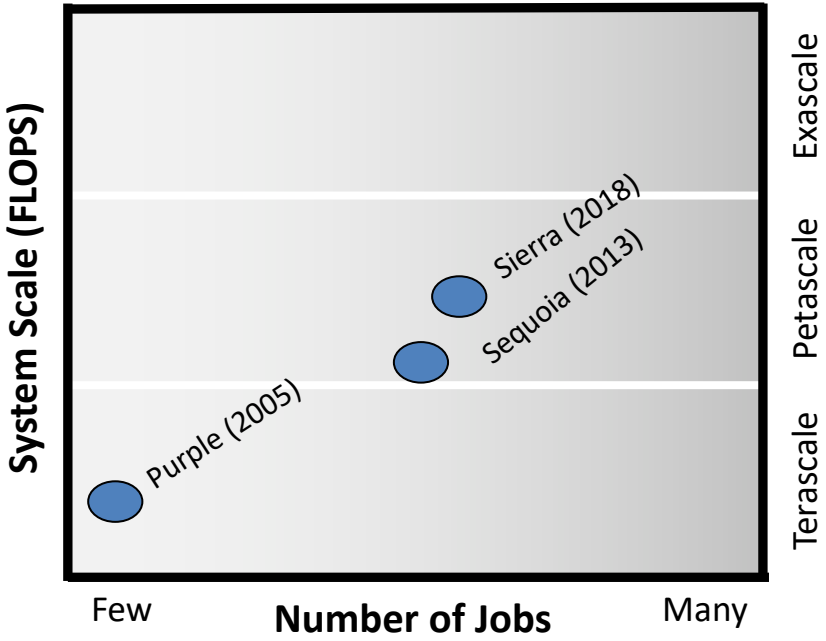
Source: https://wci.llnl.gov/simulation/computer-codes/uncertainty-quantification

# Trends in HPC: Computing-IO Gap and Ensembles of Jobs

## Widening IO Gap




## Rising Importance of Ensembles




**More generated data to move slower to storage**

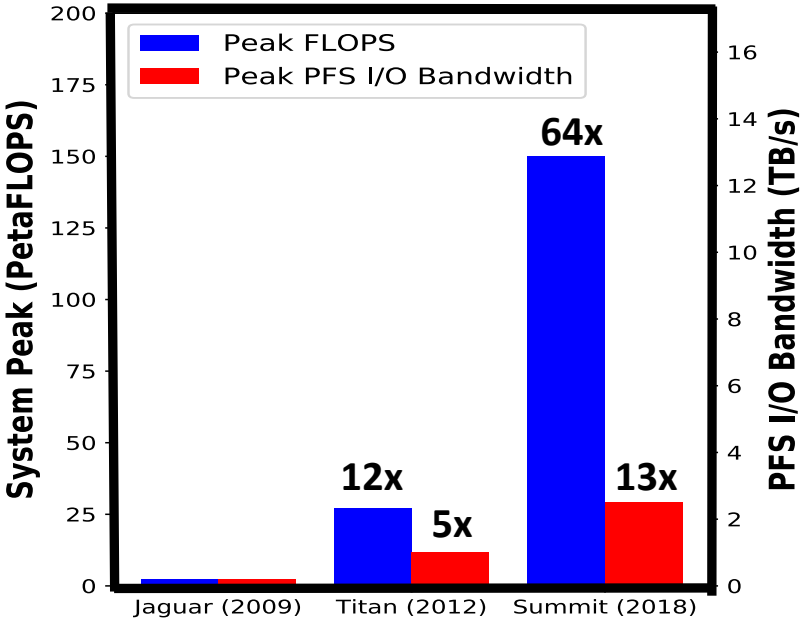


Source: Lucy Nowell (DOE)

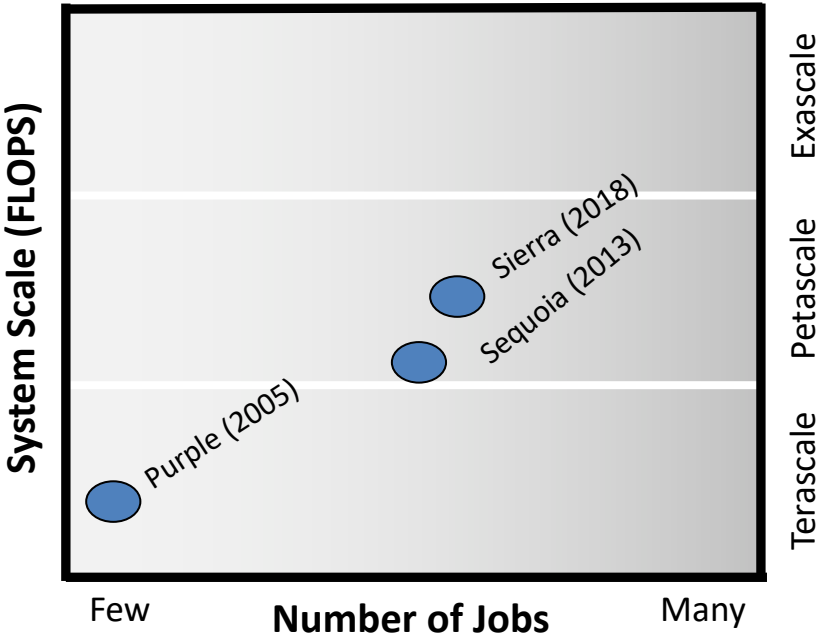
Source: https://wci.llnl.gov/simulation/computer-codes/uncertainty-quantification

# Trends in HPC: Computing-IO Gap and Ensembles of Jobs

## Widening IO Gap




## Rising Importance of Ensembles




**More local data to build a global knowledge from**

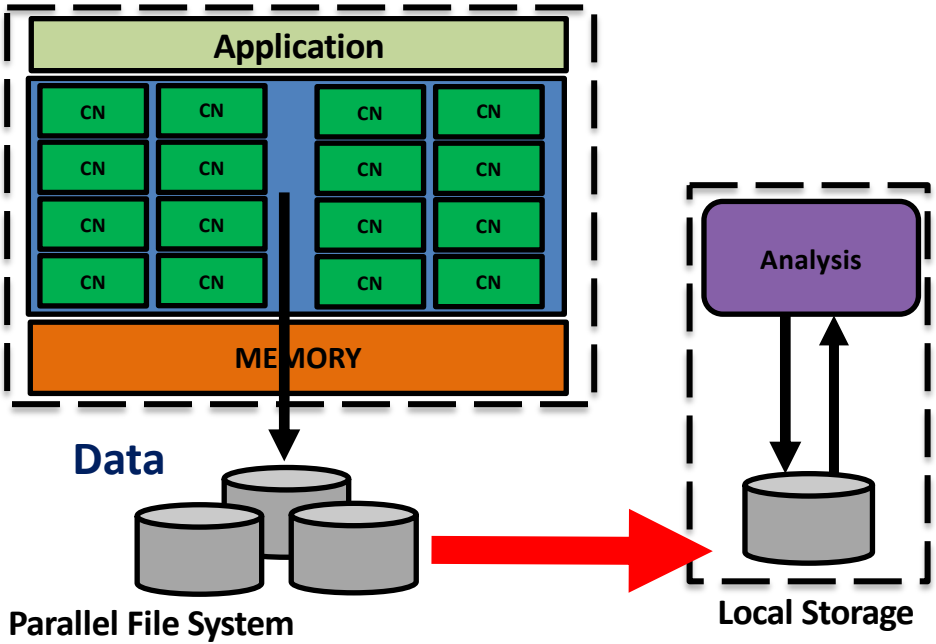


Source: Lucy Nowell (DOE)

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

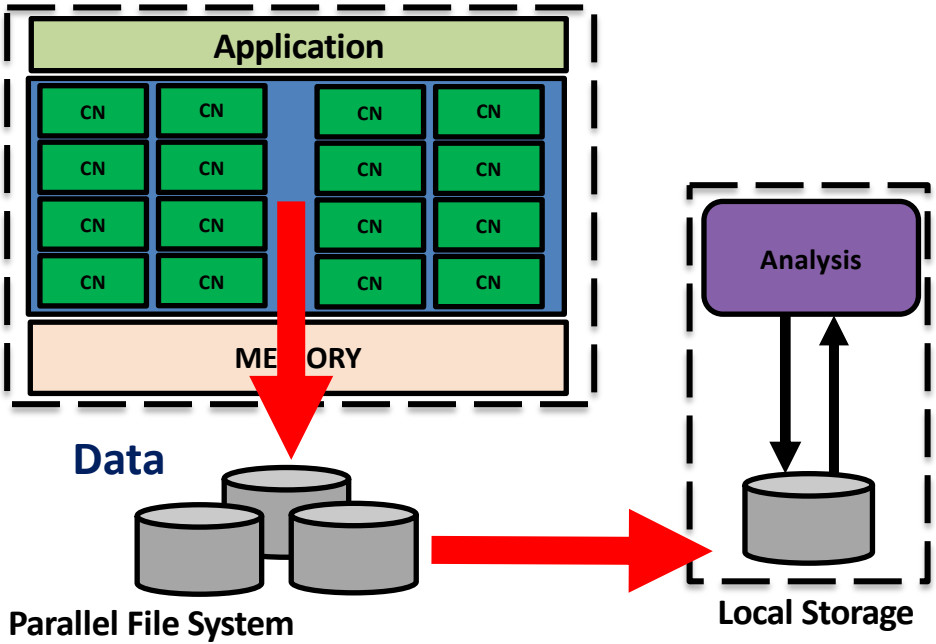
# Extending HPC to Integrate Data Analytics

**Sequential mode**:



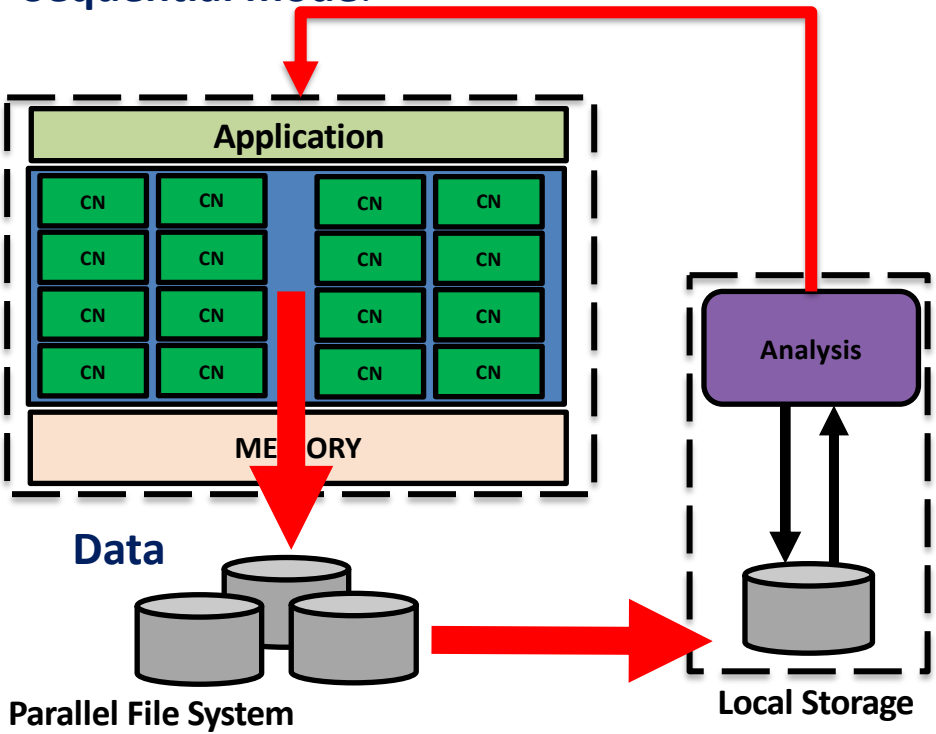
# Extending HPC to Integrate Data Analytics

**Sequential mode**:

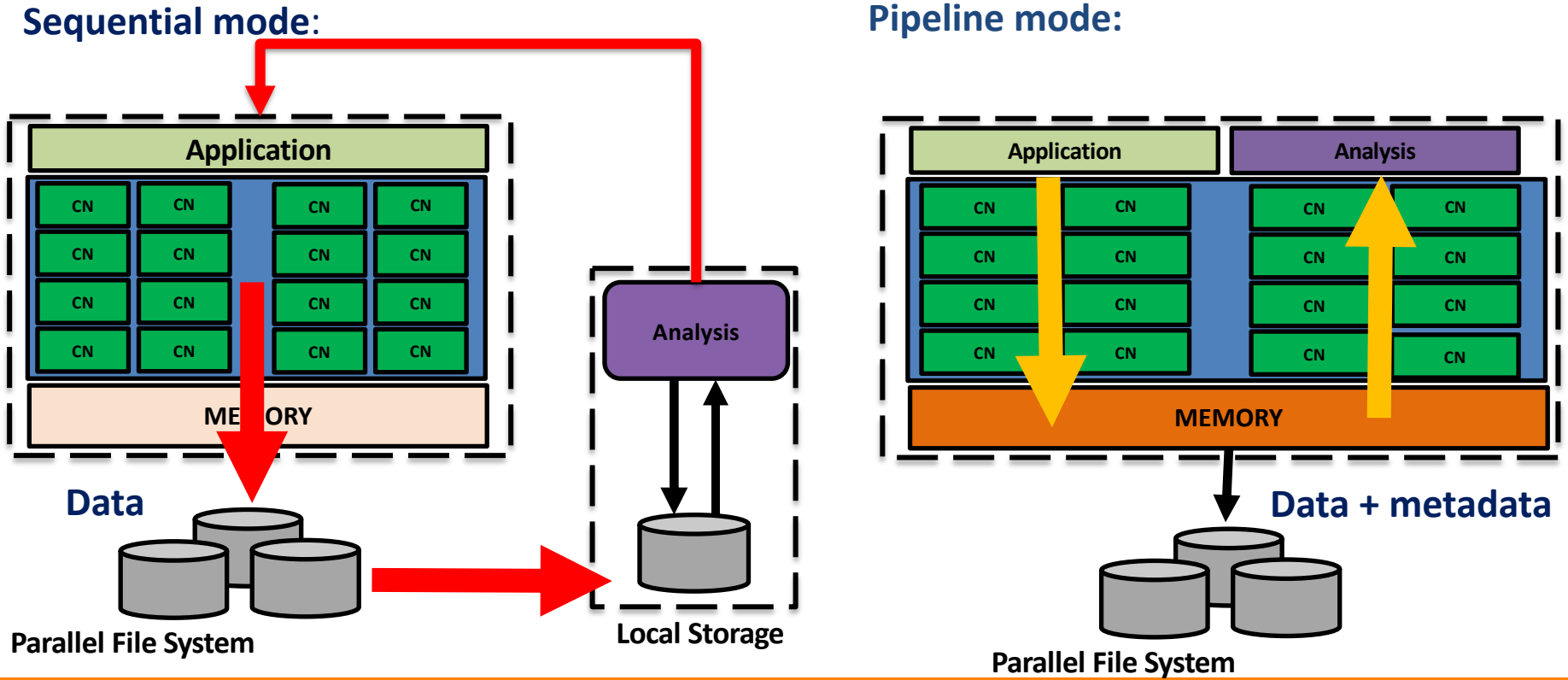


# Extending HPC to Integrate Data Analytics

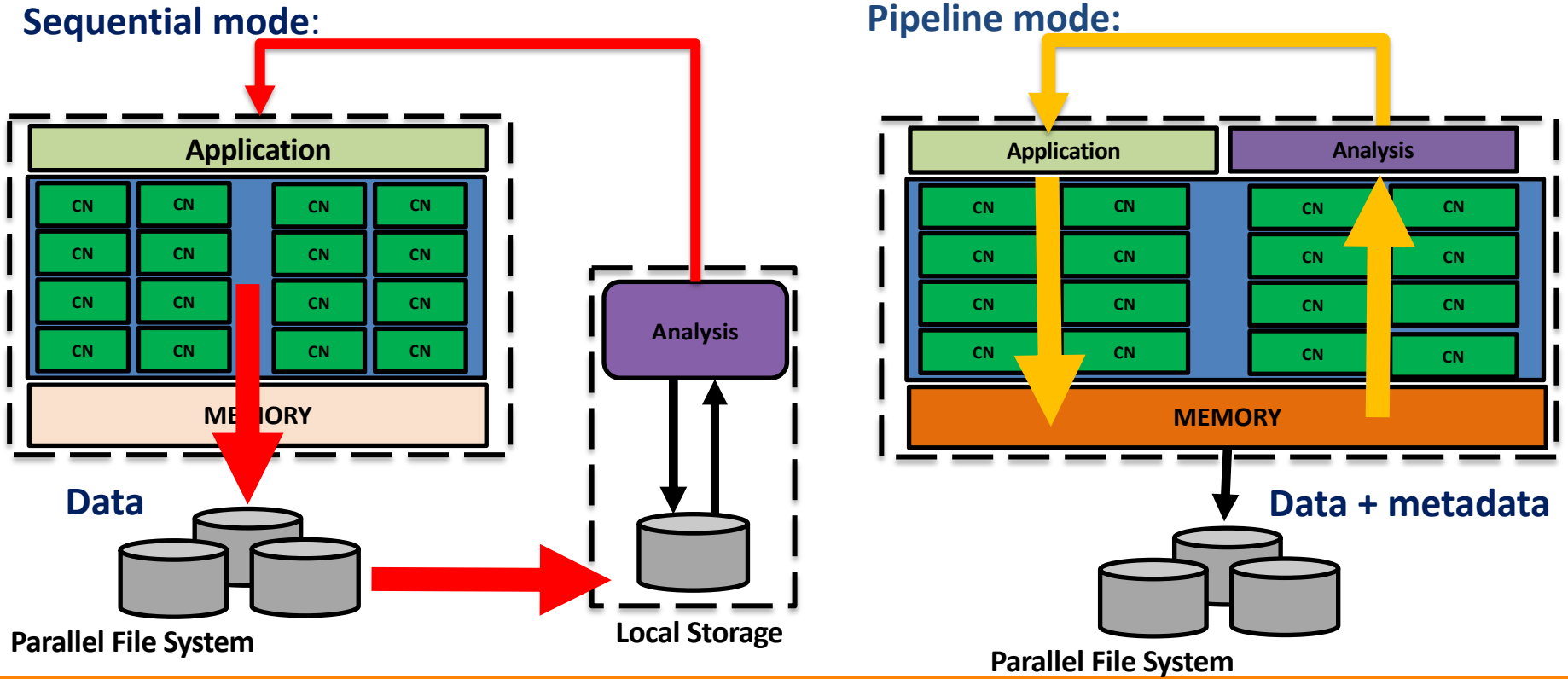
Sequential mode:



