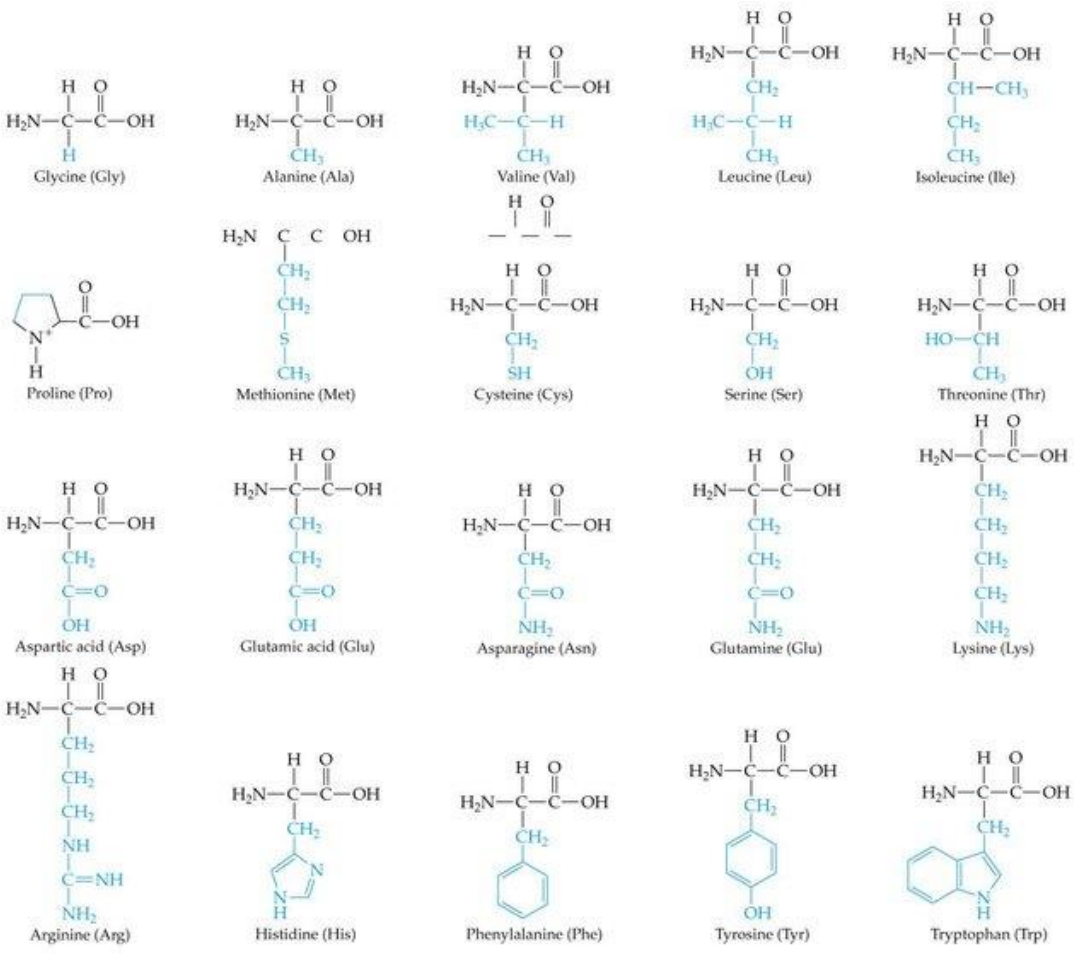
Free University of Bolzano - Bozen
Faculty of Engineering

Machine learning-driven simulation of protein folding atomistic trajectories

Alan Ianeselli

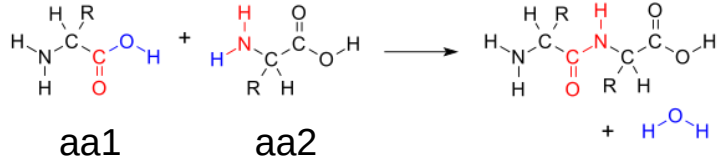
SFSCON 2023

Protein structure



~20 amino acids

They link by peptide bonds to form a polymer

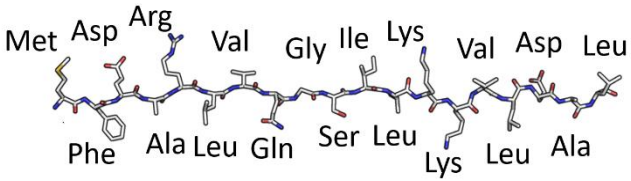


aa1 aa2

peptide bond

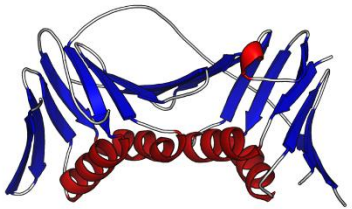
A protein is a polymer of tens, hundreds or thousands of amino acids

Protein structure

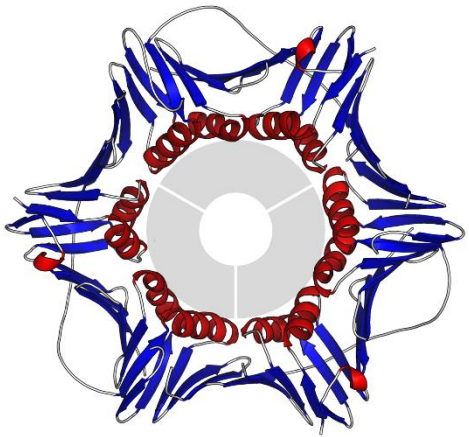


Primary structure
(aa sequence)

Secondary structure
(local structures)



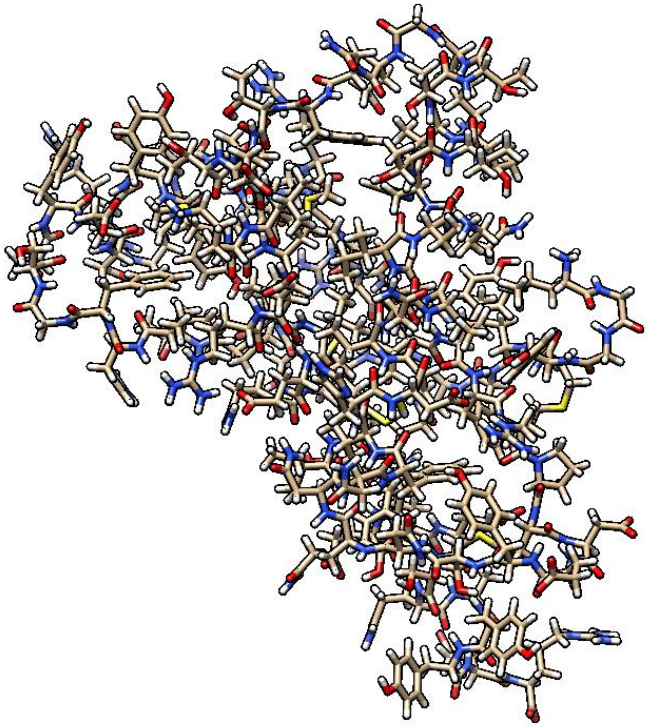
Tertiary structure
(3D conformation)



