ScaDS.AI
DRESDEN LEIPZIG

CENTER FOR SCALABLE DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE

MDannotator – Visual Labeling of MD Simulations for Classification

Daniel Wiegreffe

Gefördert durch:

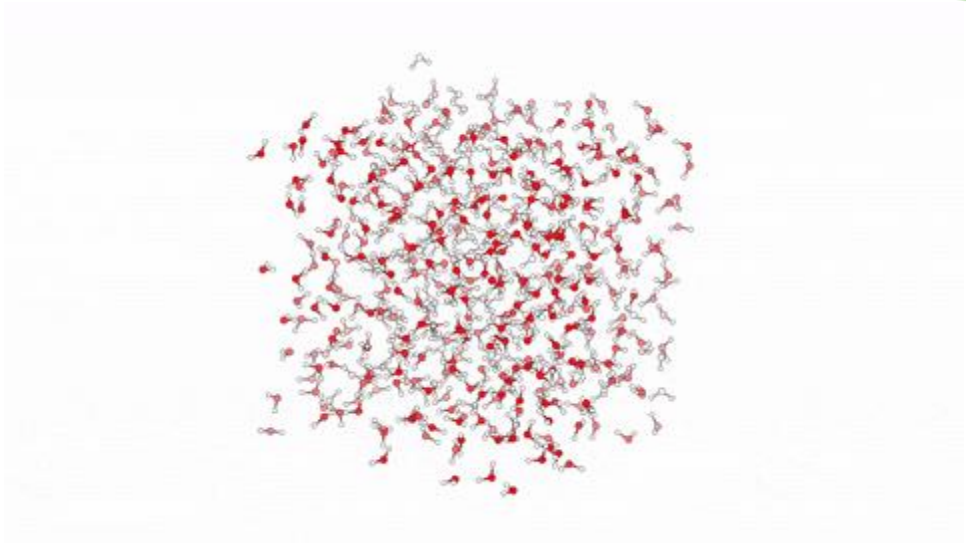
Bundesministerium für Forschung, Technologie und Raumfahrt

SACHSEN

Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

MD Simulations?

- Structure prediction of molecules is cool, but their dynamic behavior can be really interesting:
    - Interaction with other molecules
    - Stable behavior in specific environments
    - Refolding events
- Can be very time-consuming
- Computationally still massive



MD Simulations?

Simulations can contain millions of frames.

- How to analyze and explore?

- **By Hand:**
  - Using poor PhD Students to force them to watch hours of trajectory data
  - Finding the needle in a hay stack

- **Metrics and Algorithms:**
  - Using poor PhD Students to pick the right metric
  - Might miss interesting spots

MD Simulations?

Simulations can contain millions of frames.

- How to analyze and explore?

- **By Hand:**
  - Using poor PhD Students to force them to watch hours of trajectory data
  - Finding the needle in a hay stack

- **Metrics and Algorithms:**
  - Using poor PhD Students to pick the right metric/algorithm
  - Might miss interesting spots