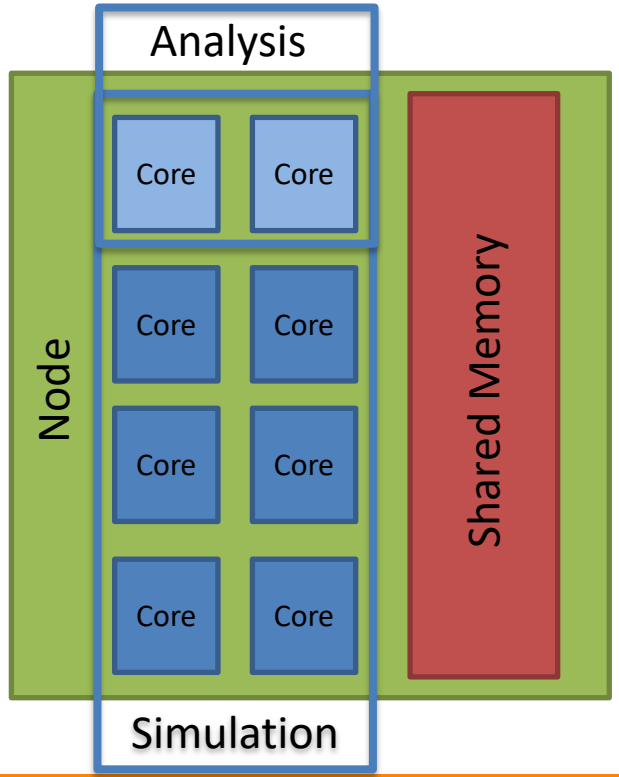
# Extending HPC to Integrate Data Analytics

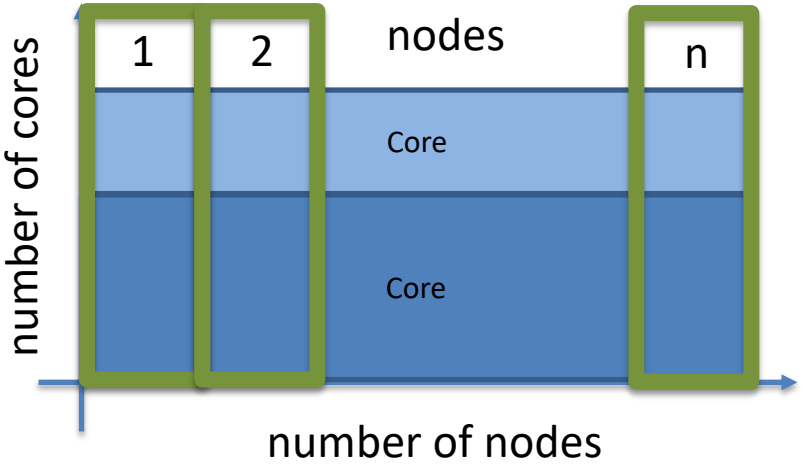


# Extending HPC to Integrate Data Analytics



# Augmenting HPC with In Situ Analytics

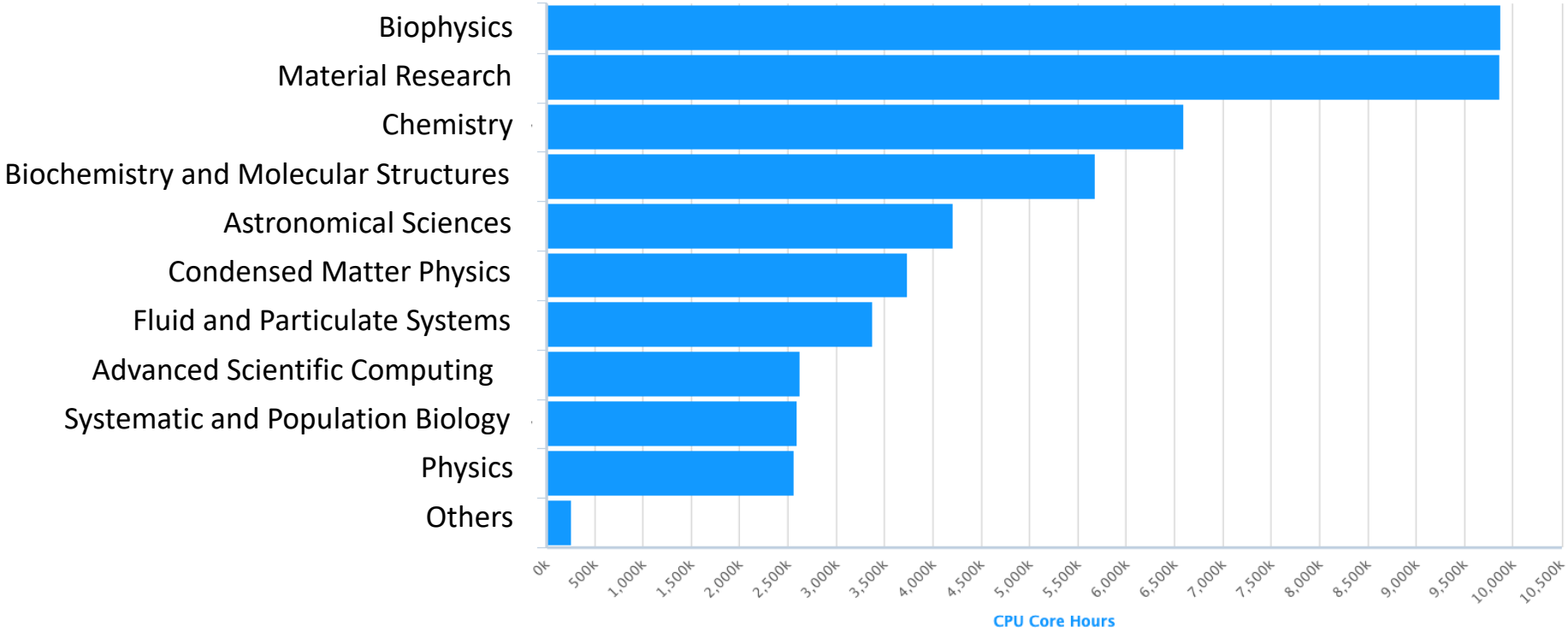



Example of software tools:

- DataSpaces (Rutgers U. / Utah U.)
- Decaf (ANL)
- DYAD (LLNL)

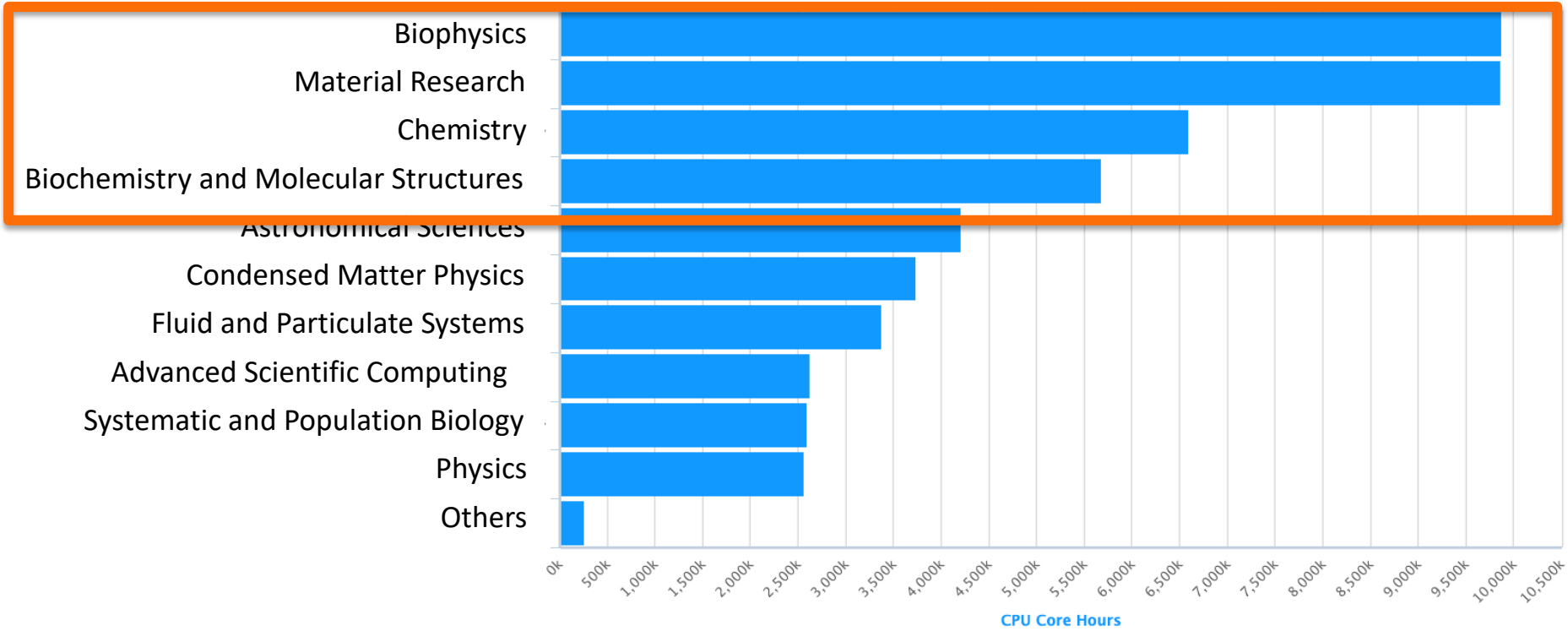
# Major Computational Domains on NSF Systems






Supported by NSF cyberinfrastructures (XSEDE). Source: https://xdmod.ccr.buffalo.edu/

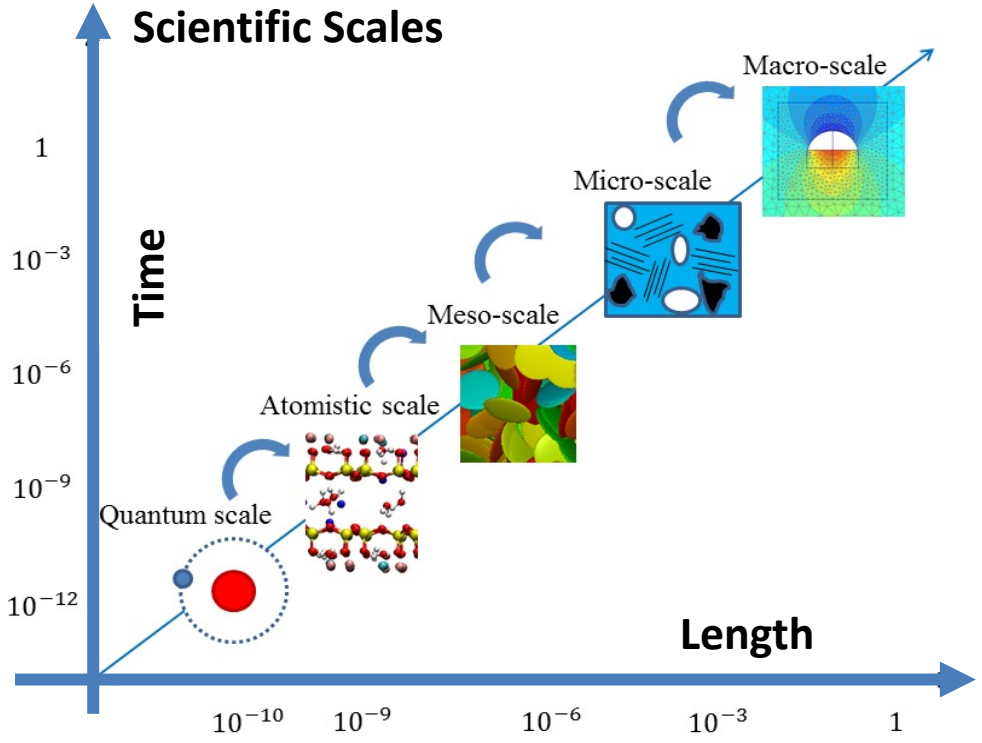
# Major Computational Domains on NSF Systems

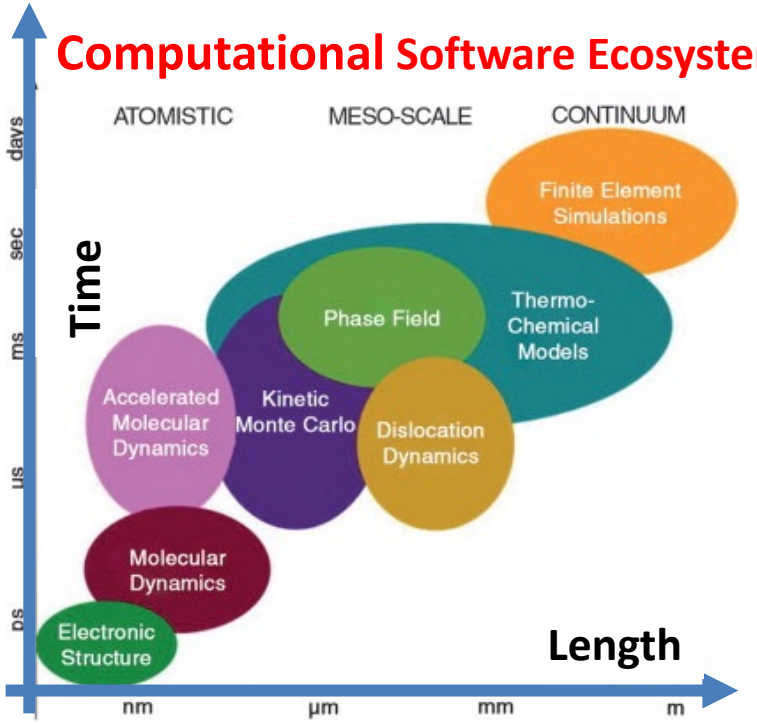





Supported by NSF cyberinfrastructures (XSEDE). Source: https://xdmod.ccr.buffalo.edu/

# Multiscale Computational Software Ecosystem



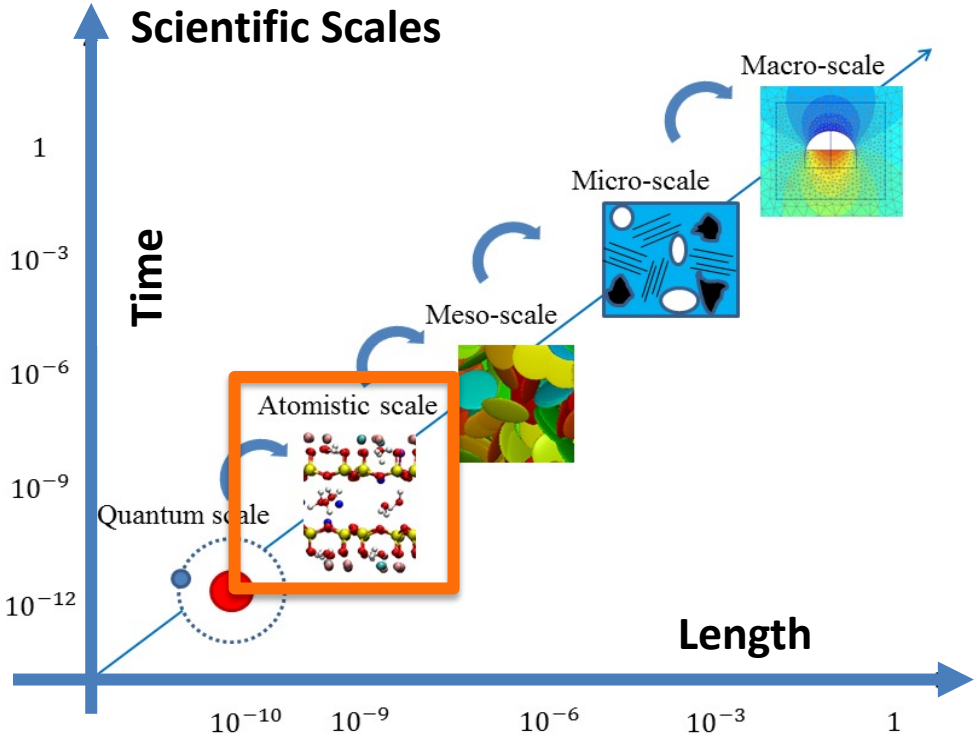



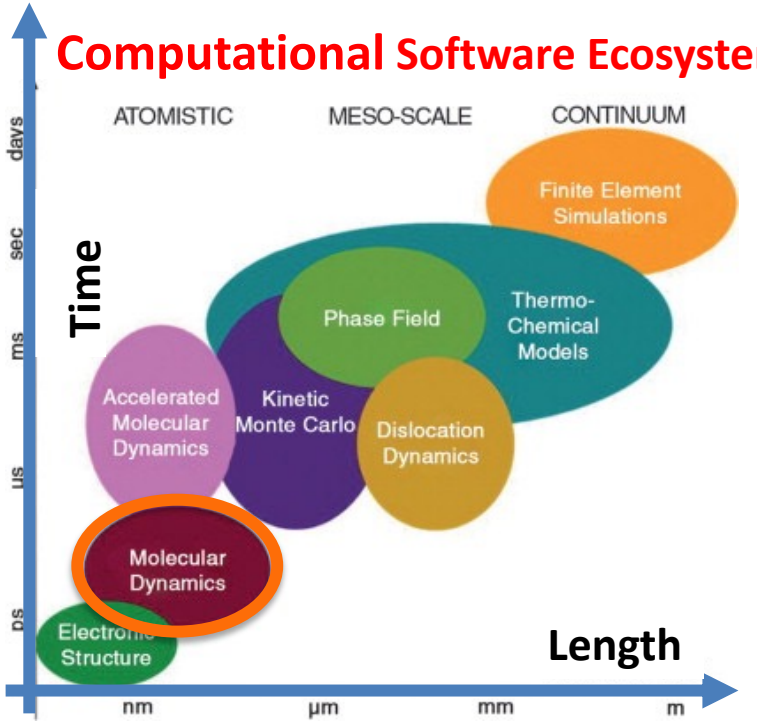



Source: https://ajw-group.mit.edu/multiscale-modeling-clays

M Stan, Material Today, 12, 2009, 20-28

# Multiscale Computational Software Ecosystem



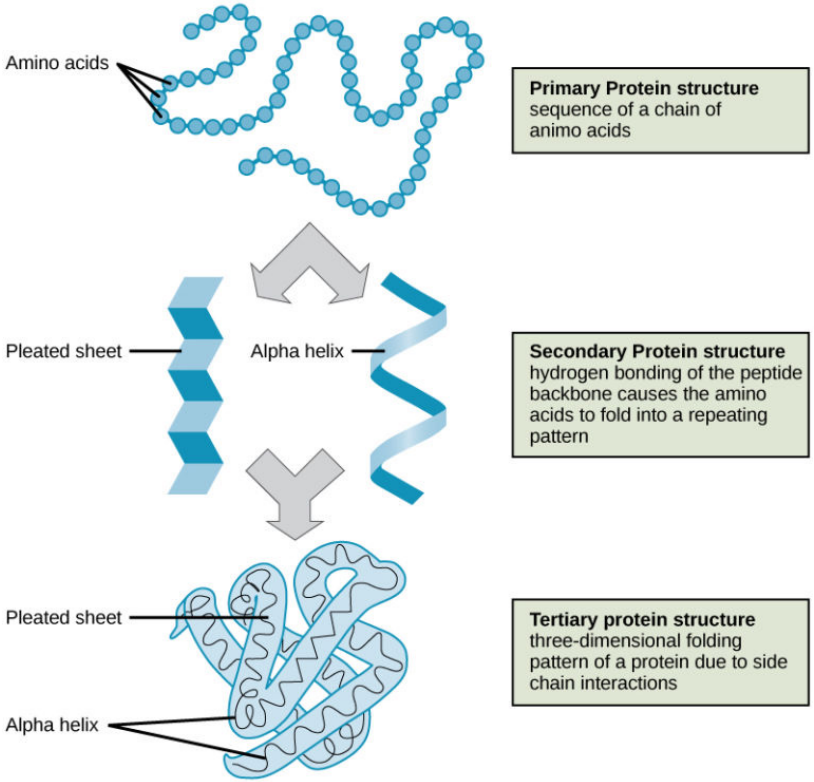






Source: https://ajw-group.mit.edu/multiscale-modeling-clays

M Stan, Material Today, 12, 2009, 20-28

# Classical Molecular Dynamics Simulations




**MD algorithm:**

Given the amino acidic atoms,

Run *np* steps:

→Forces on single atoms

→ Atom acceleration

→ Atom velocity

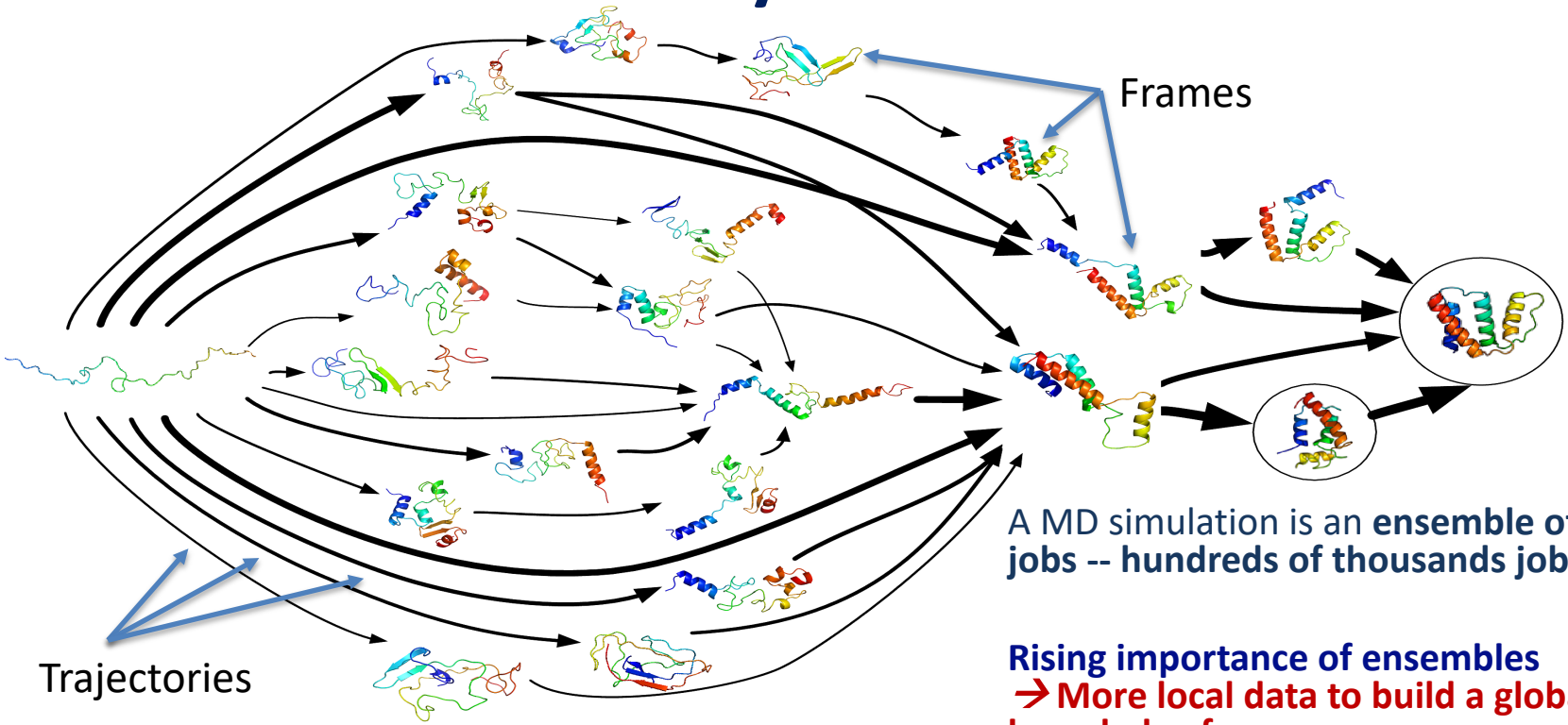
→ New position

Every *n* steps (n << np),

Output atom coordinates (frame)

**Widening IO Gap → *more data are generated and moved to the slower storage***

# Classical Molecular Dynamics Simulations

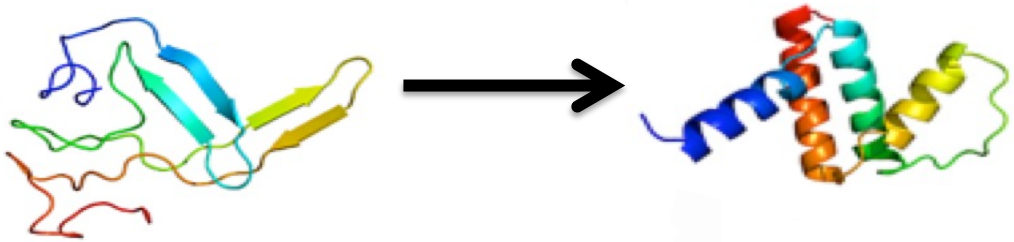



A MD simulation is an **ensemble of MD jobs -- hundreds of thousands jobs**

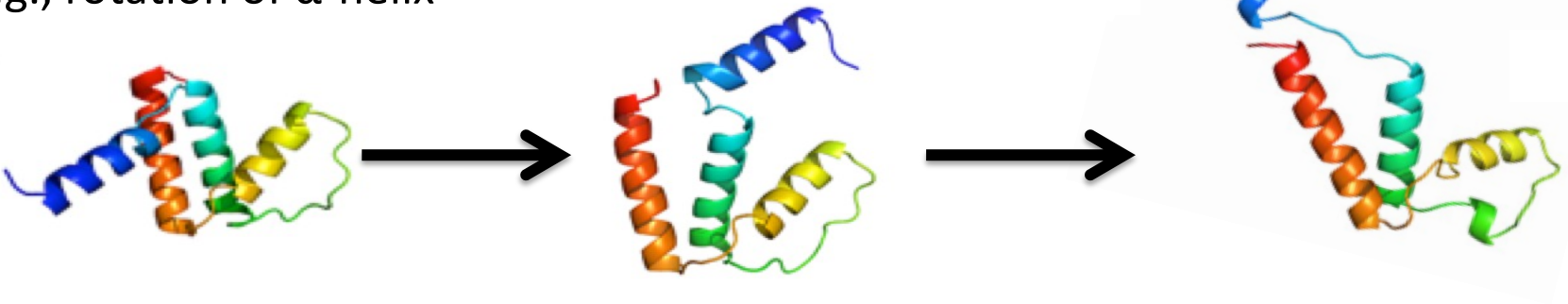
**Rising importance of ensembles**
**→More local data to build a global knowledge from**

# Capturing Rare Events in Trajectories

*Transformations:*
e.g., from β-sheets
to α-helixes

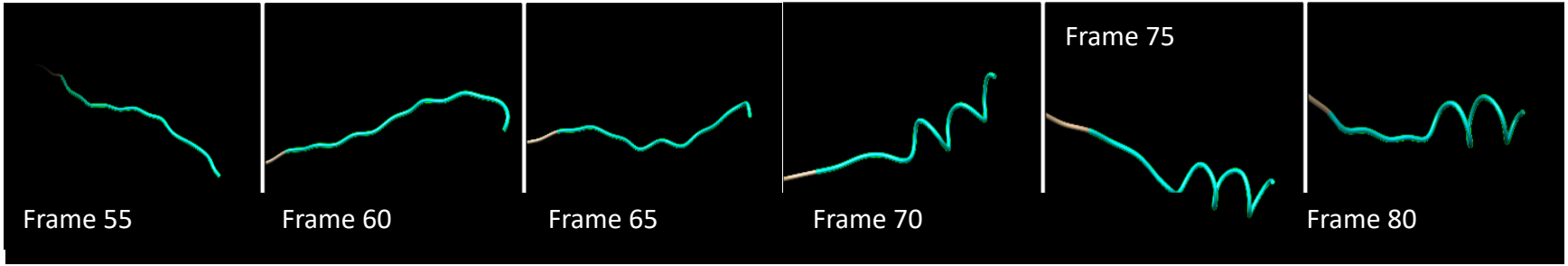


*Movements:*
e.g., rotation of α-helix



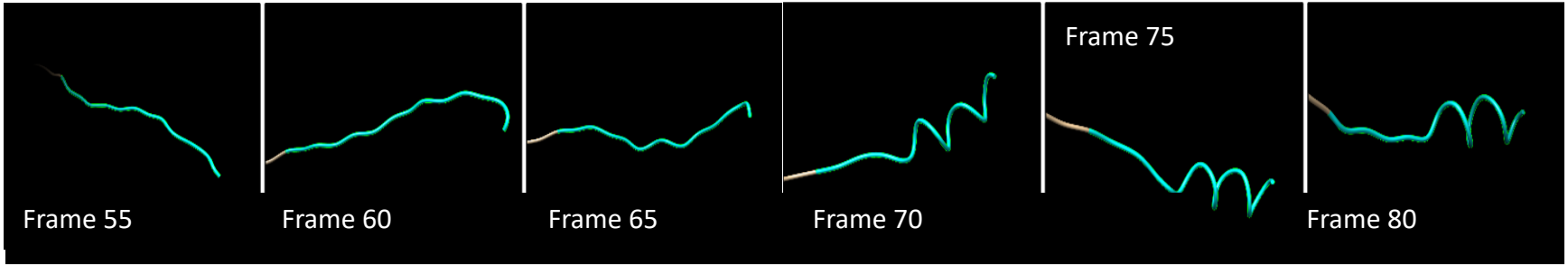
# From Visualization-driven Analytics of MD Trajectories …

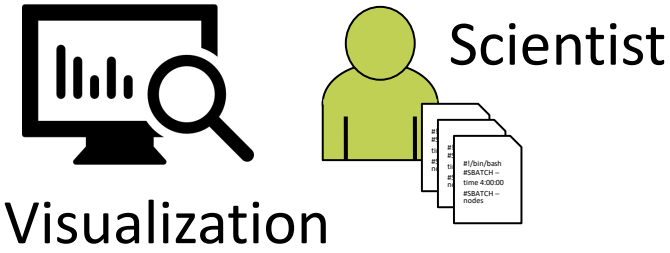
Frames (or snapshots) of an MD trajectory with a stride of 5 steps:



# From Visualization-driven Analytics of MD Trajectories …

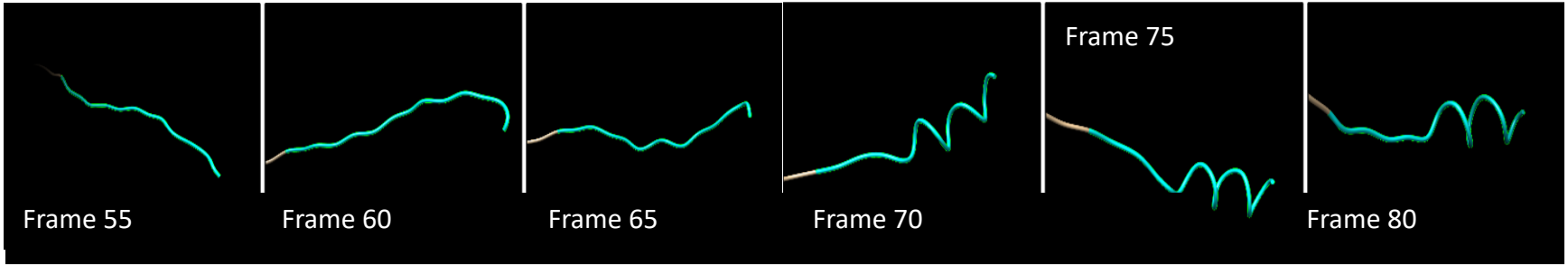
Frames (or snapshots) of an MD trajectory with a stride of 5 steps:



# … To In Situ Analytics of MD Simulations

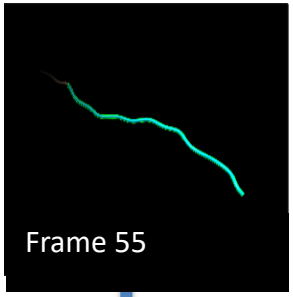
Frames (or snapshots) of an MD trajectory with a stride of 5 steps:




- 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

# In Situ Analytics of MD Trajectories

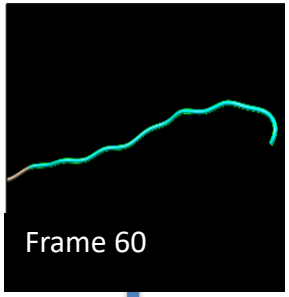
Frames (or snapshots) of an MD trajectory with a stride of 5 steps:




**Collective variables (time 55)**

# In Situ Analytics of MD Trajectories

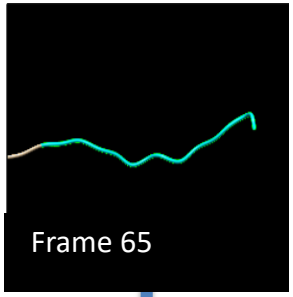
Frames (or snapshots) of an MD trajectory with a stride of 5 steps:




**Collective variables (time 60)**

# In Situ Analytics of MD Trajectories

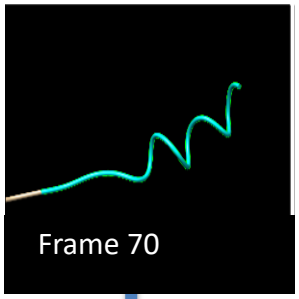
Frames (or snapshots) of an MD trajectory with a stride of 5 steps:




**Collective variables (time 65)**

# In Situ Analytics of MD Trajectories

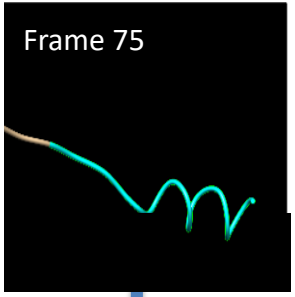
Frames (or snapshots) of an MD trajectory with a stride of 5 steps:




**Collective variables (time 70)**

# In Situ Analytics of MD Trajectories

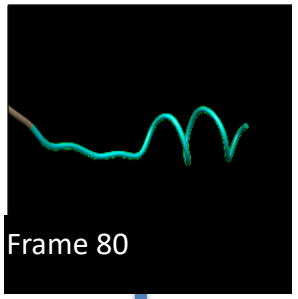
Frames (or snapshots) of an MD trajectory with a stride of 5 steps:




**Collective variables (time 75)**

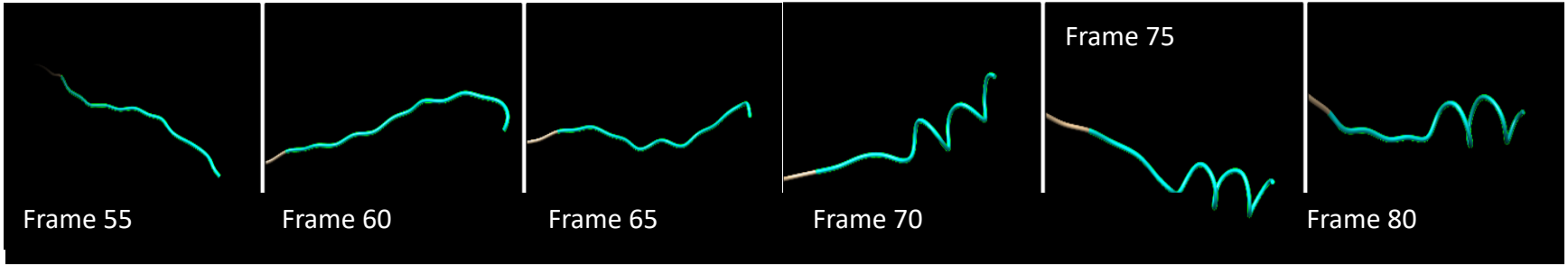
# In Situ Analytics of MD Trajectories

Frames (or snapshots) of an MD trajectory with a stride of 5 steps:




**Collective variables (time 80)**

# In Situ Analytics of MD Trajectories




**Collective variables(55)** **Collective variables(60)** **Collective variables(65)** **Collective variables(70)** **Collective variables(75)** **Collective variables(80)**



Travis Johnston, Buyu Zhang, Adam Liwo, Silvia Crivelli, and Michela Taufer. "***In-Situ Data Analysis and Indexing of Protein Trajectories,***" JCC 2017.