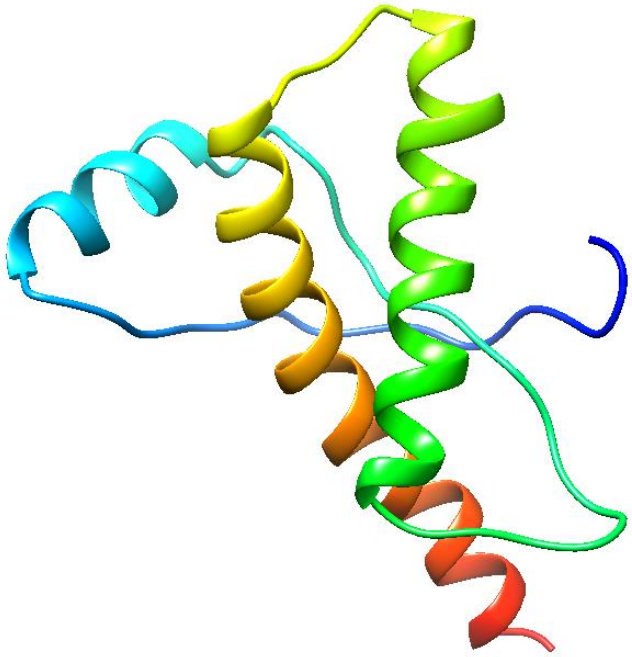
Quaternary structure
(protein complexes)

→ Unique conformation given a specific aminoacidic sequence
= the **protein folding problem**

Protein structure

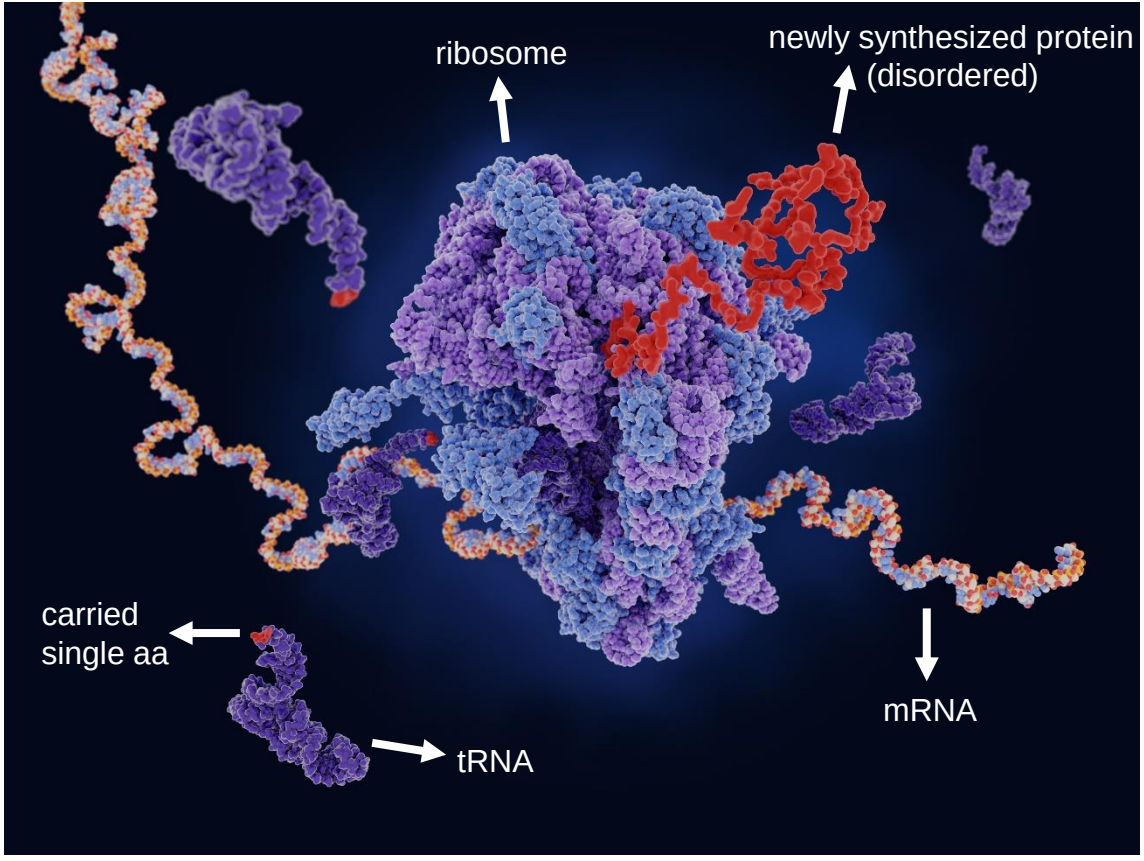


All-atom
representation



Ribbon
representation

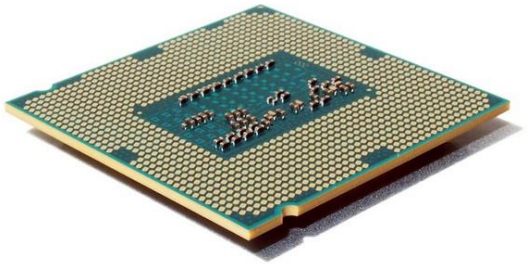
Protein synthesis



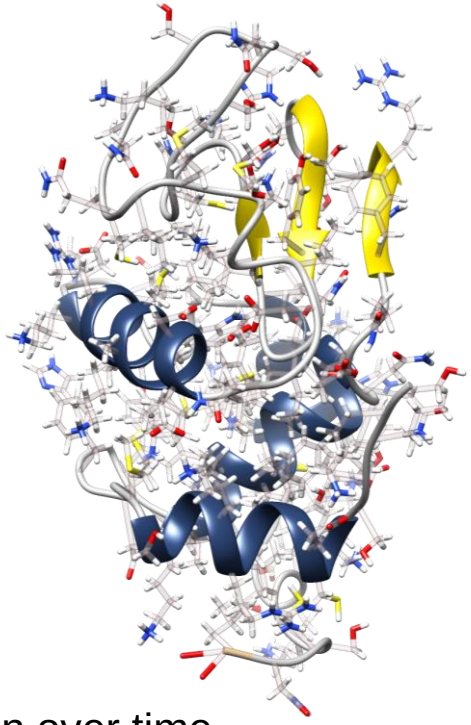
→ How does the newly synthesized disordered protein achieve its final conformation?