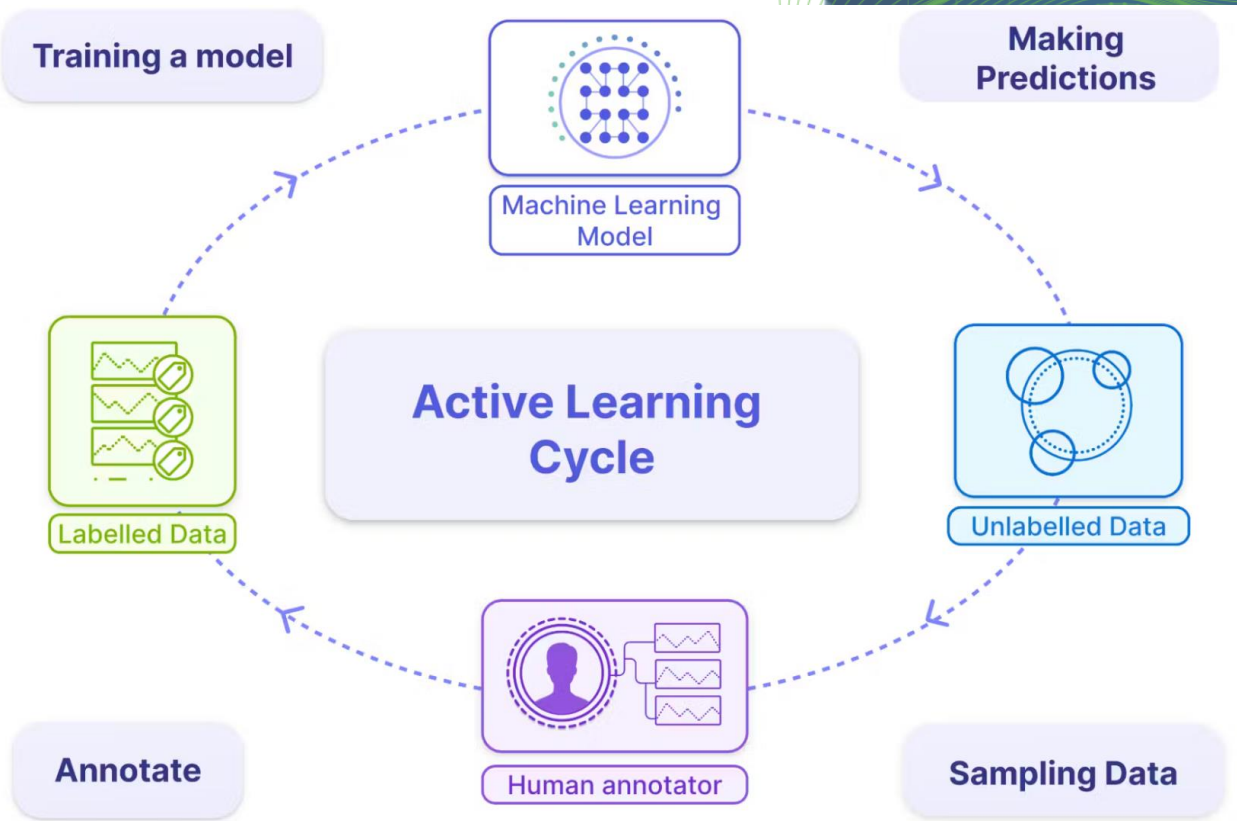
Might miss an interesting spot in the data due to boredom!

→ Can we use AI for making PhD students happy?

Data Sources?