# In Situ Analytics of MD Trajectories

**Collective variables(55)** **Collective variables(60)** **Collective variables(65)** **Collective variables(70)** **Collective variables(75)** **Collective variables(80)**

***Collective variables serve as proxy for structural and conformational changes***



Travis Johnston, Buyu Zhang, Adam Liwo, Silvia Crivelli, and Michela Taufer. "***In-Situ Data Analysis and Indexing of Protein Trajectories,***" JCC 2017.

# The Machine Learning Myth



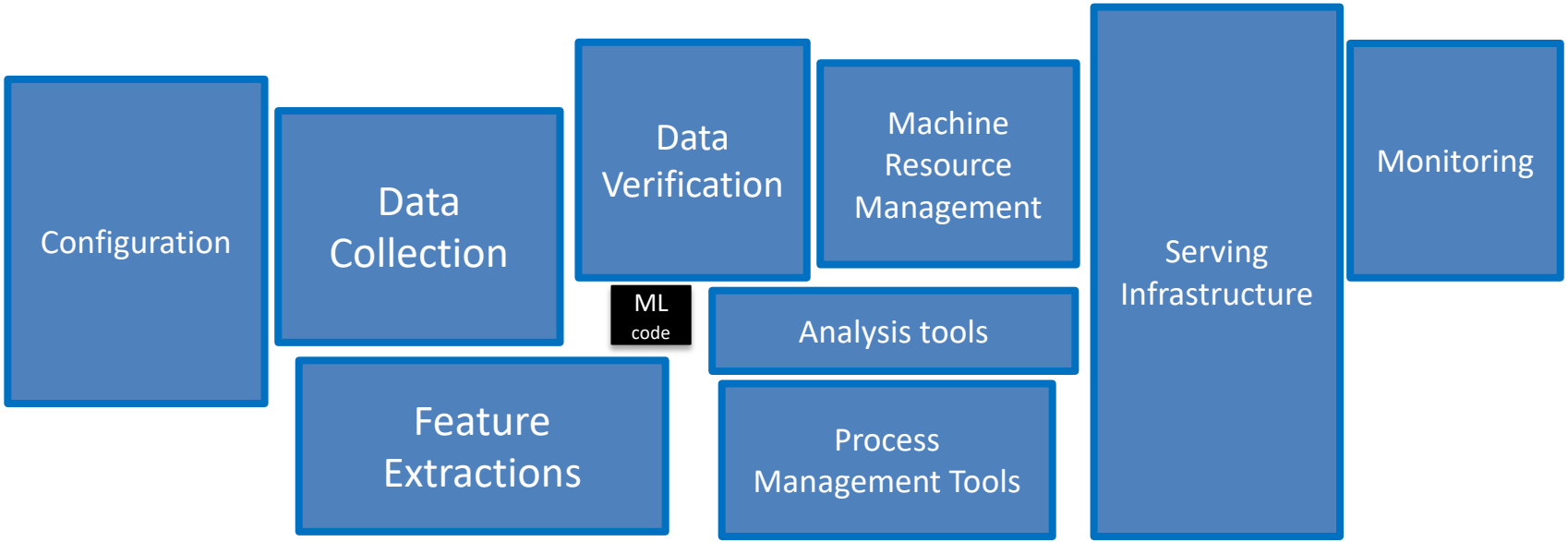
30

# The Machine Learning Myth



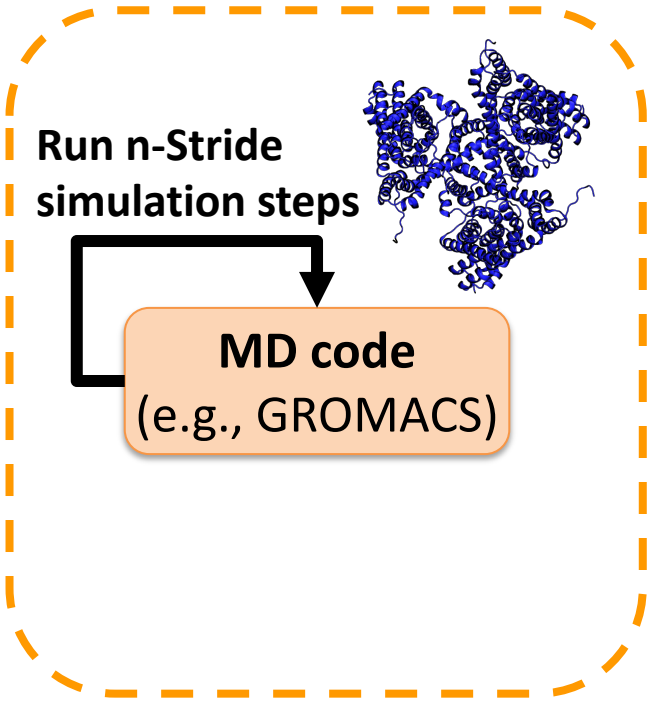
31

# (Missing) Data / Software Ecosystem

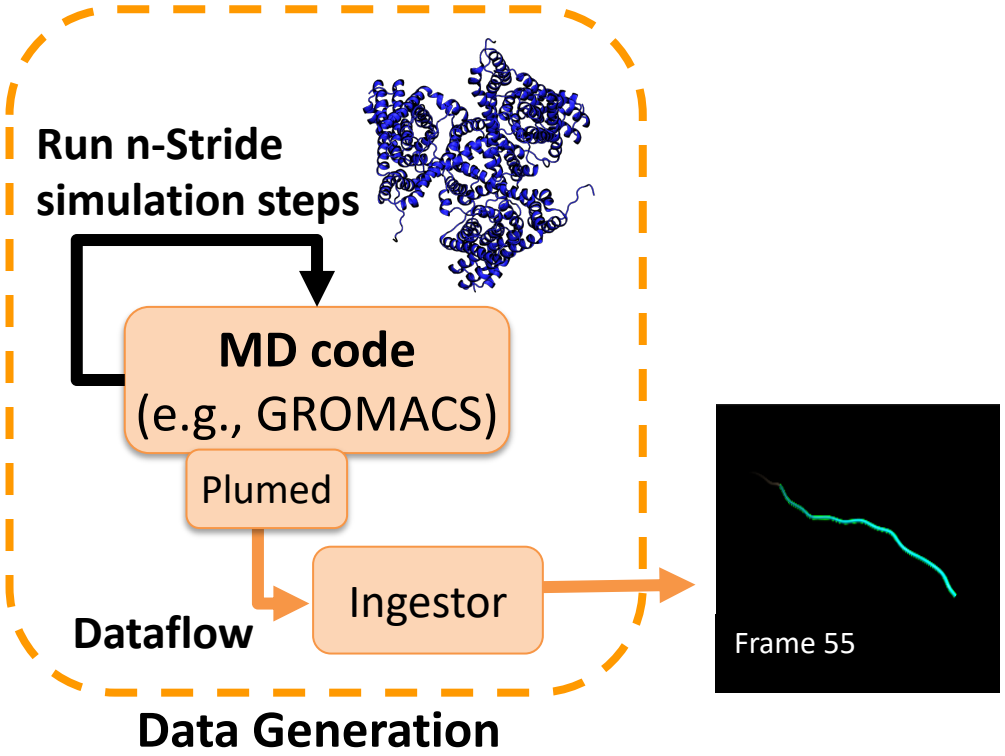



"**Only a small fraction of real-world ML systems is composed of the ML code**" D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips Hidden Technical Debt in Machine Learning Systems