Author: Juan Gaertner

Molecular dynamics (MD) simulations



$$\vec{F} = m\vec{a}$$

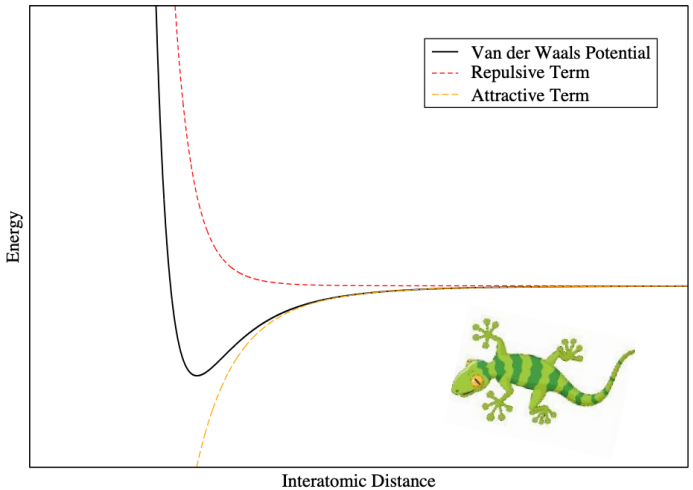


- Numerically solve Newton's equations of motion over time for each atom of the system