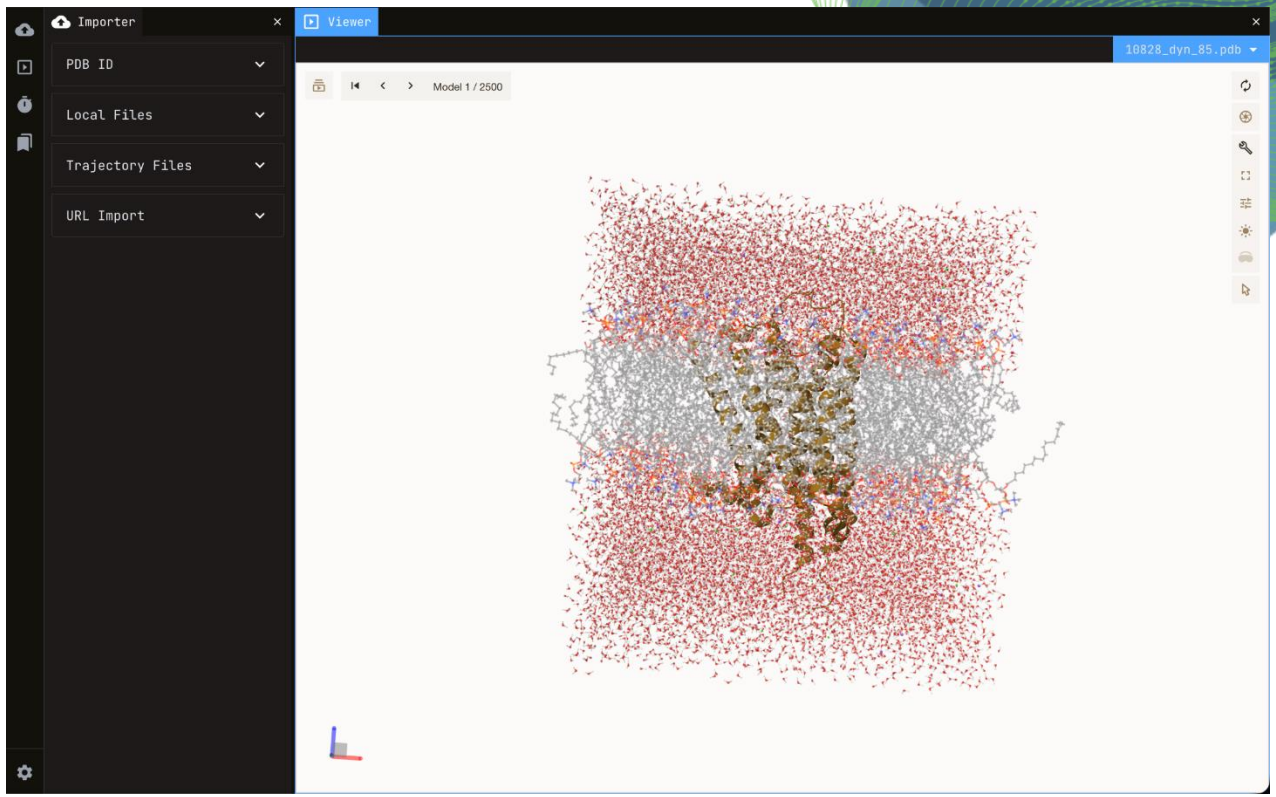
- We have large databases for MD simulations
- But: No labeled data for machine learning
- Use PhD students to create labeled data for a ML model:
  - Iteration loop with zero shot model and human annotator
  - Use metrics to detect uncertain areas in the latent space

We need an easy to use web based tool to convince people to do this → MDannotator