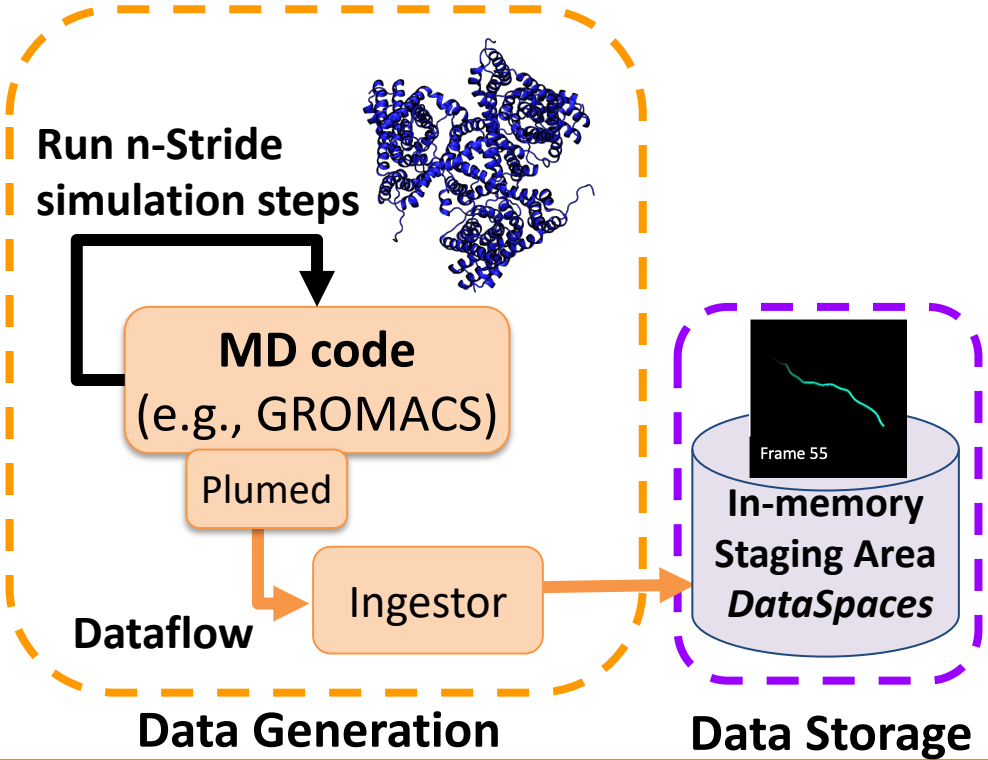
# Building the MD Ecosystem



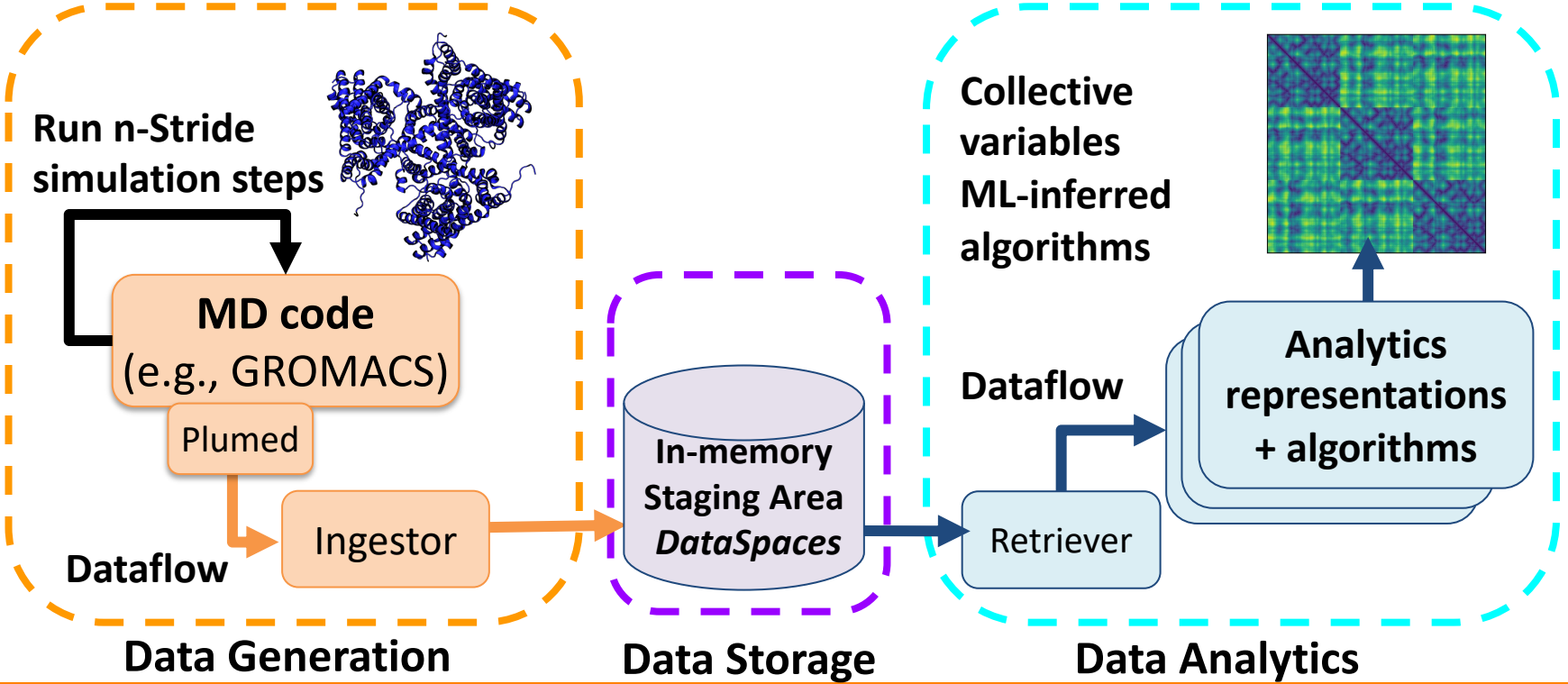
# Building the MD Ecosystem



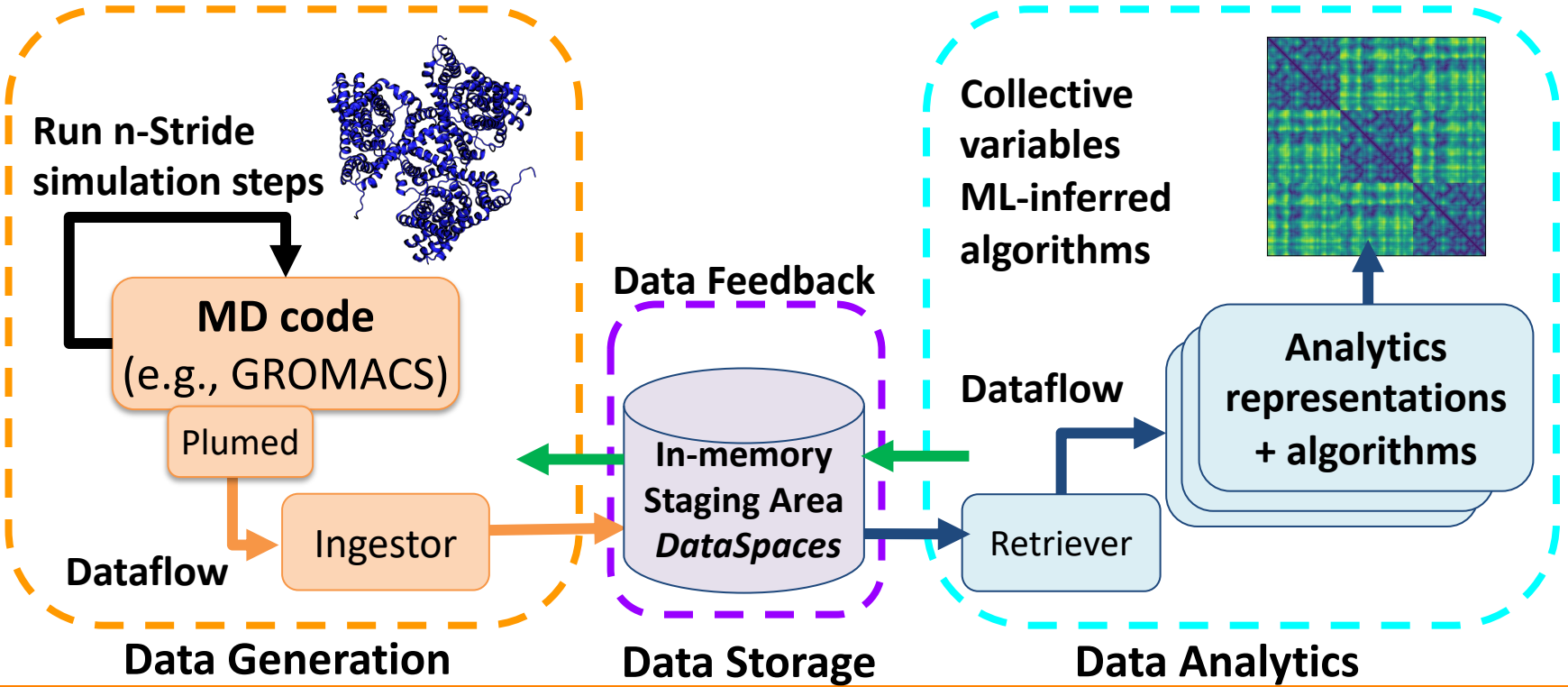
# Building the MD Ecosystem



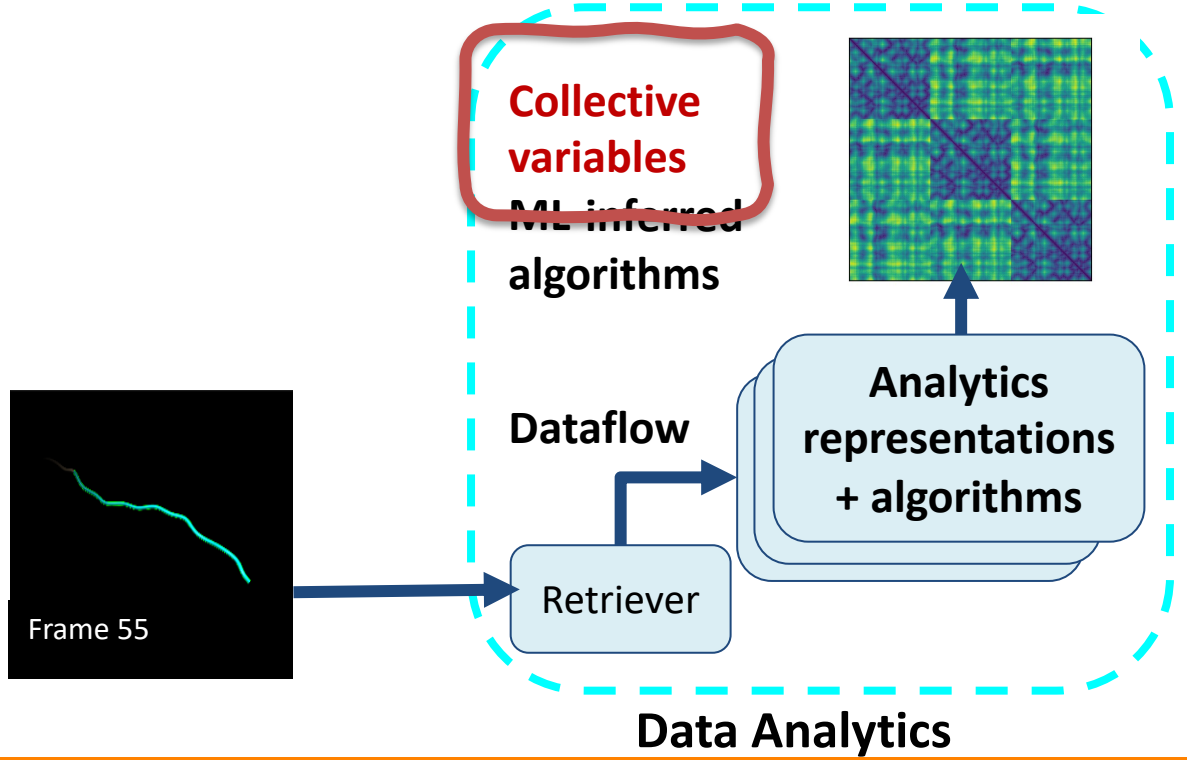
# Building the MD Ecosystem



# Building the MD Ecosystem



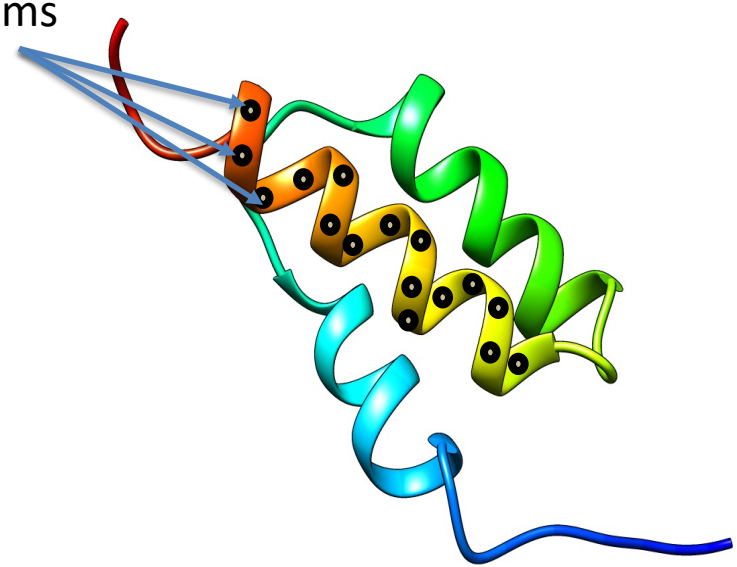
# Capturing Knowledge through Collective Variables



# CV: Largest Eigenvalue of a Secondary Structures

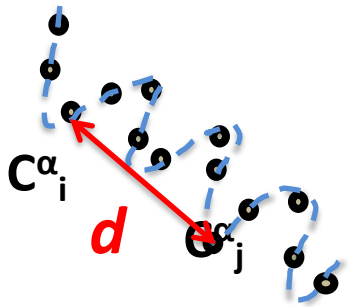
Given the **single, local frame** of a MD job **at time *t***

$C^{\alpha}$ atoms



# CV: Largest Eigenvalue of a Secondary Structures

Measure the distance between $C^{\alpha}_j$ and $C^{\alpha}_i$



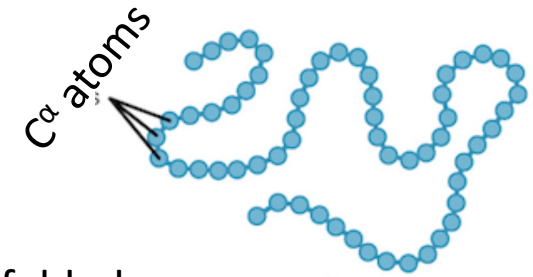

Build the **substructure Euclidean Distance Matrix (D)**

$$D = \begin{array}{c} \\ C^{\alpha}_j \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccc} C^{\alpha}_i & & & & & \end{array} \\ \begin{bmatrix} 0 & x & x & x & x & x \\ x & 0 & d & x & x & x \\ x & d & 0 & x & x & x \\ x & x & x & 0 & x & x \\ x & x & x & x & 0 & x \\ x & x & x & x & x & 0 \end{bmatrix} \end{array}$$

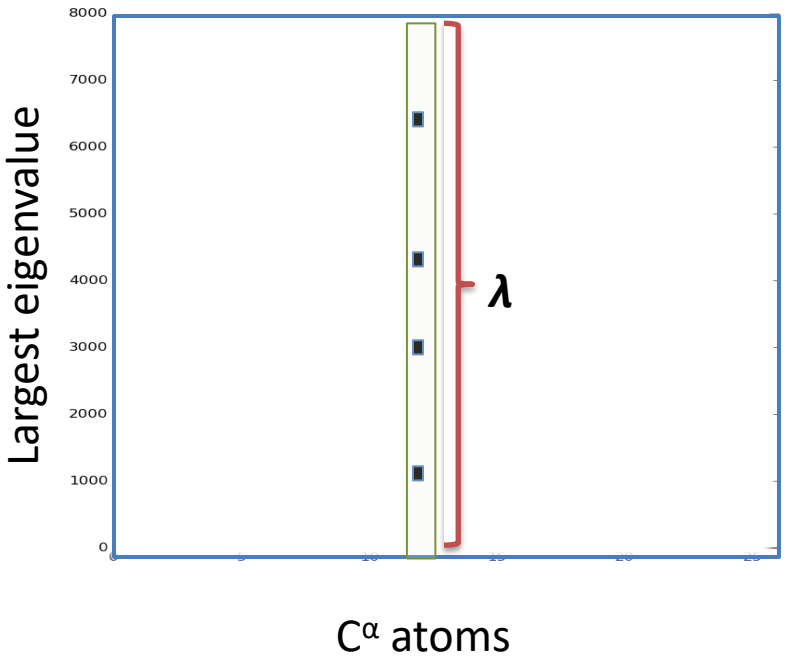
Compute largest eigenvalue → $\lambda_{max}$

# CV: Largest Eigenvalue of a Secondary Structures

- Largest eigenvalues of the Cartesian distance matrix for the $C^{\alpha}$ atoms of a secondary structure at a given time t ($frame_t$)




Unfolded

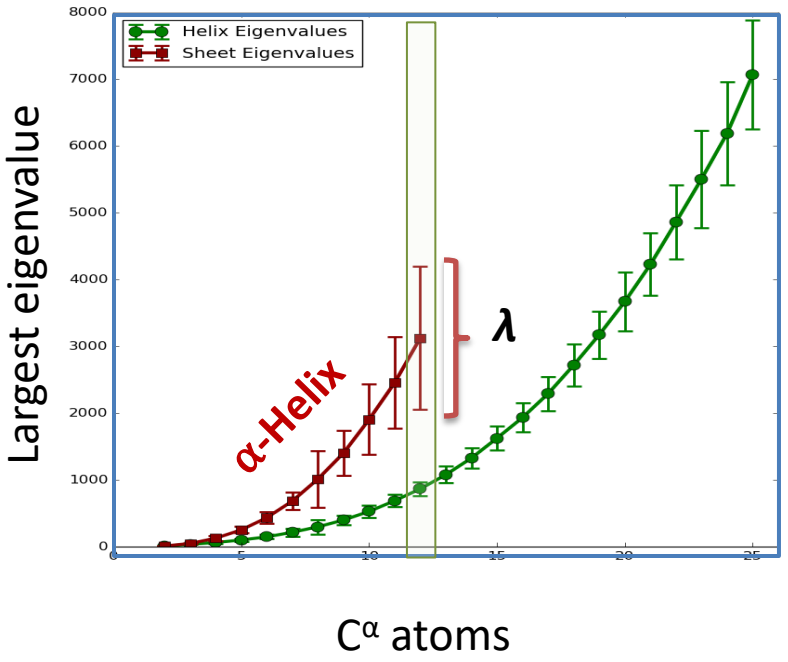


# CV: Largest Eigenvalue of a Secondary Structure

- Largest eigenvalues of the Cartesian distance matrix for the $C^{\alpha}$ atoms of a secondary structure at a given time t ($frame_t$)

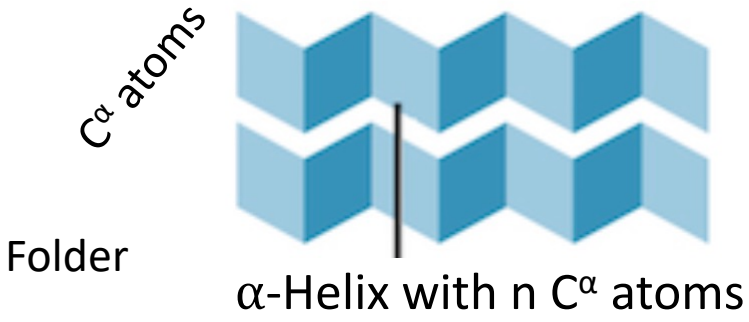

α-Helix with n $C^{\alpha}$ atoms

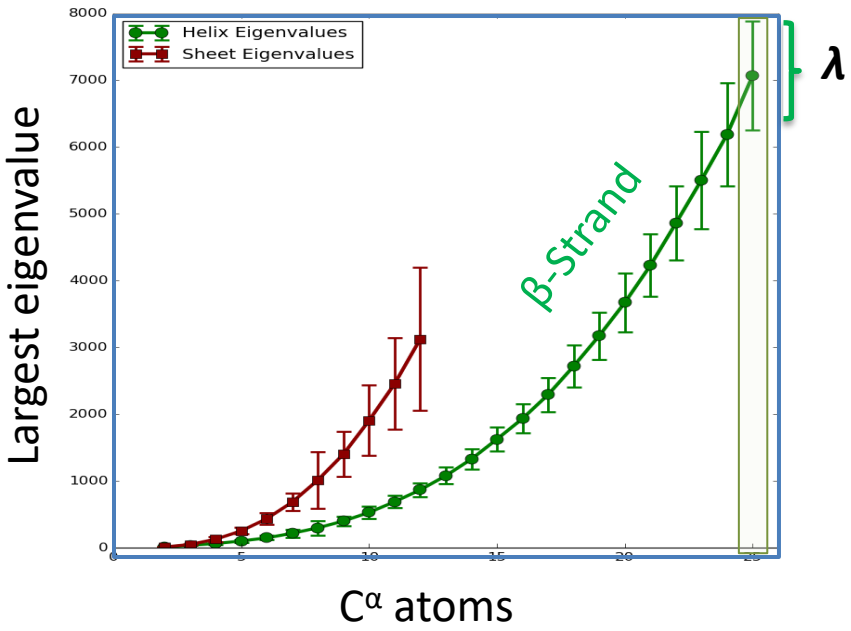
Folder



# CV: Largest Eigenvalue of a Secondary Structure

- Largest eigenvalues of the Cartesian distance matrix for the $C^{\alpha}$ atoms of a secondary structure at a given time t ($frame_t$)

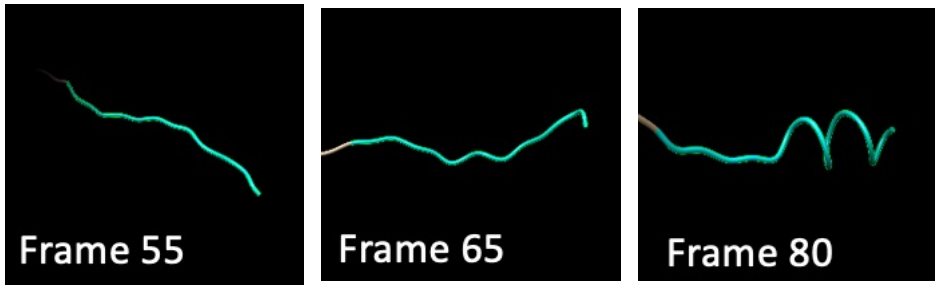


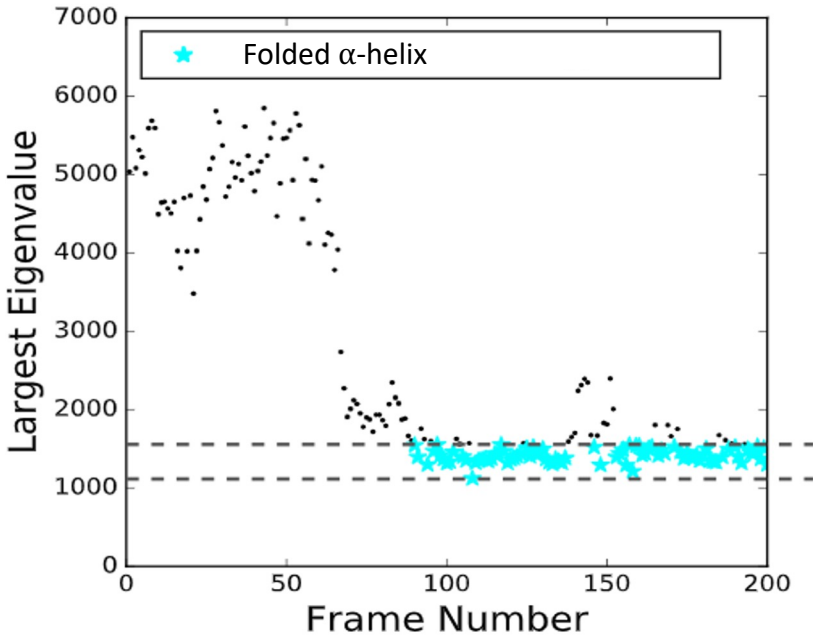
# CV: Largest Eigenvalue of a Secondary Structure

Use case: folding of a simple α-Helix with 10 $C^{\alpha}$ atoms

***Largest eigenvalue → λ***



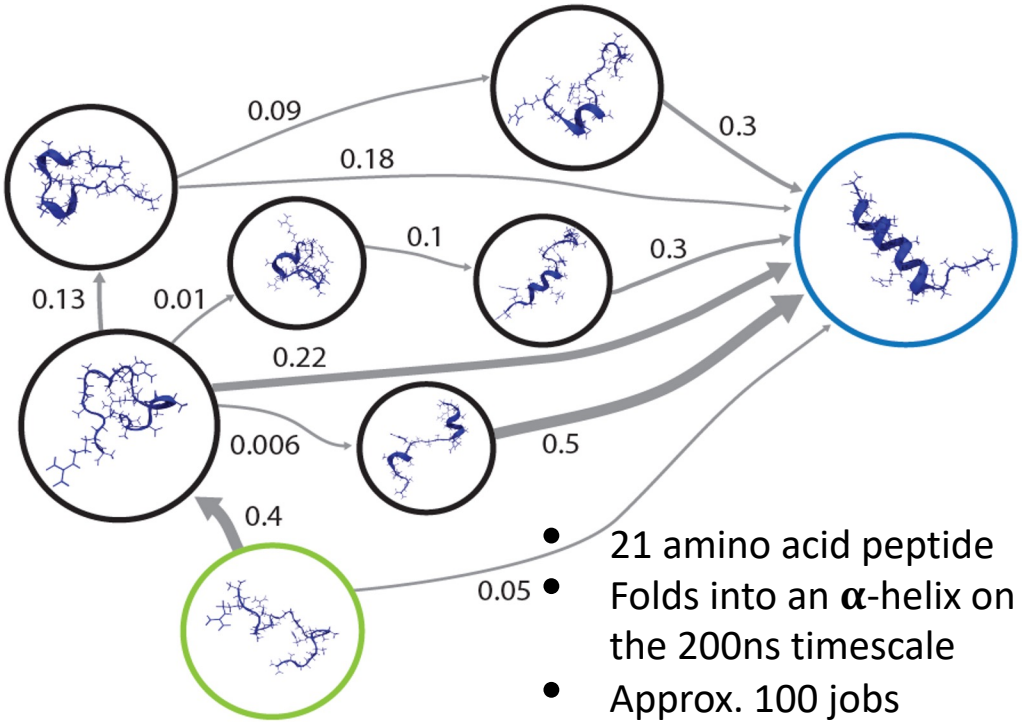

$\lambda_{55}$ $\lambda_{65}$ $\lambda_{80}$






Travis Johnston, Buyu Zhang, Adam Liwo, Silvia Crivelli, and Michela Taufer. "***In-Situ Data Analysis and Indexing of Protein Trajectories,***" JCC 2017.