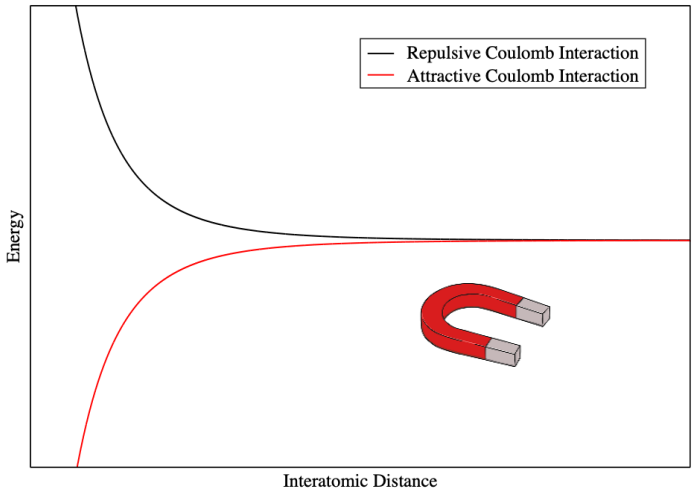
Molecular dynamics (MD) simulations

- Forces are computed from *force fields*

Non-bonded interactions



Van der Waals

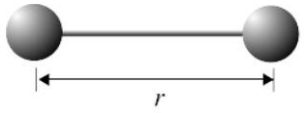


Coulomb

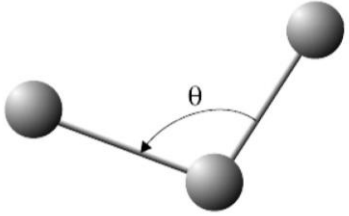
Molecular dynamics (MD) simulations

- Forces are computed from *force fields*

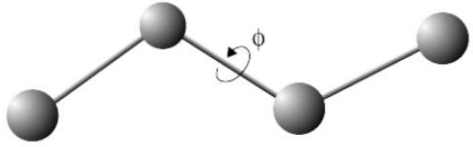
Bonded interactions



Bond stretching

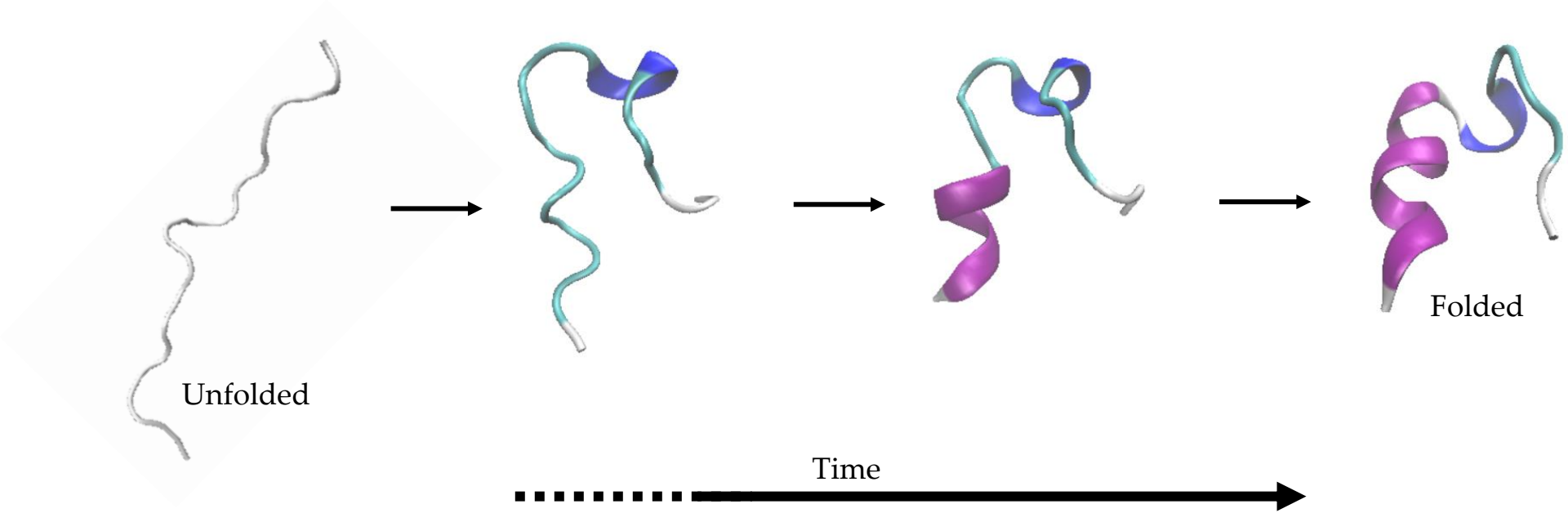


Angle bending



Dihedral torsions

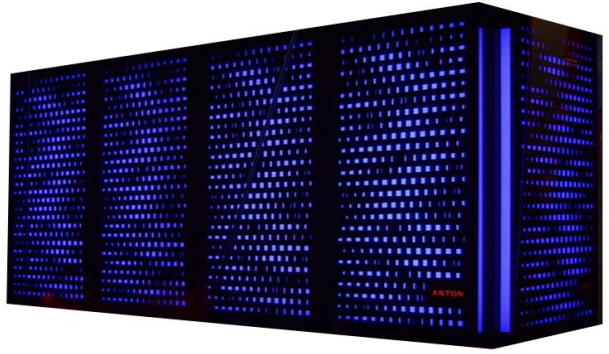
Molecular dynamics (MD) simulations



Folding time ~ microseconds (for small proteins)
= Weeks of simulation on supercomputers

Molecular dynamics (MD) simulations

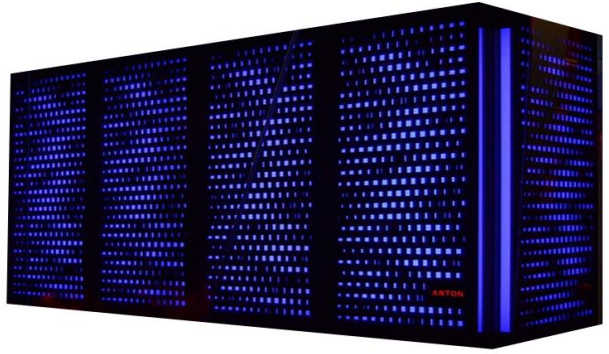
Anton Supercomputer
~50 μs/day for ~100'000 atoms



DE Shaw et al., 2009

Molecular dynamics (MD) simulations

Anton Supercomputer
~50 μs/day for ~100’000 atoms



DE Shaw et al., 2009

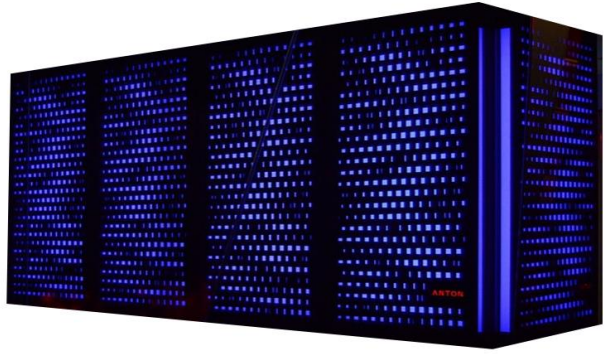
For example, Lysozyme in water (~100’000 atoms) requires SECONDS to fold

↓

~100 years of simulation!!

Molecular dynamics (MD) simulations

Anton Supercomputer
~50 μs/day for ~100’000 atoms



DE Shaw et al., 2009

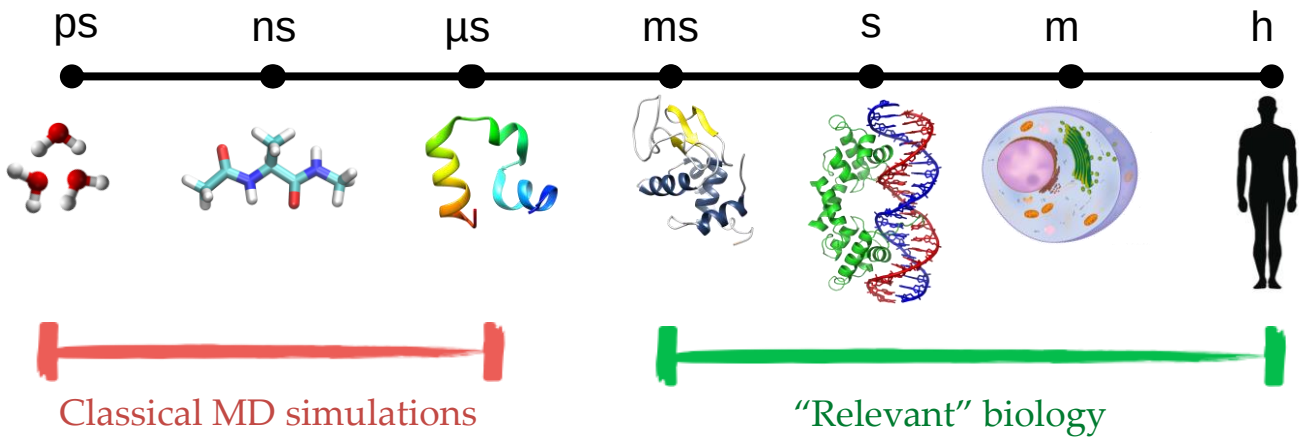
For example, Lysozyme in water (~100’000 atoms) requires SECONDS to fold

↓

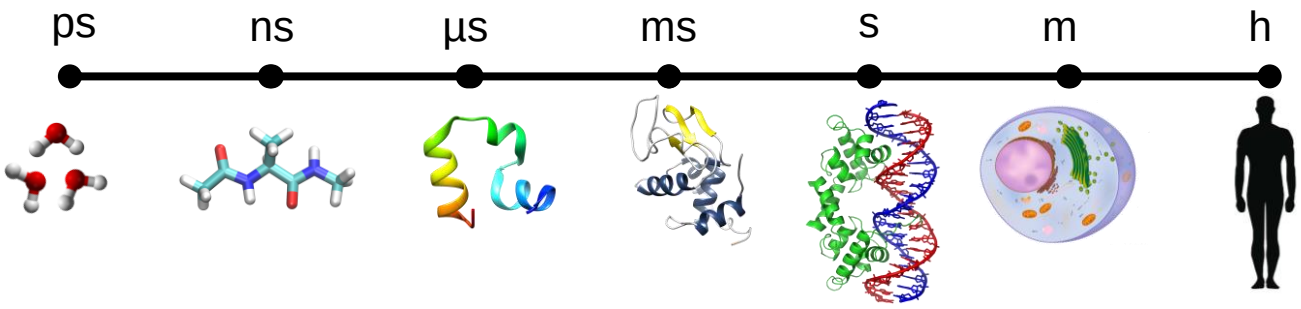
~100 years of simulation!!