MDannotator

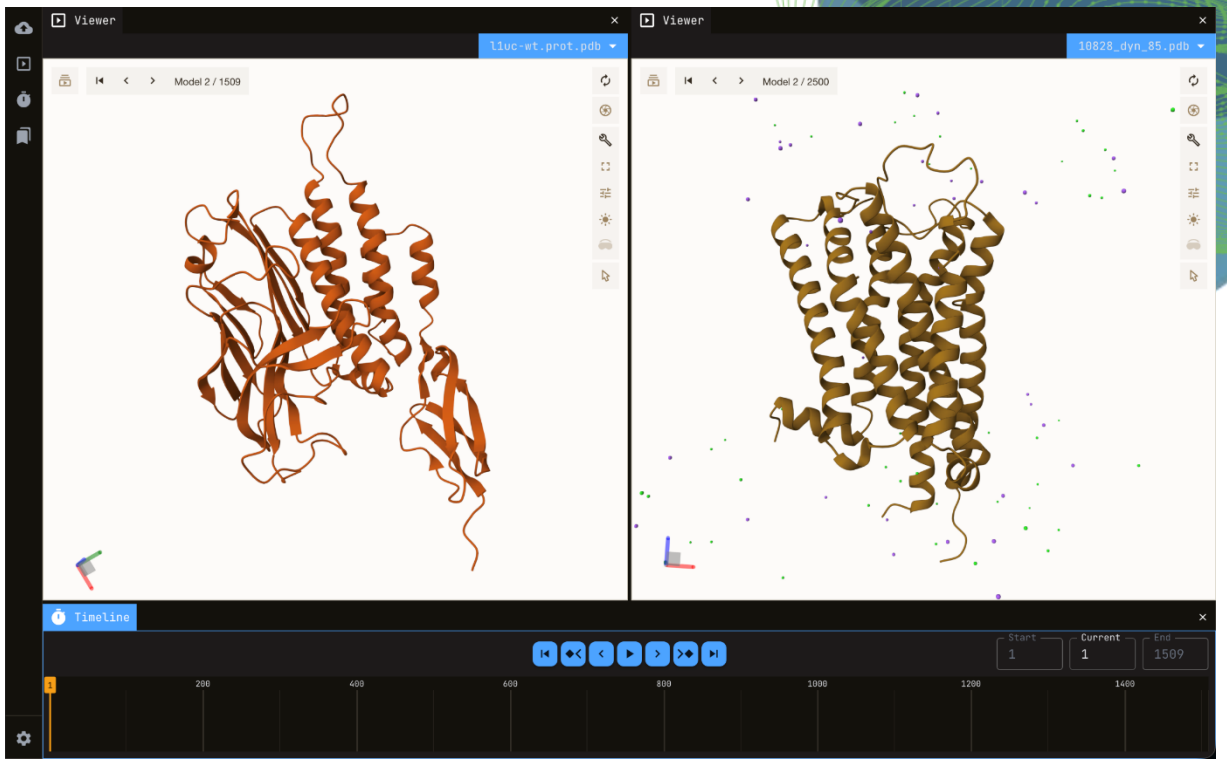
- **Using MDsrv [1] as starting point:**
  - Implementing a collaboration platform to share MD simulations
  - Comparison of simulation runs
  - Labeling mechanisms for presentation
  - Integrate metrics for exploration to support users
  - Develop a common format to describe events in MD simulations