# Application: Minimizing Folding Time of FS-Peptide

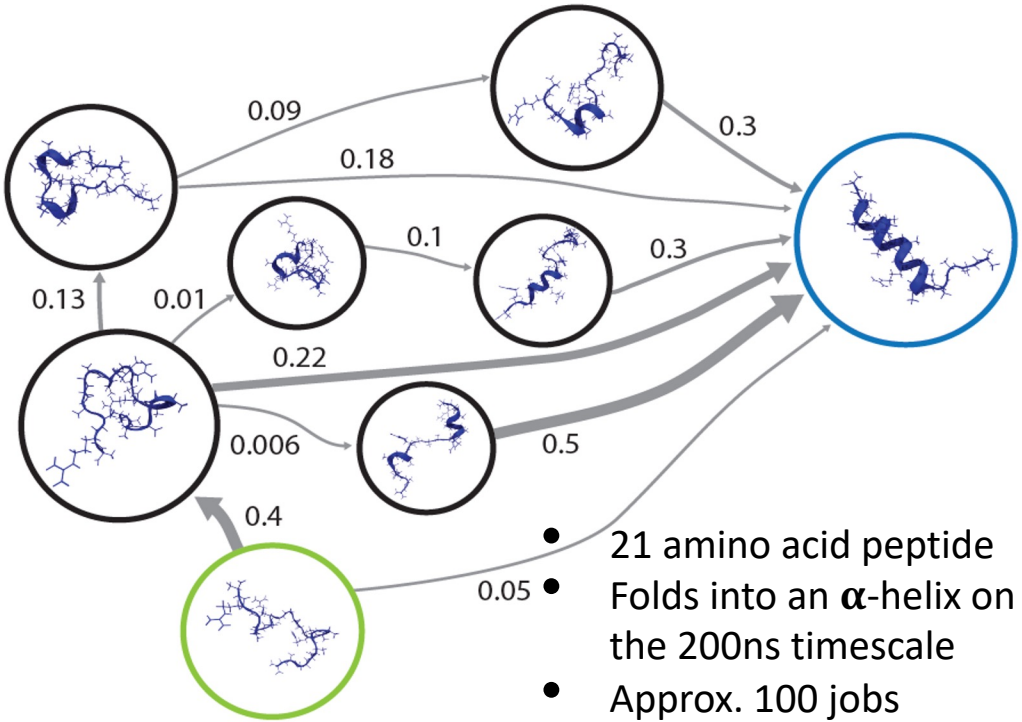



- 21 amino acid peptide
- Folds into an $\alpha$-helix on the 200ns timescale
- Approx. 100 jobs

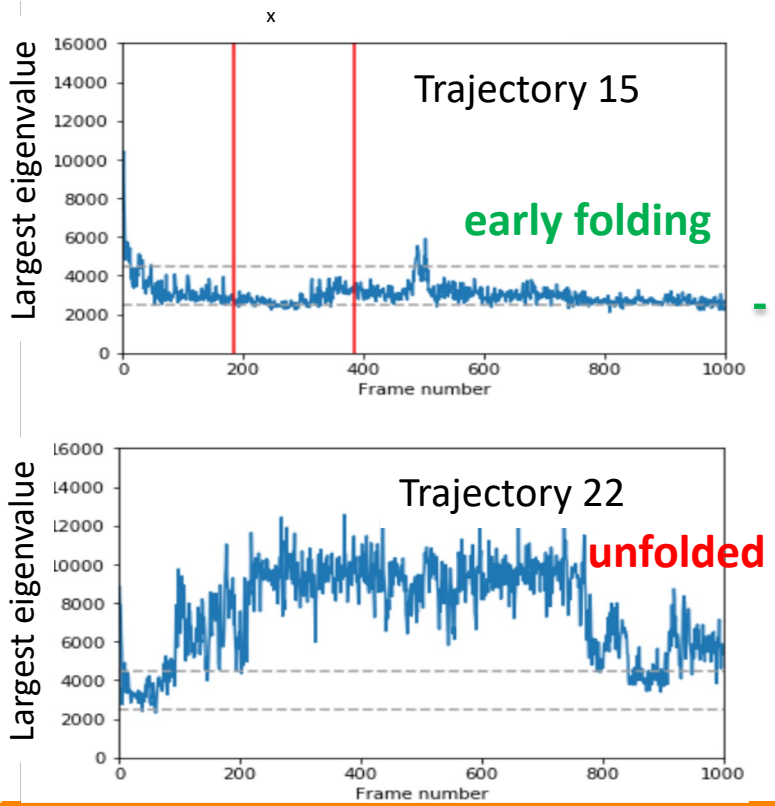


Pronk *et al.*, *Molecular Simulation Workflows as Parallel Algorithms: The Execution Engine of Copernicus, a Distributed High-Performance Computing Platform*, Journal of Chemical Theory and Computation, 2015

# Application: Minimizing Folding Time of FS-Peptide



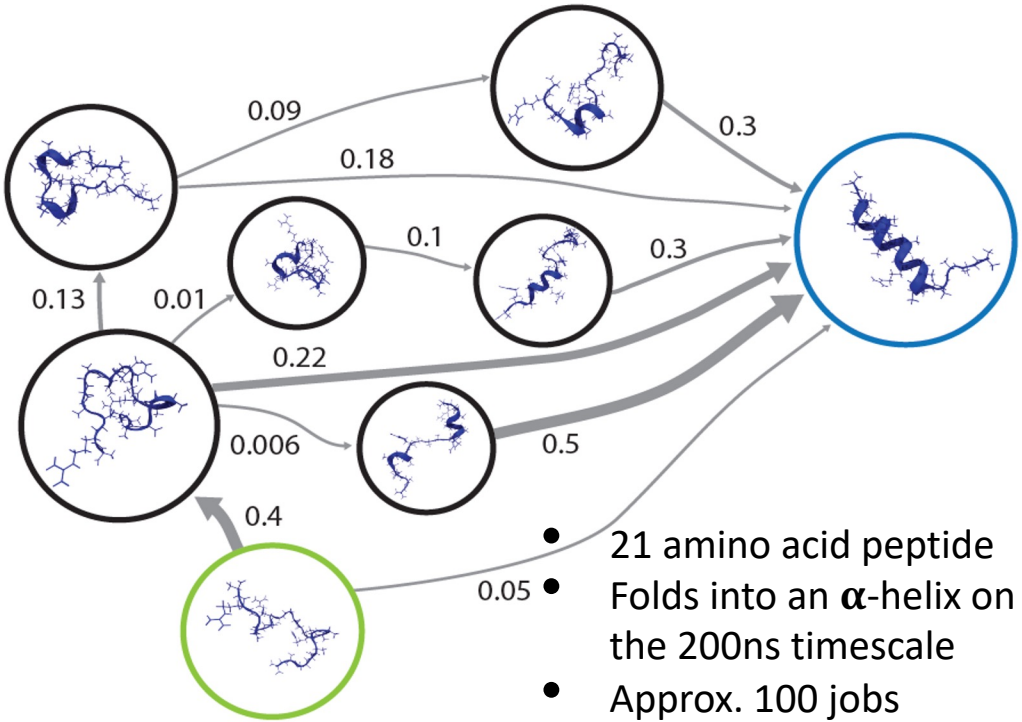

- 21 amino acid peptide
- Folds into an $\alpha$-helix on the 200ns timescale
- Approx. 100 jobs



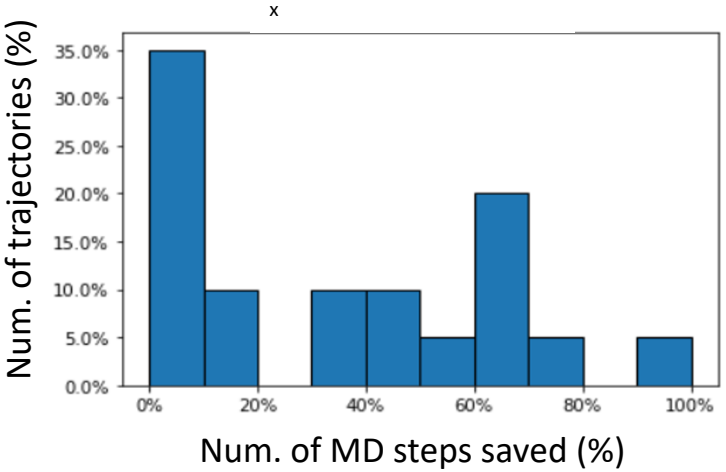



Pronk *et al.*, *Molecular Simulation Workflows as Parallel Algorithms: The Execution Engine of Copernicus, a Distributed High-Performance Computing Platform*, Journal of Chemical Theory and Computation, 2015

# Application: Minimizing Folding Time of FS-Peptide




- 21 amino acid peptide
- Folds into an **α**-helix on the 200ns timescale
- Approx. 100 jobs



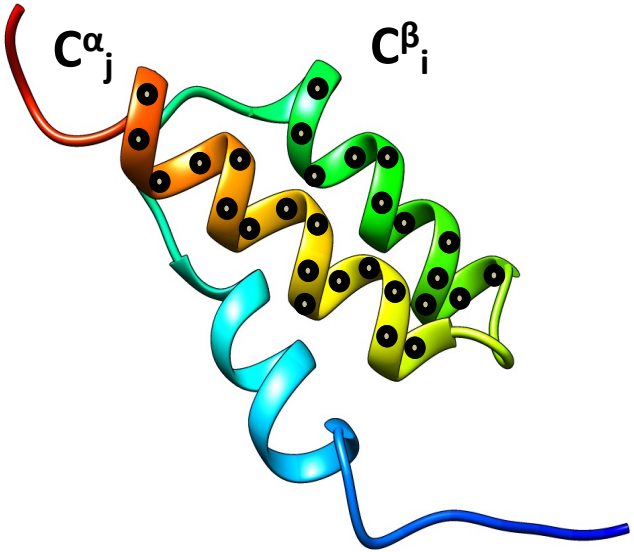

- **70.0% of trajectories terminate early**
- Avg. term. time for all trajectories: 264 ns
- Avg. term. time for preempted trajectories: 207 ns
- Total steps saved: 135,054 (33.76% of ensemble)



Pronk *et al.*, *Molecular Simulation Workflows as Parallel Algorithms: The Execution Engine of Copernicus, a Distributed High-Performance Computing Platform*, Journal of Chemical Theory and Computation, 2015

# CV: Larger Eigenvalues of Ternary Structures

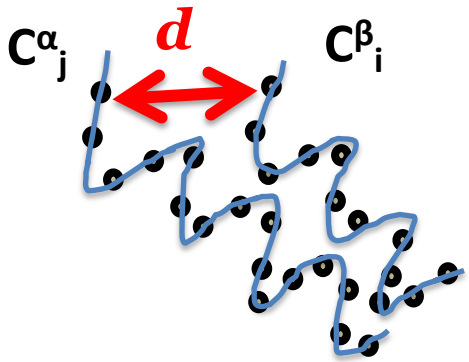
Given the **single, local frame** of a MD job **at time *t***




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.

# CV: Larger Eigenvalues of Ternary Structures

Measure the distance between $C^{\alpha}_{j}$ and $C^{\beta}_{i}$




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

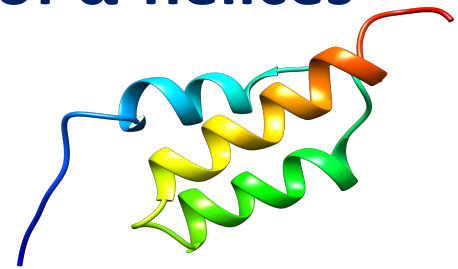
$$D = \begin{array}{c} \\ j \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccc} & & & i & & \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & X & X & X \\ 0 & 0 & 0 & d & X & X \\ 0 & 0 & 0 & X & X & X \\ X & d & X & 0 & 0 & 0 \\ X & X & X & 0 & 0 & 0 \\ X & X & X & 0 & 0 & 0 \end{bmatrix} \end{array}$$

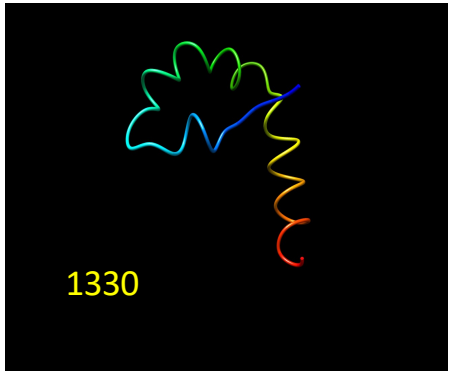
Compute largest eigenvalue → $\lambda_{max}$

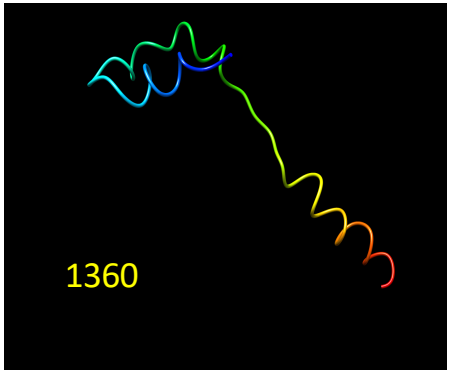

T. Johnston, B. Zhang, A. Liwo, S. Crivelli, and M. Taufer. In-Situ Data Analytics and Indexing of Protein Trajectories. *Journal of Computational Chemistry (JCC),* 38(16):1419-1430, 2017.

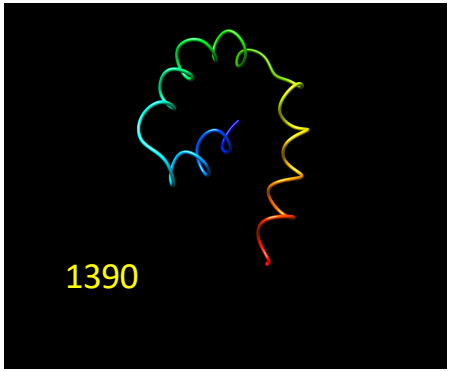
# Application: Capturing Movement of α-helices

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





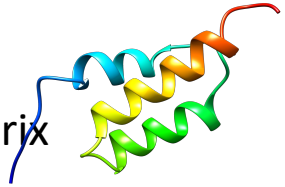

T. Johnston, B. Zhang, A. Liwo, S. Crivelli, and M. Taufer. In-Situ Data Analytics and Indexing of Protein Trajectories. *Journal of Computational Chemistry (JCC),* 38(16):1419-1430, 2017.

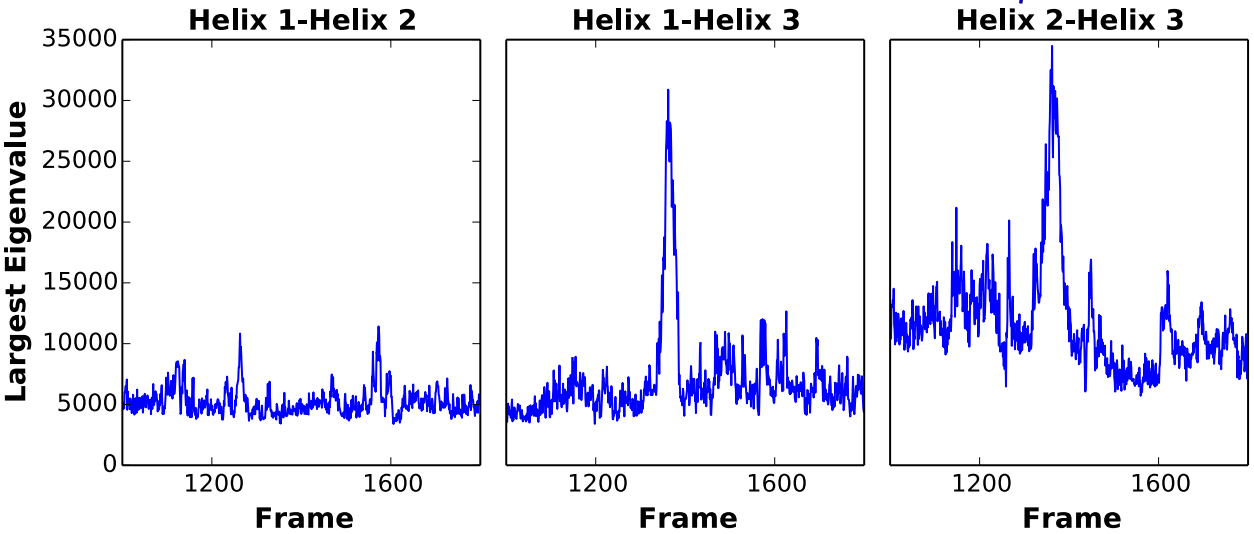


50

# Application: Capturing Movement of α-helices

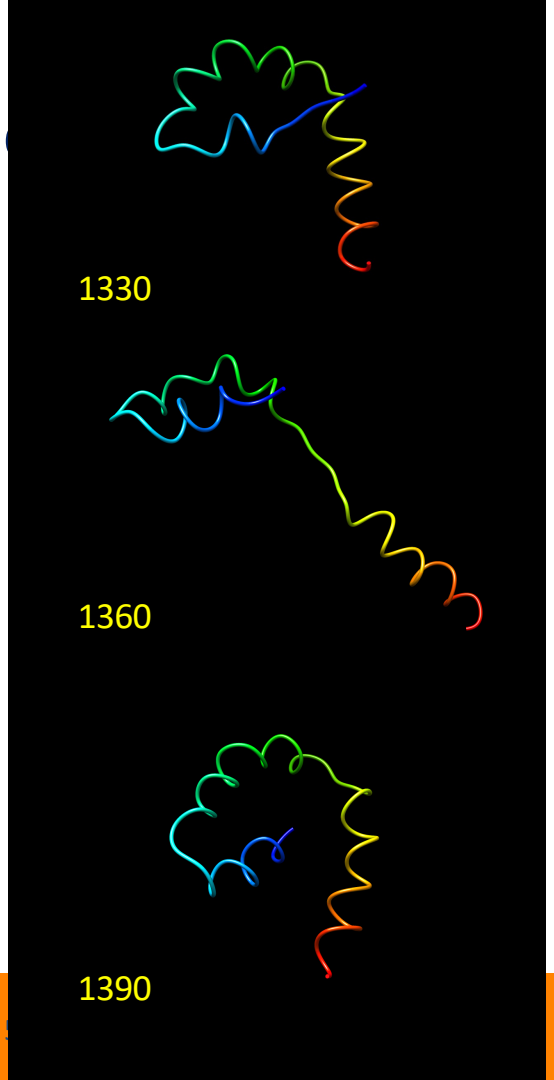
Monitor largest eigenvalue of bipartite distance matrix