- Conventional MD approaches are unfeasible

Timescales of macromolecules



Timescales of macromolecules

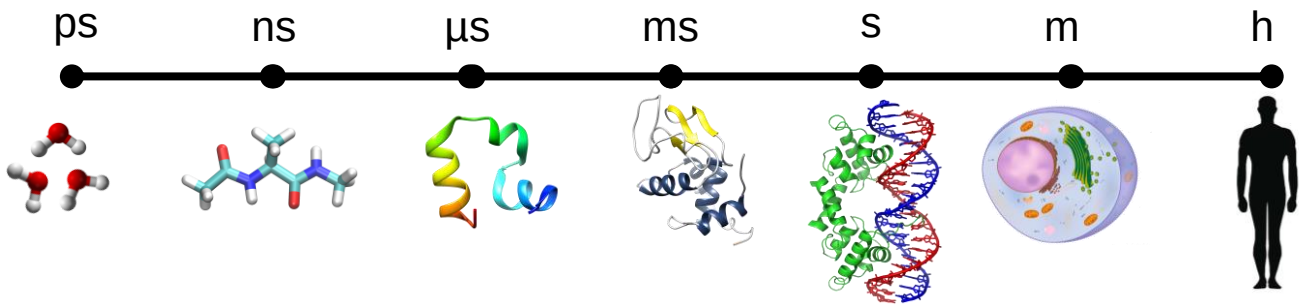


Classical MD simulations

"Relevant" biology

Coarse grained and approximated models

Timescales of macromolecules



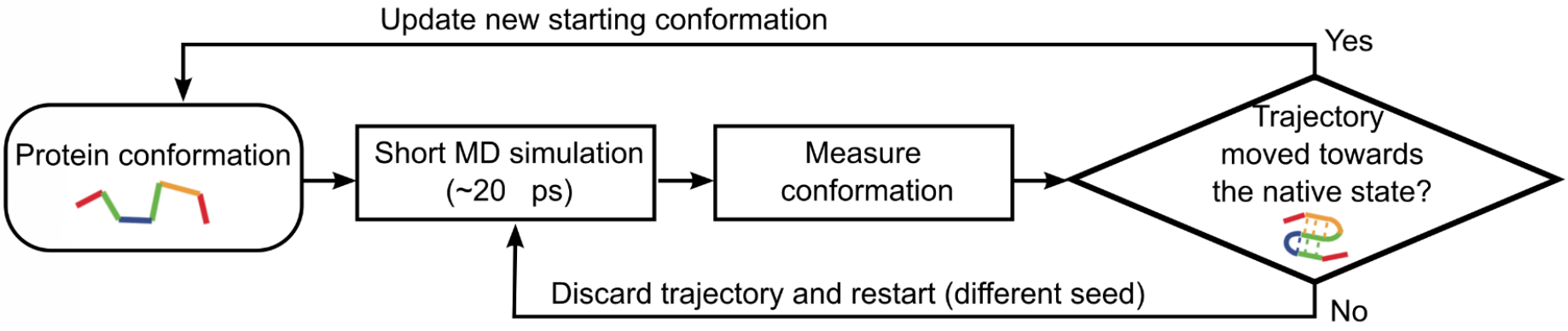
Classical MD simulations

"Relevant" biology

Coarse grained and approximated models
→ Sometimes unrealistic or unfalsifiable

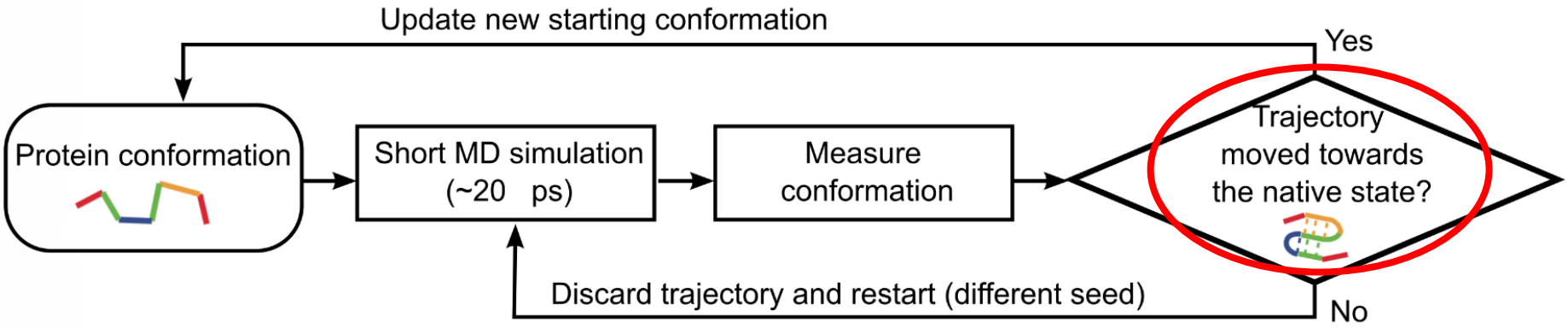
What I have in mind?