[1] Kampfrath, Michelle, et al. "MDsrv: visual sharing and analysis of molecular dynamics simulations." *Nucleic Acids Research* 50.W1 (2022): W483-W489.

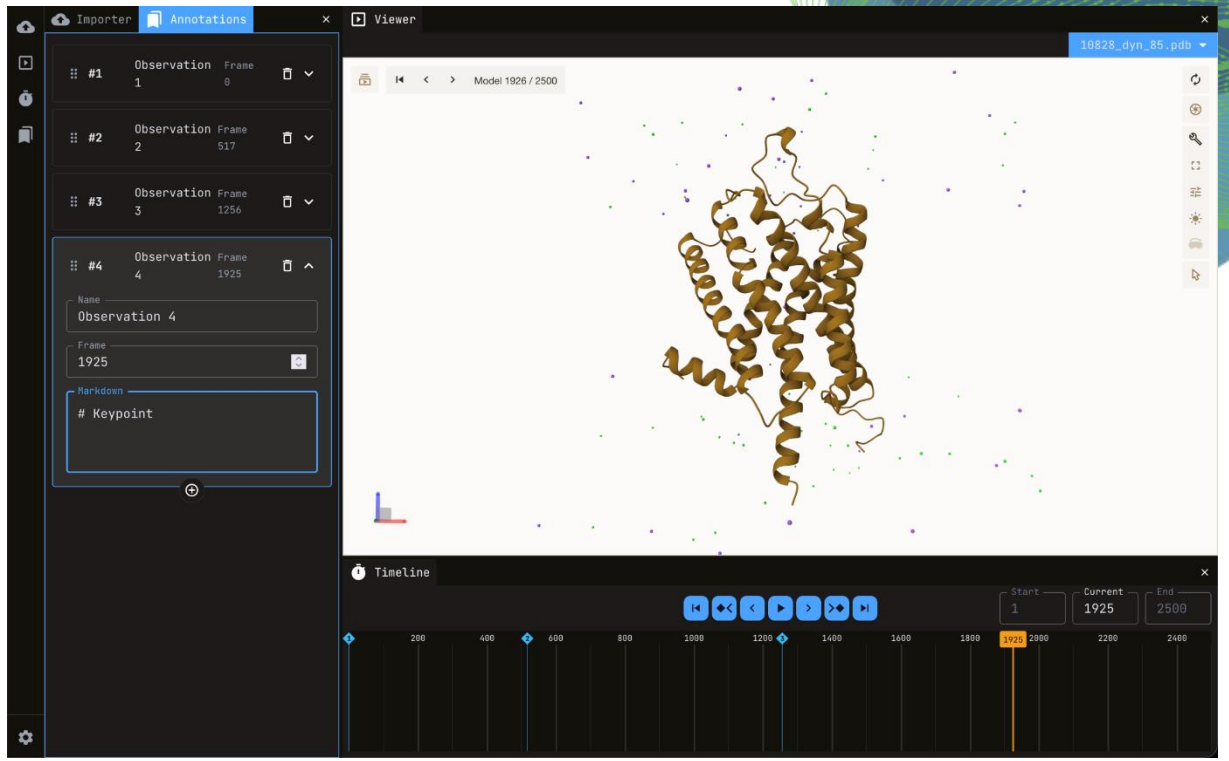
MDannotator – Comparison of Simulations

- **Coupled view with shared interactions:**
  - Shared timeline as coordinate system between different simulations
    - Using it as an overview for labeled events
  - Interpolation between different space coordinate systems