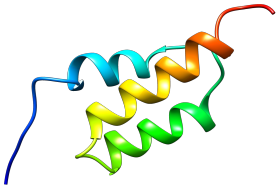


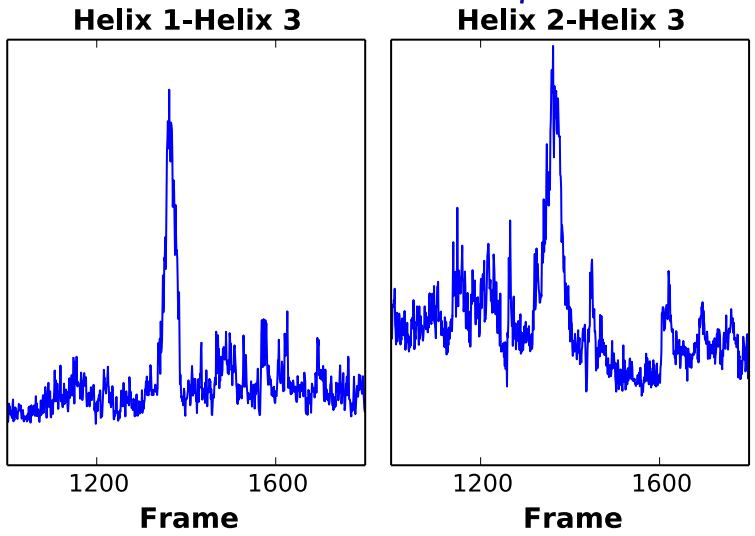

First and second α-helices appear stable; third helix moves

# Appli




# Movement of α-helices



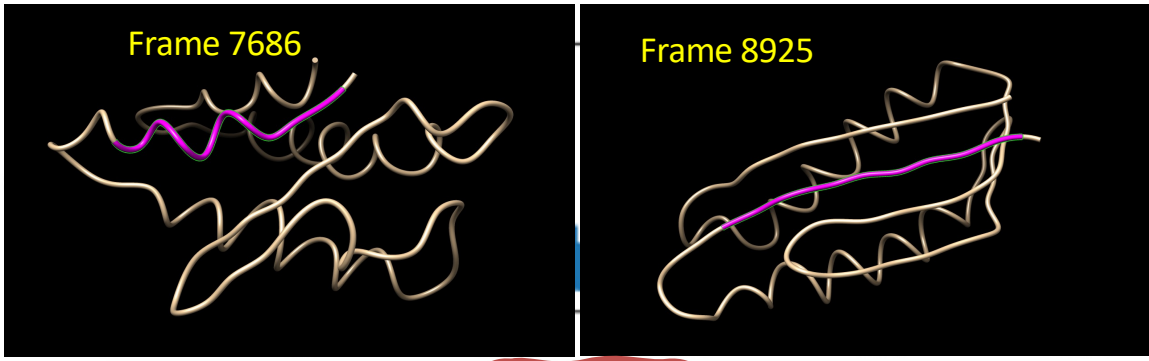


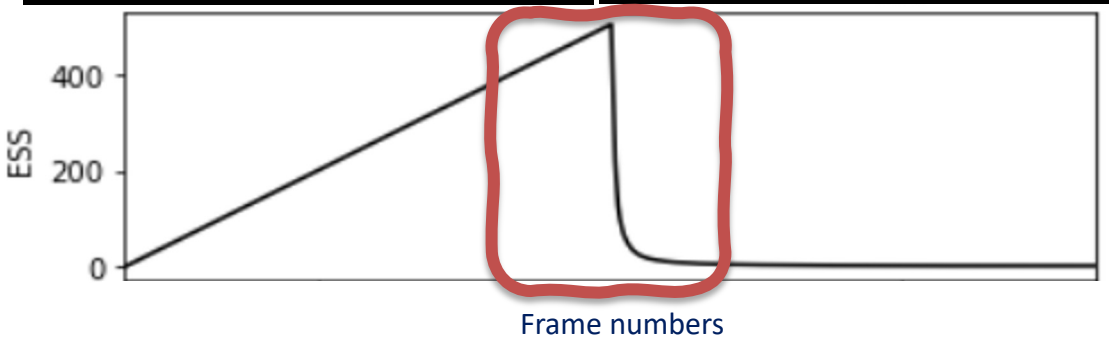
# CV: Effective Sample Size (ESS)

## Effective Sample Size (ESS)

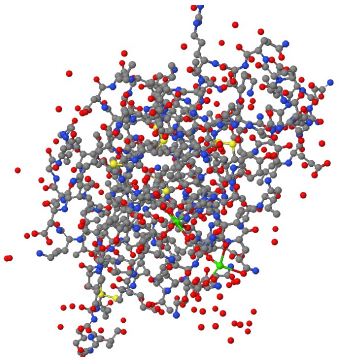
- Metric capturing structural changes over a sample of *n* observations in the trajectory (n << total num. steps)
- If observations are autocorrelated:
  → no major changes in sequence of frames
  → else rare event has occurred

2MQ8 protein: two of the four helices transform into strands



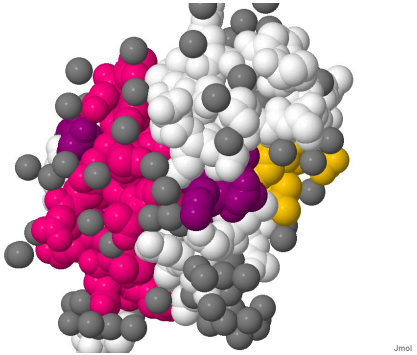



# CV: Graphic Encoding



3D Cartesian representation

Multi-fold representation

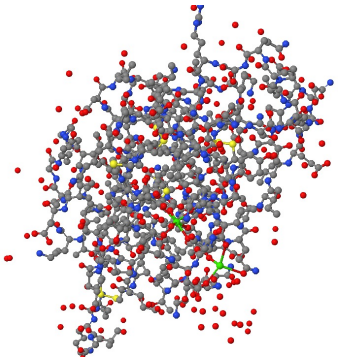


Surface representation



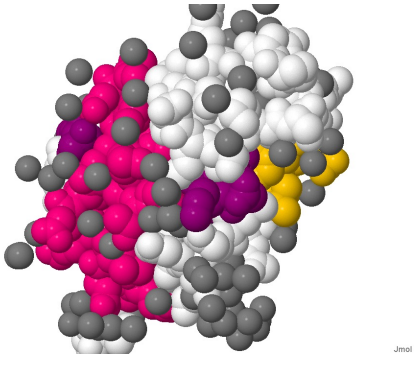
T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer.
***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

# CV: Graphic Encoding

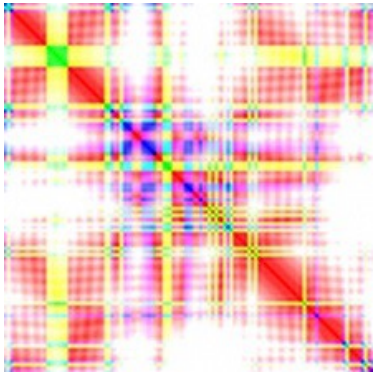


3D Cartesian representation

Multi-fold representation



Surface representation

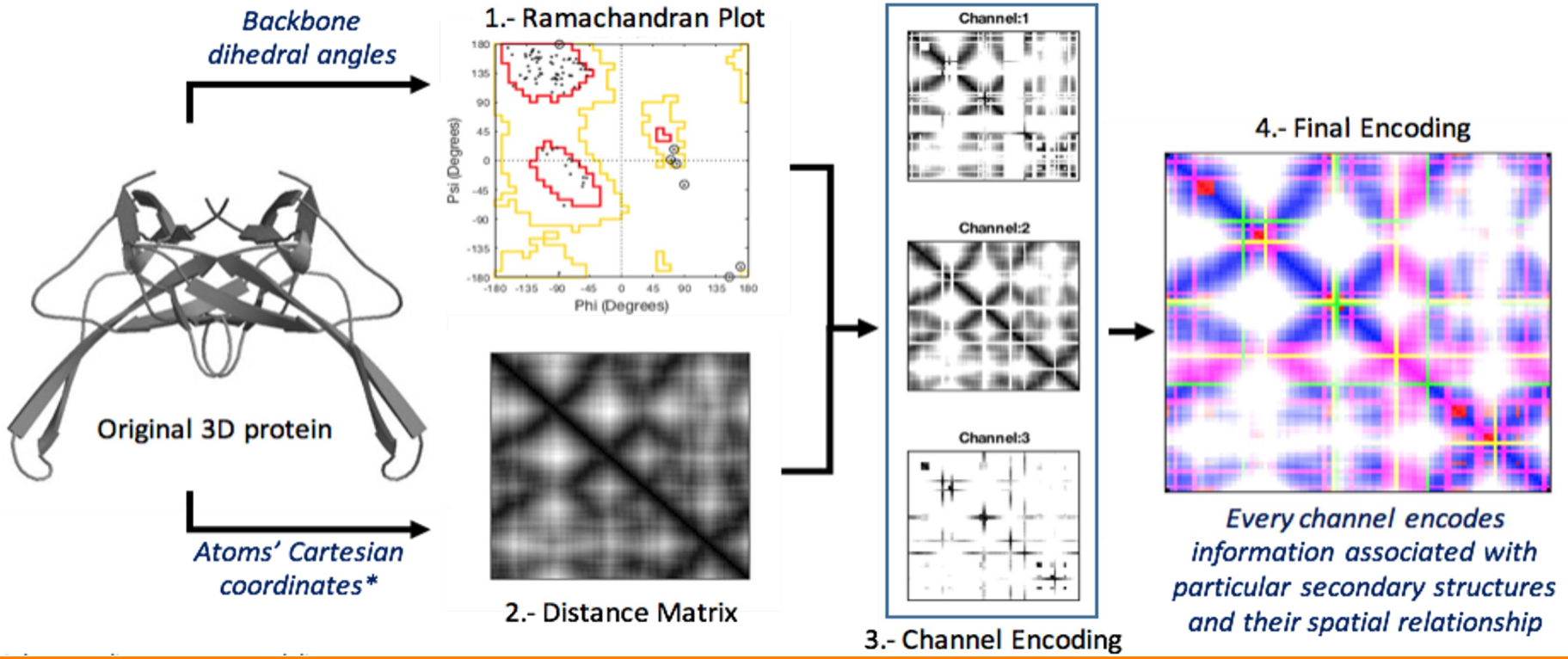


Graphic encoding



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.*

# CV: Graphic Encoding Algorithm

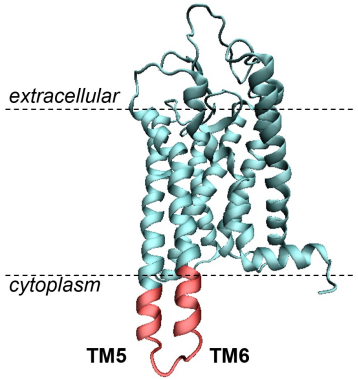


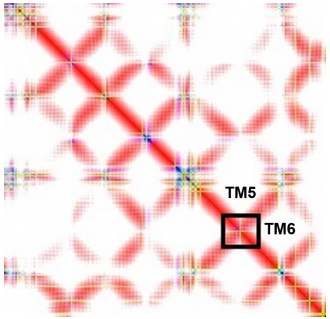



T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer.
***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

# Application: Capturing Unraveling of Loop TM6

Protein: Opsin




Opsins are proteins involved in the vision process, supporting the conversion of light into electrochemical signals

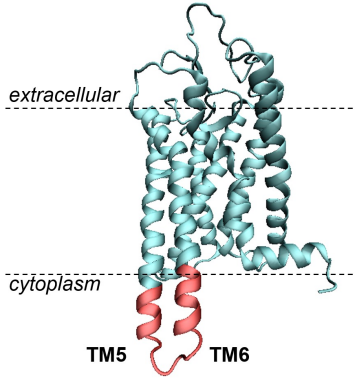
Use graphical representations as indicators of what areas of a protein are relevant for particular conformational changes

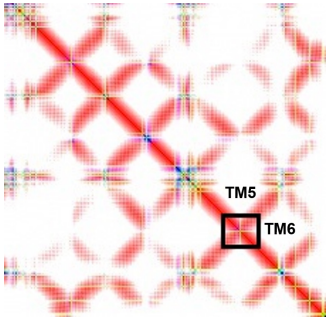


T. Estrada,et al. **A Graphic Encoding Method for Quantitative Classification of Protein Structure and Representation of Conformational Changes**. IEEE TBCB, 2019

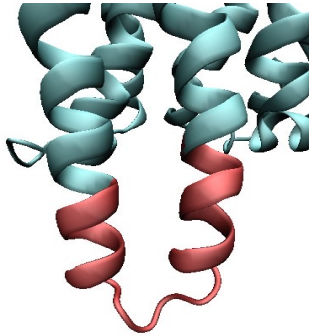
# Application: Capturing Unraveling of Loop TM6

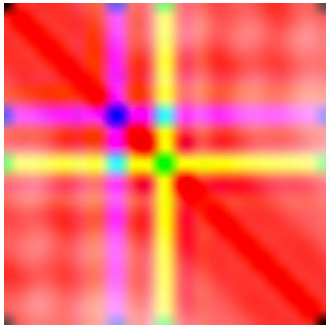
Protein: Opsin



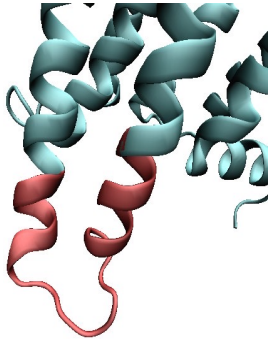



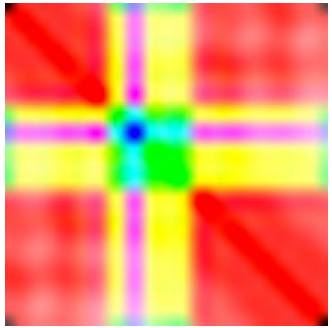

Frame 50



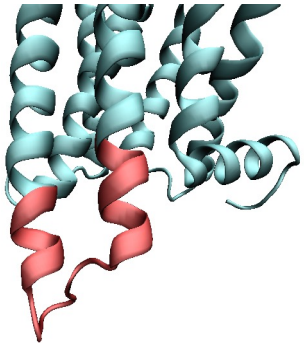


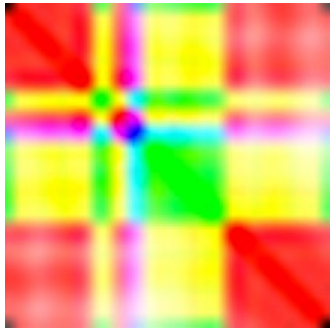
Frame 1500





Frame 1950



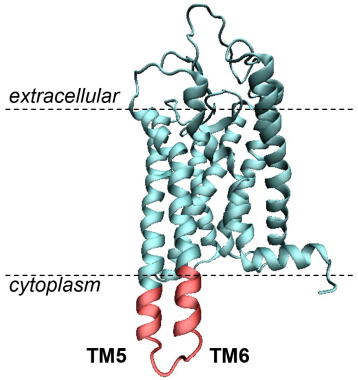


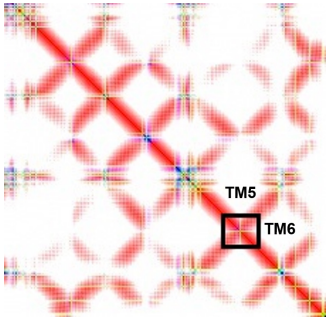


T. Estrada,et al. **A Graphic Encoding Method for Quantitative Classification of Protein Structure and Representation of Conformational Changes**. IEEE TBCB, 2019

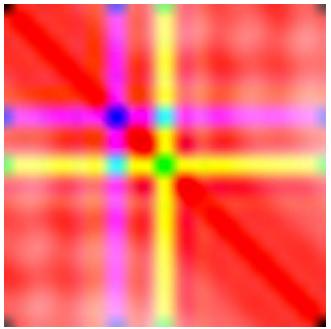
# Application: Capturing Unraveling of Loop TM6

Protein: Opsin

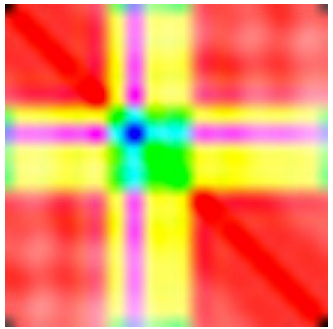


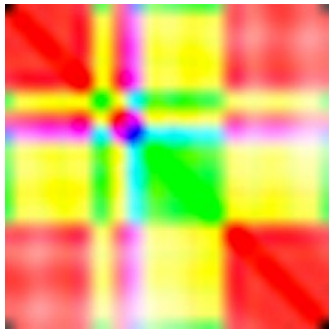

Frame 50



Frame 1500



Frame 1950

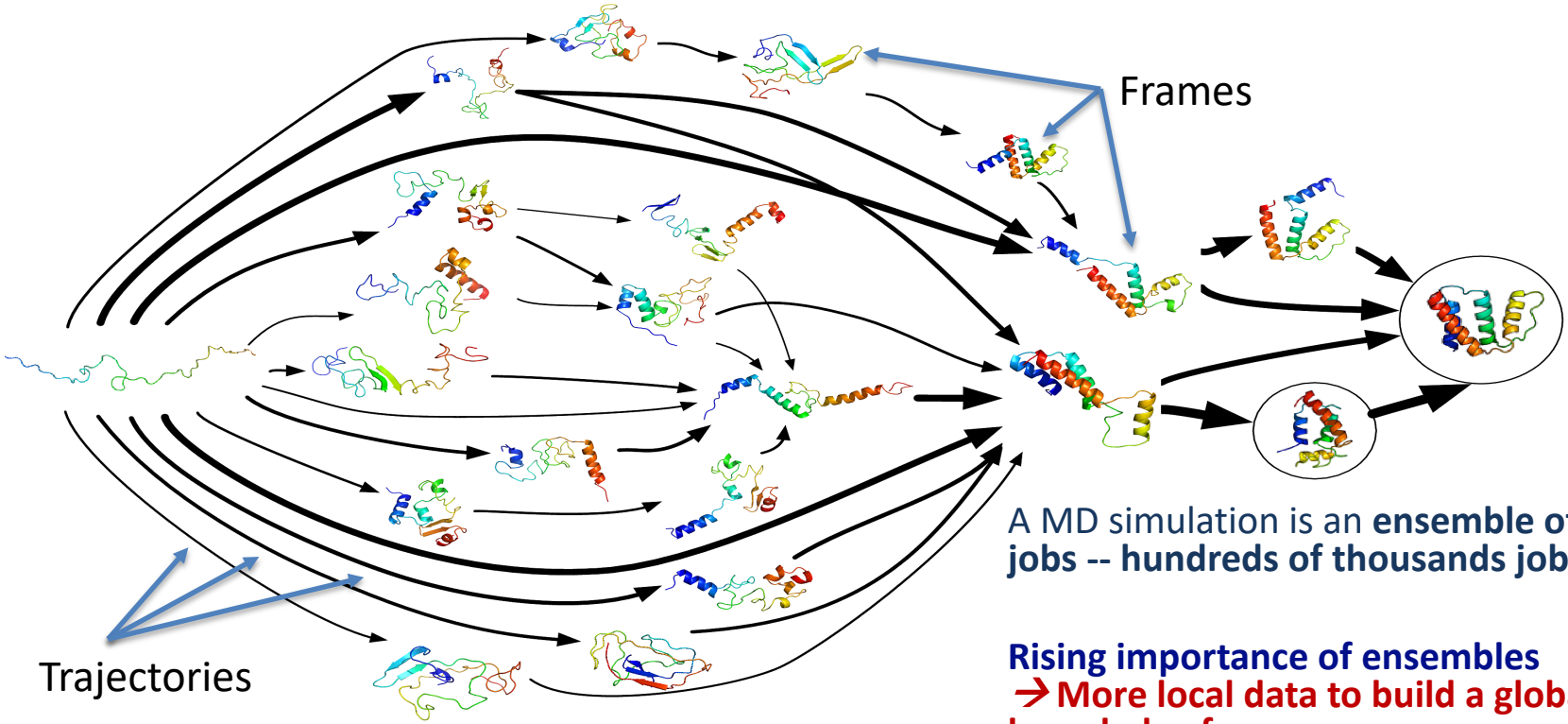


Example of **transfer learning**: we can take an existing neural network such as **GEM-net** that has been trained on some dataset and re-purposing it **for the analysis of trajectories**