A smart **algorithm** to study protein folding trajectories



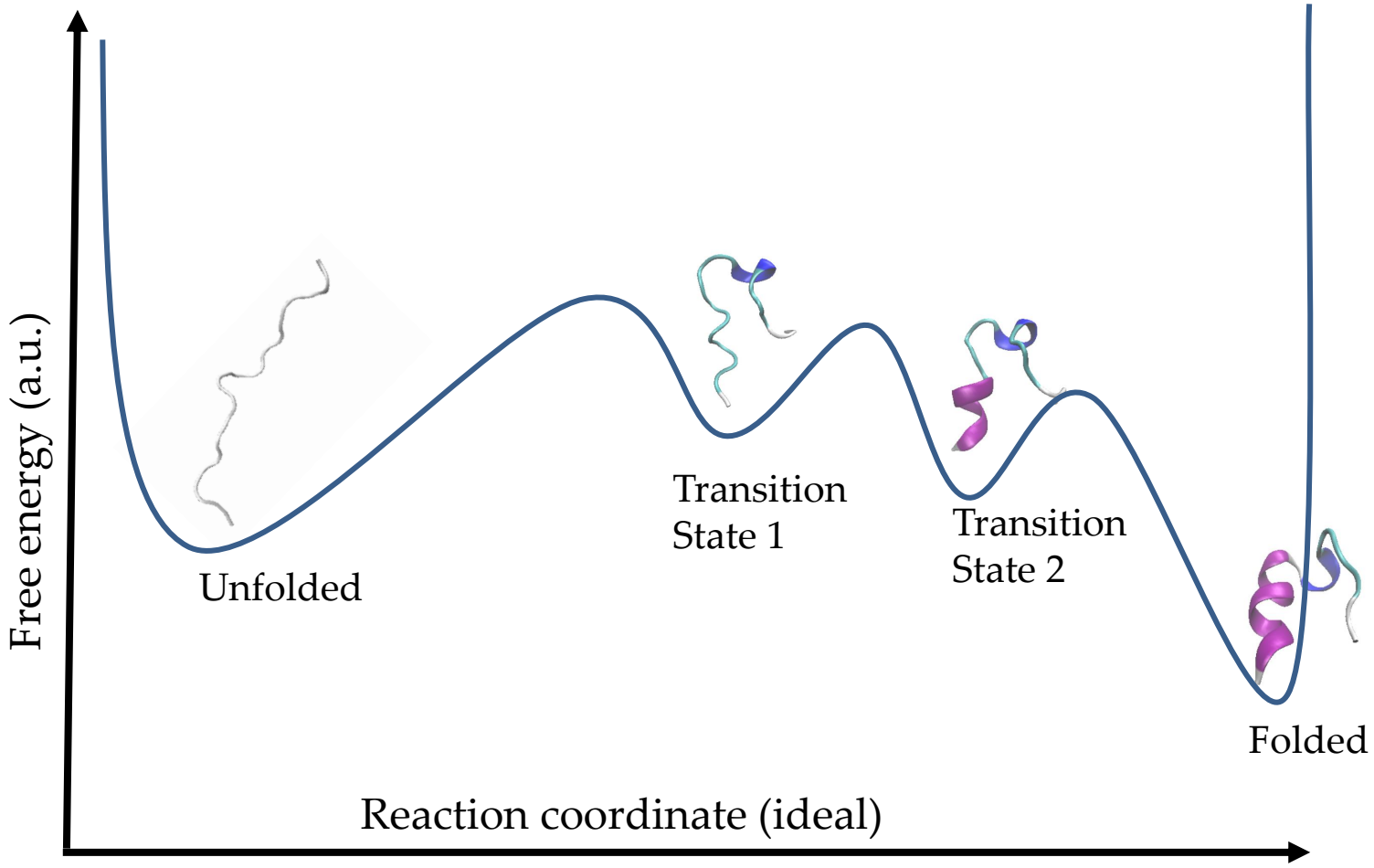
What I have in mind?

A smart **algorithm** to study protein folding trajectories

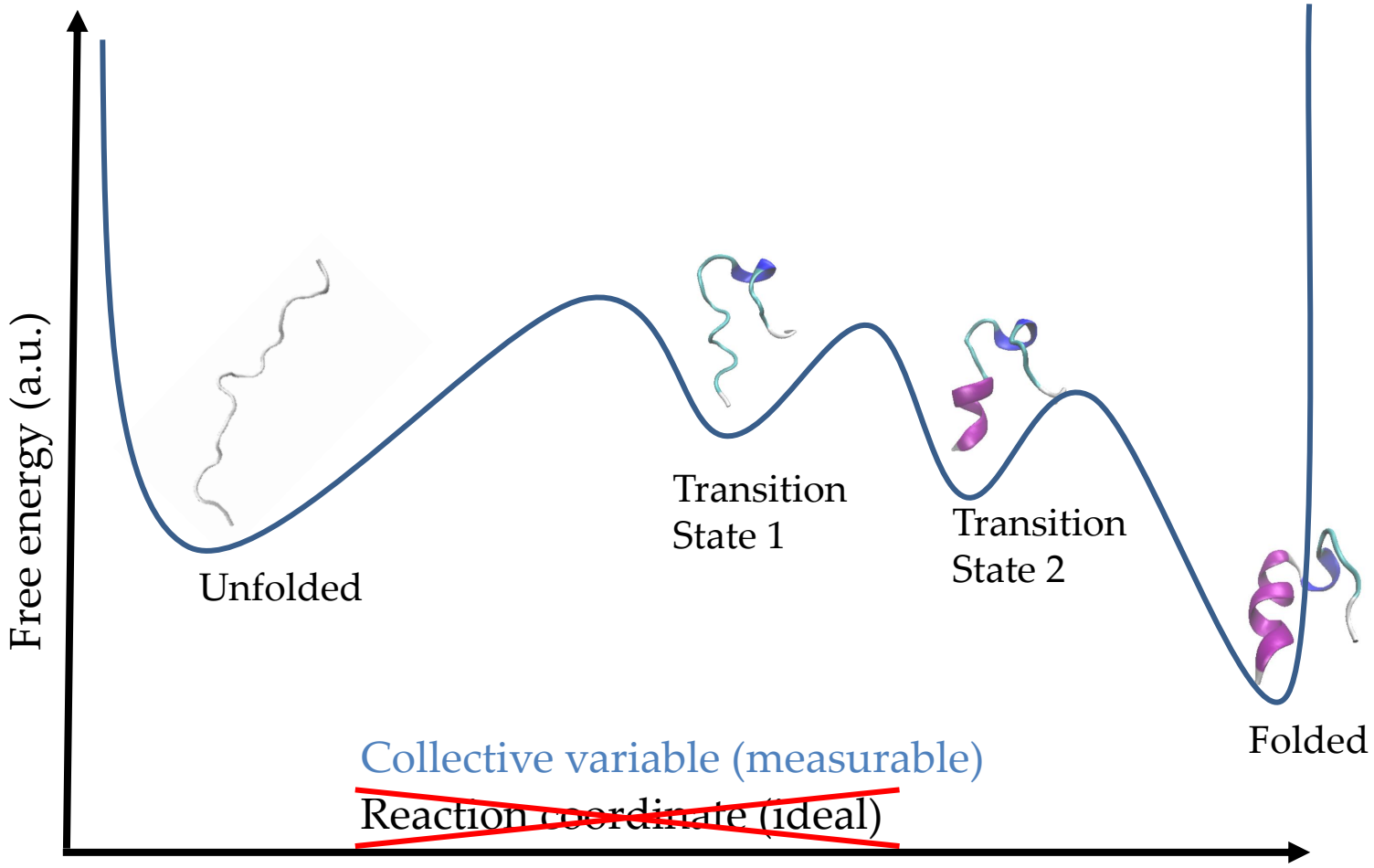


How do you measure if it went “forward”?