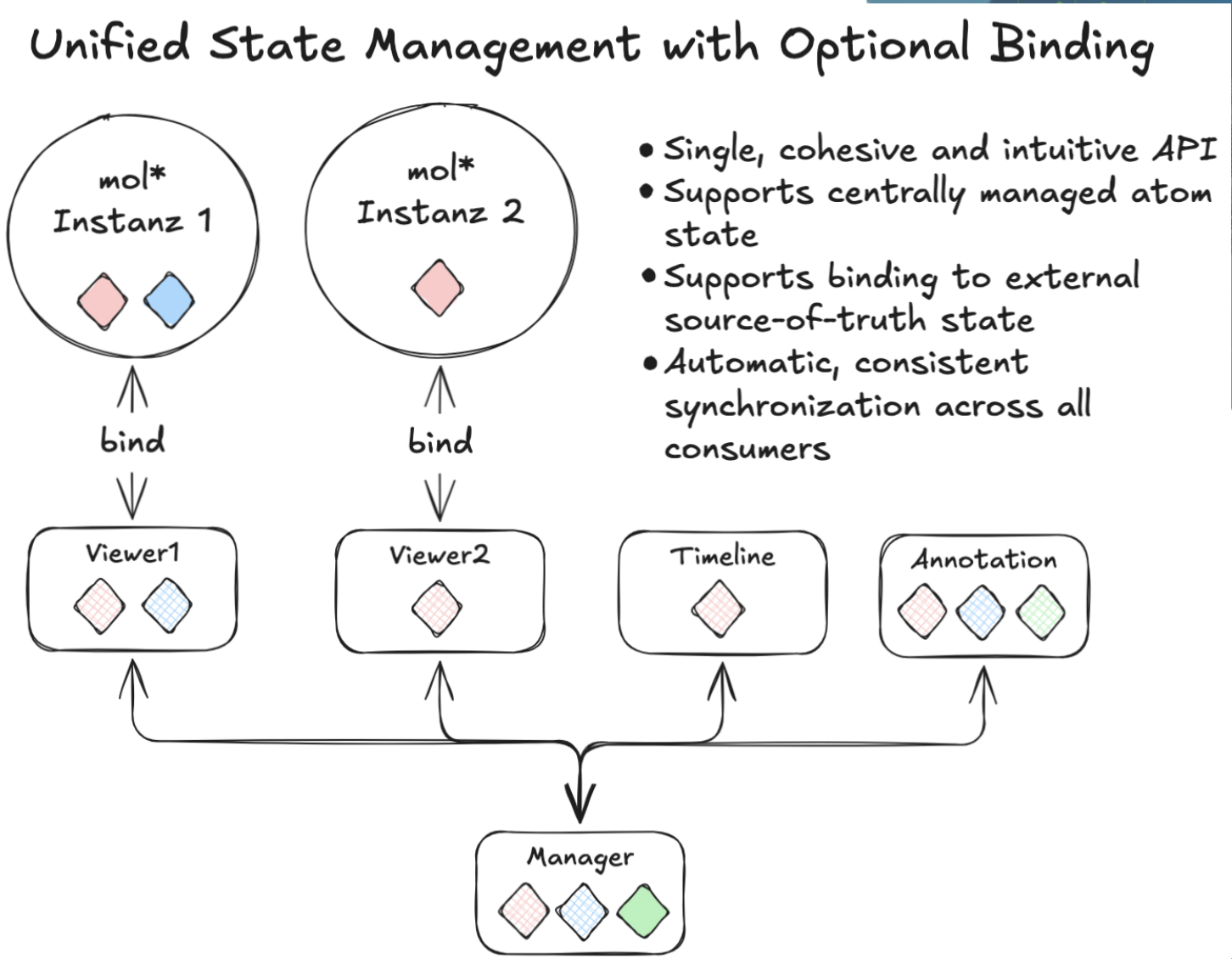
MDannotator – Labeling

- **Labels:**
  - Single or multiple frames
  - Metric values like RMSD
  - Binding events of local components
  - ….



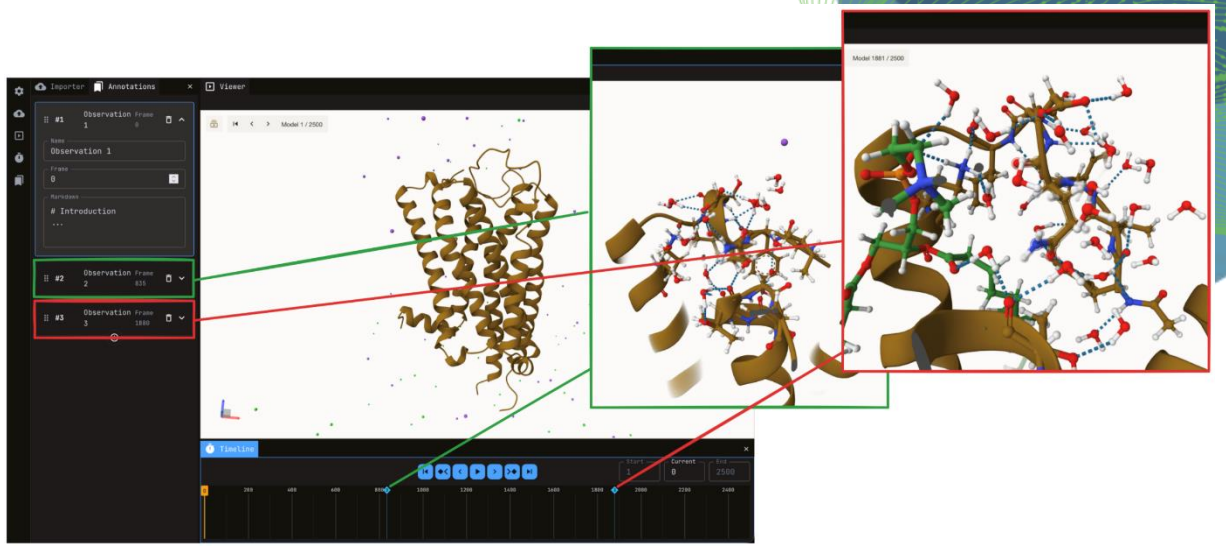
MDannotator - Architecture

- **We need a state management:**
  - Multiple views with different data
  - Interaction can modify multiple states
  - States can exist for varying lengths of time



MDannotator - Outlook

- Use user defined labels for the Active Learning approach to fine-tune a zero-shot approach
- Visualize predicted labels for inspection
- Generate uncertain labels for user correction
- Establish an machine readable format for MD simulation events for presentation and classification



Thanks for listening!

Questions?

Daniel Wiegreffe

Data Science Center ScaDS.AI Leipzig
Image and Signal Processing Group