T. Estrada,et al. **A Graphic Encoding Method for Quantitative Classification of Protein Structure and Representation of Conformational Changes**. IEEE TBCB, 2019

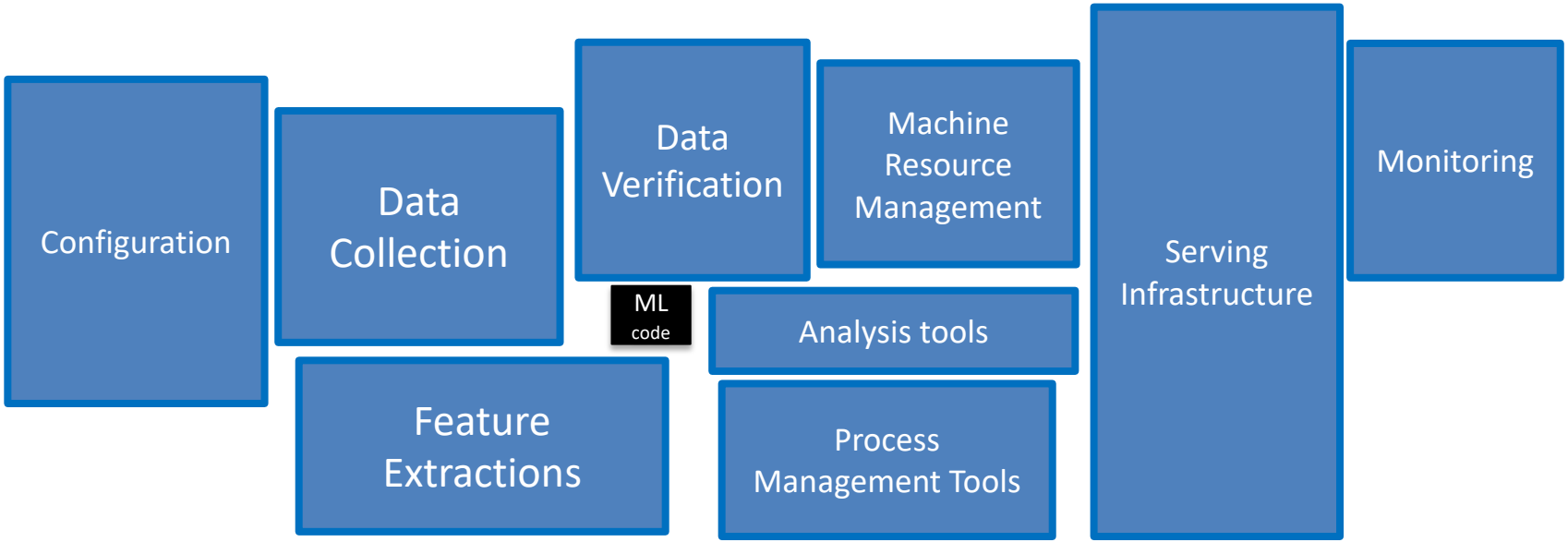
# MD: an Ensemble of Distributed CVs



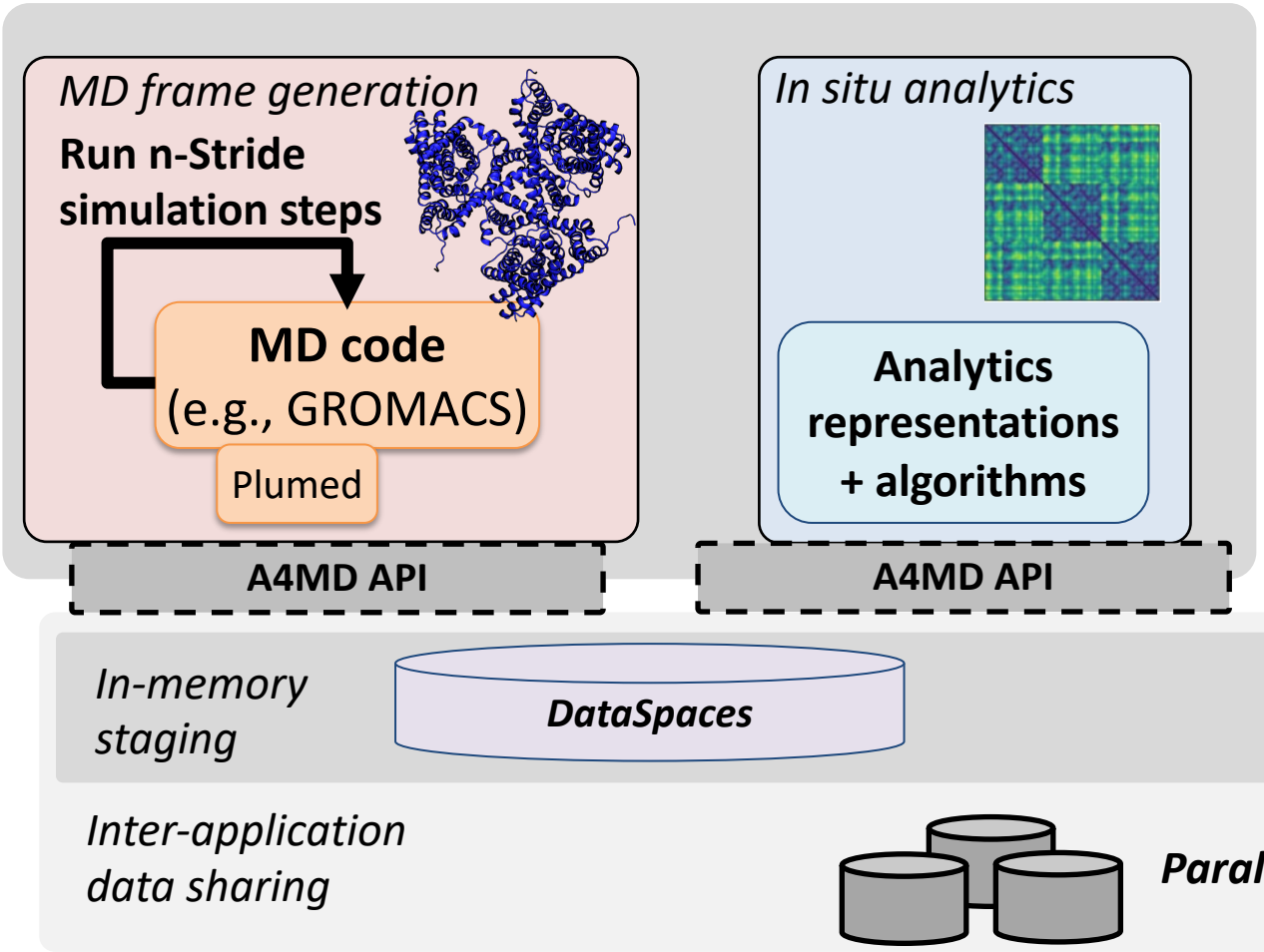

A MD simulation is an **ensemble of MD jobs -- hundreds of thousands jobs**

**Rising importance of ensembles**
**→More local data to build a global knowledge from**

# (Missing) Data / Software Ecosystem

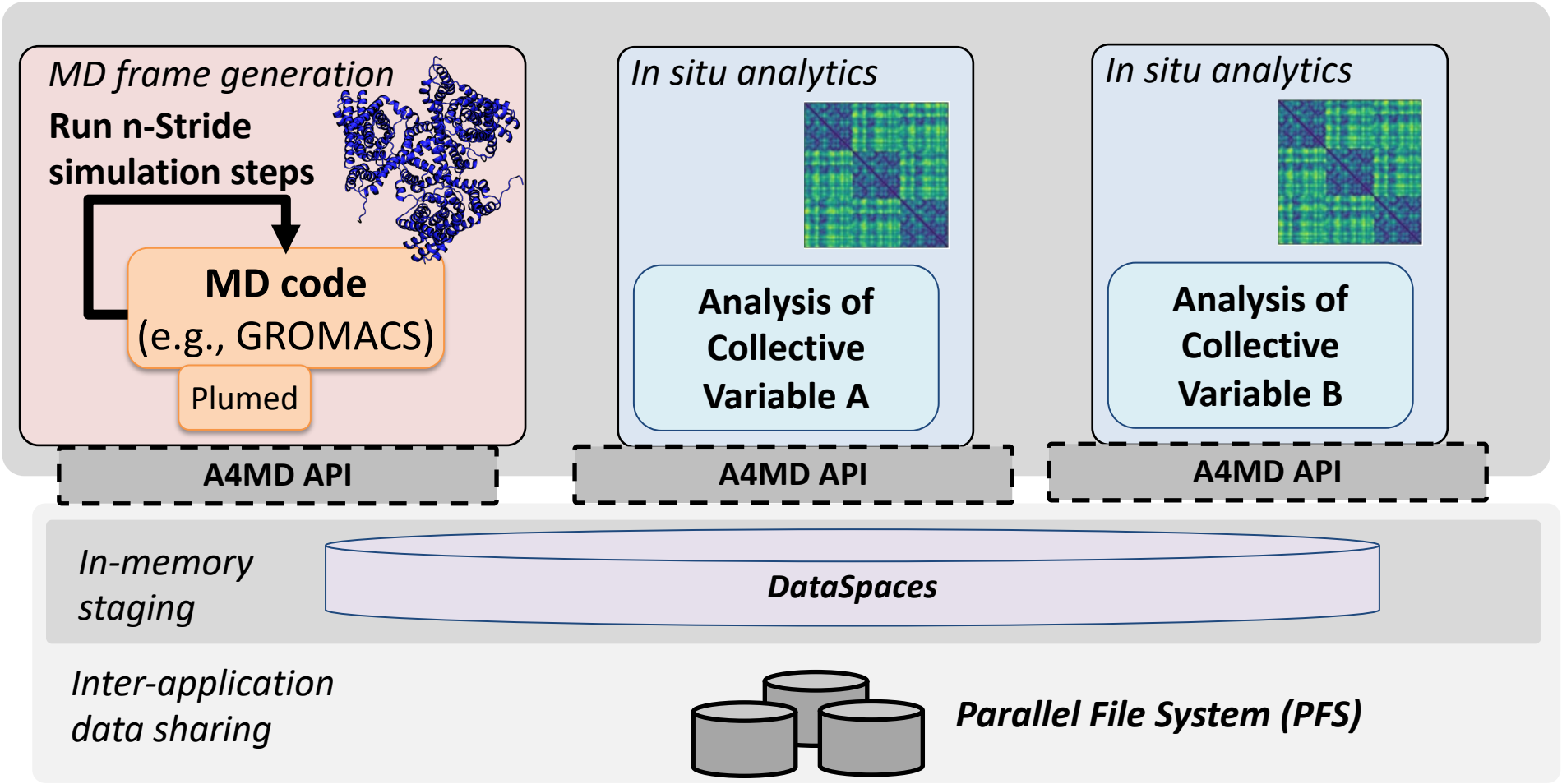


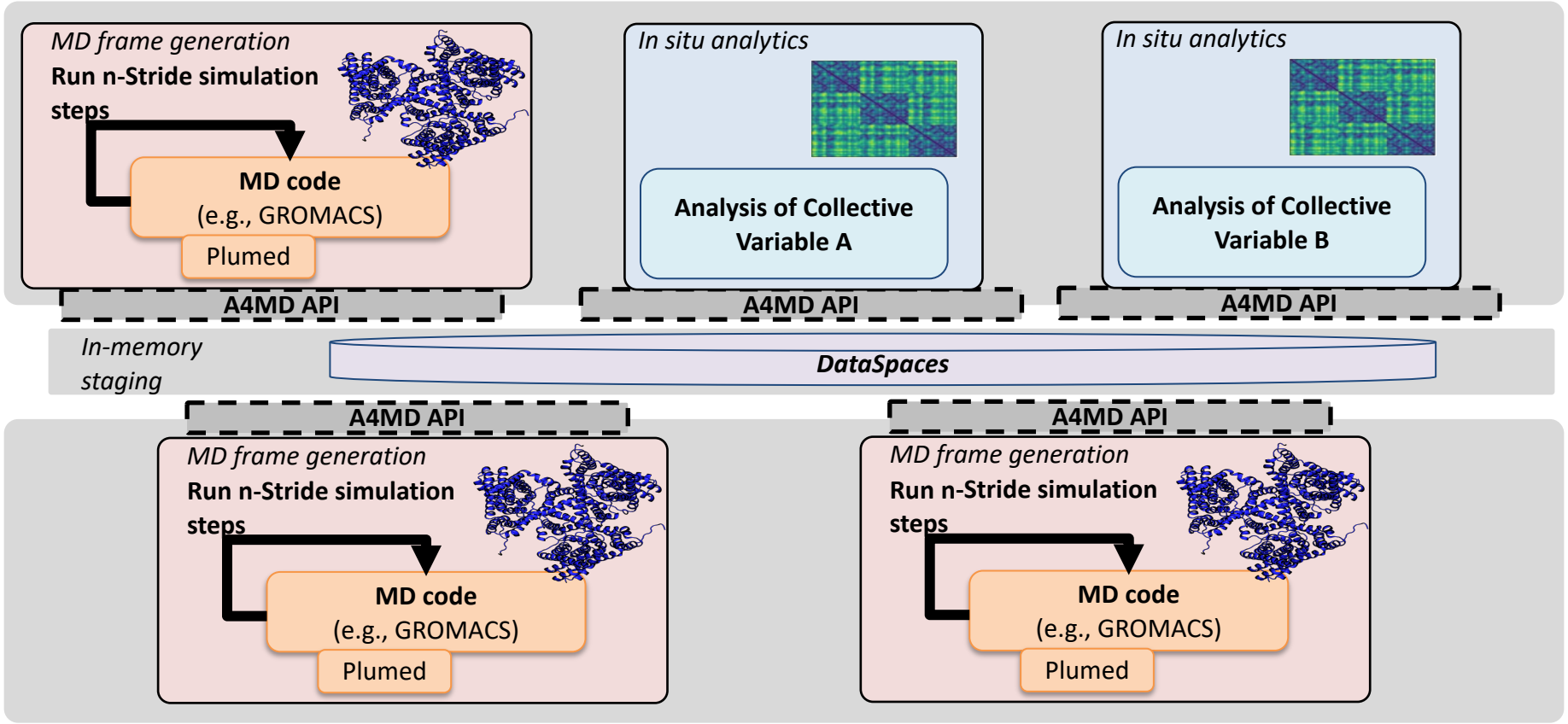

"**Only a small fraction of real-world ML systems is composed of the ML code**" D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips Hidden Technical Debt in Machine Learning Systems




analytics for molecular dynamics

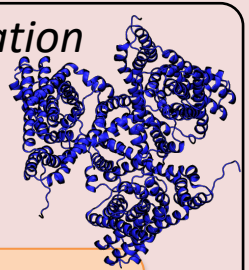
https://analytics4md.org/


MD frame generation
Run n-Stride simulation steps
MD code (e.g., GROMACS)
Plumed
A4MD API
In situ analytics
Analysis of Collective Variable A
A4MD API
In situ analytics
Analysis of Collective Variable B
A4MD API
In-memory staging
DataSpaces
Inter-application data sharing
Parallel File System (PFS)



*MD frame generation*

**Run n-Stride simulation steps**

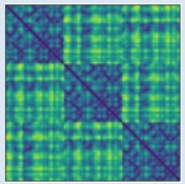


**MD code** (e.g., GROMACS)

Plumed

**A4MD API**

*In situ analytics*



**Analytics representations + algorithms**

**A4MD API**

*Predictions / Steering*

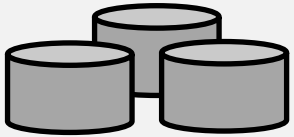
- Understand and annotate dynamics in MD trajectories
- Enhance adaptive sampling in MD trajectories
- Enable on-the-fly tuning of MD workflows (i.e., stop, start, and fork MD jobs)

**A4MD API**

*In-memory staging*

***DataSpaces***

*Inter-application data sharing*



***PFS***

https://analytics4md.org/

analytics for molecular dynamics

# Mission of the *analytics4md* Project

- Formulate **new in situ methods to trace molecular events** by locally reducing knowledge of high-dimensional molecular organization into collective variables (CVs)
- Design **new data representations and extend unsupervised machine learning techniques** to build an explicit global organization of structural and temporal CVs
- Develop **new in situ workflow management that integrates simulation and analytics** into complex MD workflows
- Train a diverse community in the use of the **new MD workflows**


https://analytics4md.org/

NSF IIS:
1741057, 1740990, 1741040, 1841758

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®