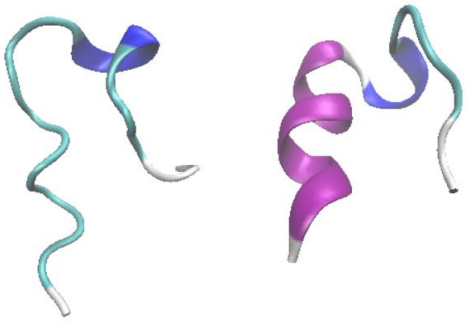
Reaction Coordinate and Collective Variables



Reaction Coordinate and Collective Variables



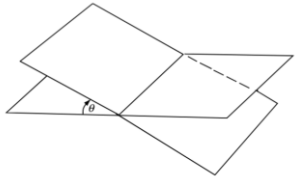
„Best“ collective variables (CVs) for protein folding



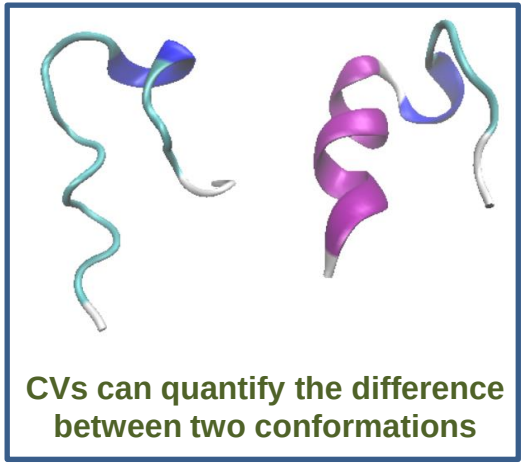
CVs can quantify the difference between two conformations

„Best“ collective variables (CVs) for protein folding

1. Dihedral angles deviation from the native state

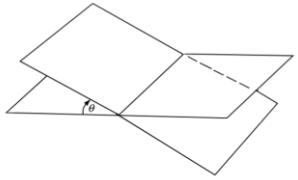


Syzonenko *et al*,
J. Chem. Inf. Model. 2020

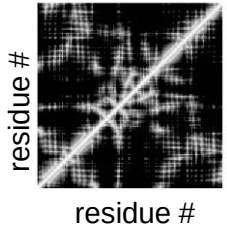


CVs can quantify the difference between two conformations

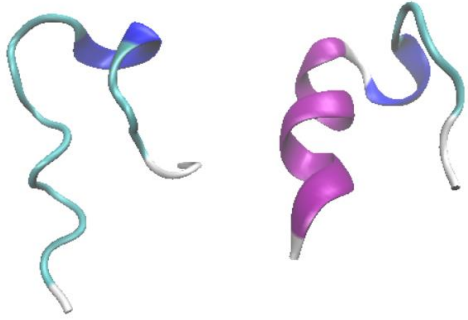
„Best“ collective variables (CVs) for protein folding



1. Dihedral angles deviation from the native state
   Syzonenko *et al*,
   J. Chem. Inf. Model. 2020

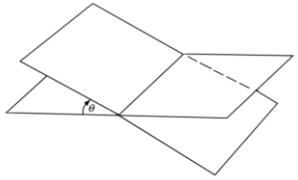


2. Inter-aa contact deviation from the native state
   Beccara *et al*,
   Phys. Rev. Lett. 2015

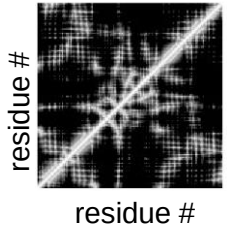


CVs can quantify the difference between two conformations

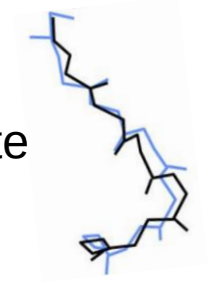
„Best“ collective variables (CVs) for protein folding



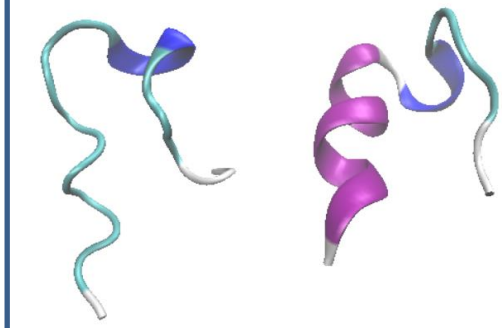
1. Dihedral angles deviation from the native state
   Syzonenko *et al*,
   J. Chem. Inf. Model. 2020



2. Inter-aa contact deviation from the native state
   Beccara *et al*,
   Phys. Rev. Lett. 2015



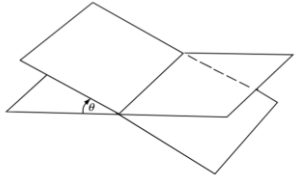
3. Geometrical difference from the native state



CVs can quantify the difference between two conformations

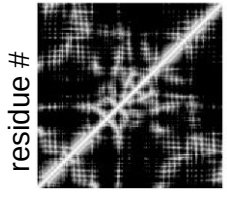
„Best“ collective variables (CVs) for protein folding

1. Dihedral angles deviation from the native state



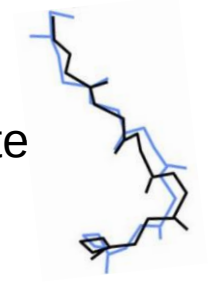
Syzonenko *et al*,
J. Chem. Inf. Model. 2020

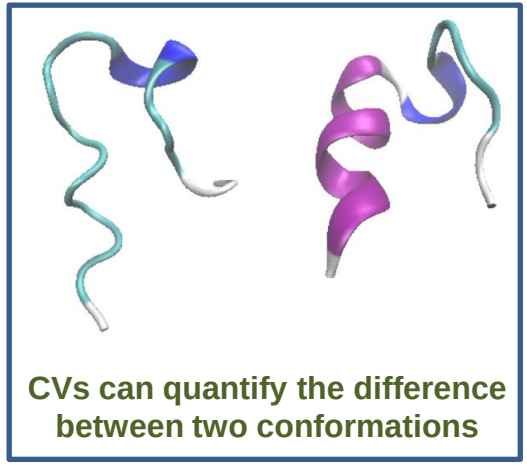
2. Inter-aa contact deviation from the native state



Beccara *et al*,
Phys. Rev. Lett. 2015

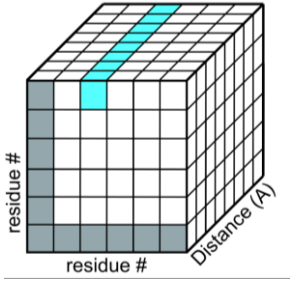
3. Geometrical difference from the native state





CVs can quantify the difference between two conformations

4. Deviation from **Google’s Deepmind AlphaFold** tensor



Alphafold

4. Deviation from **Google's Deepmind AlphaFold** tensor

Google's Deepmind Alphafold:

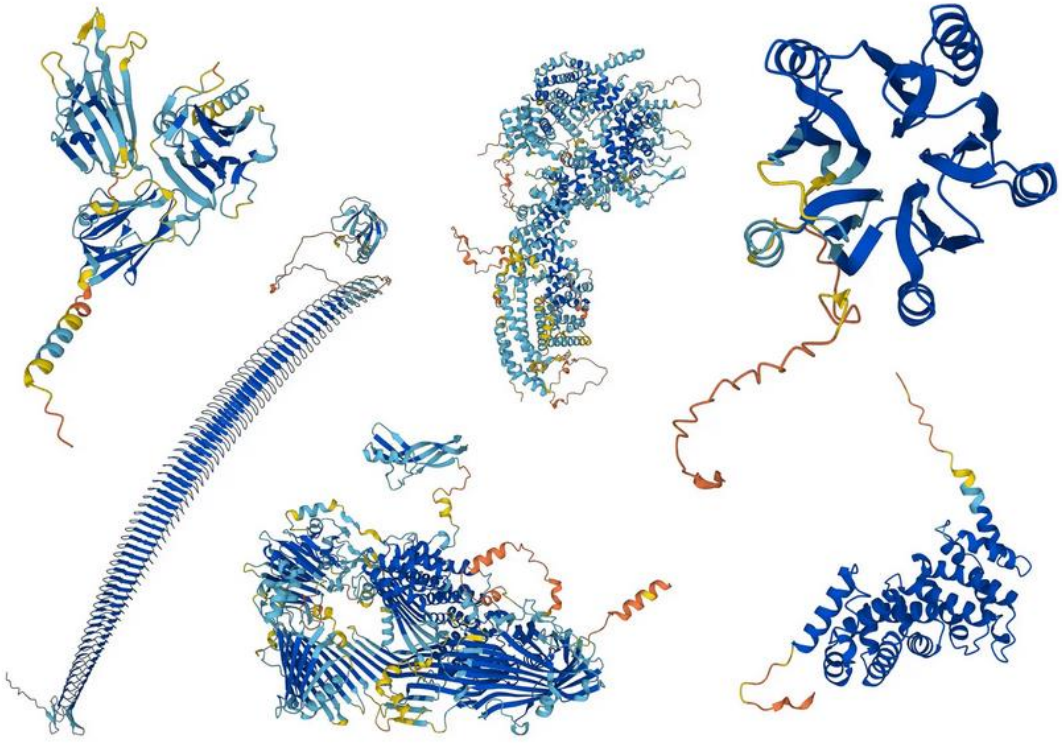
Latest AI milestone in the protein folding field

Training set: **170k** protein structures

Able to predict more than **200 million** of structures

Unprecedented accuracy

→ able to predict the final conformation of any aminoacidic sequence

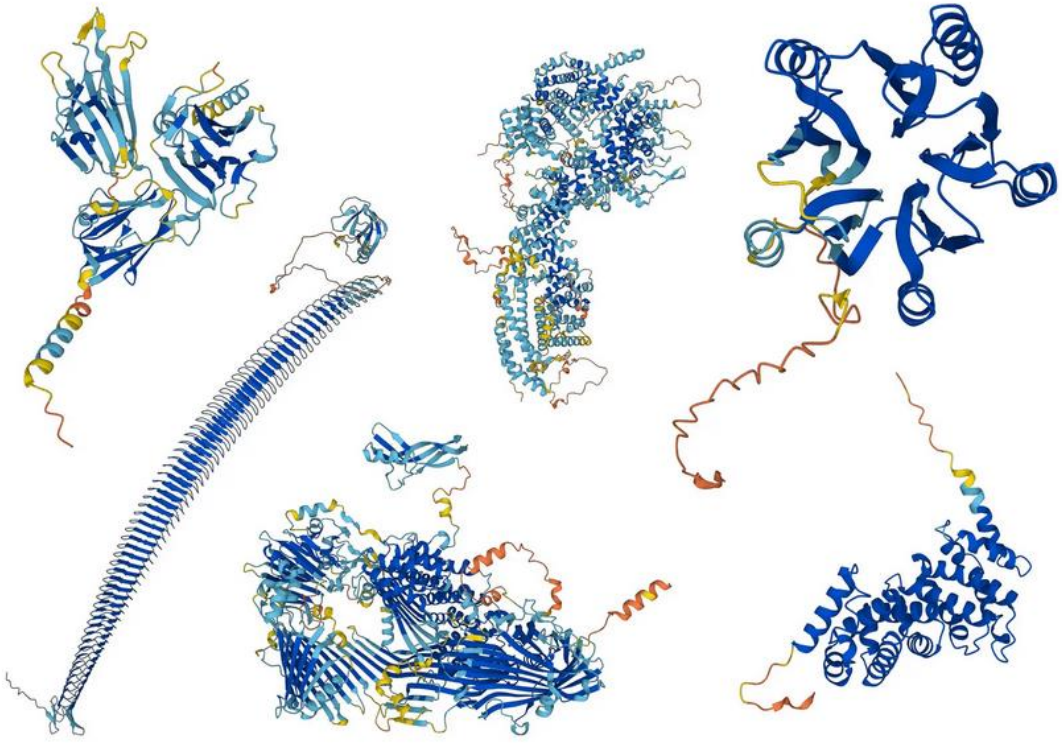


Alphafold

4. Deviation from **Google's Deepmind AlphaFold** tensor

Google's Deepmind Alphafold:

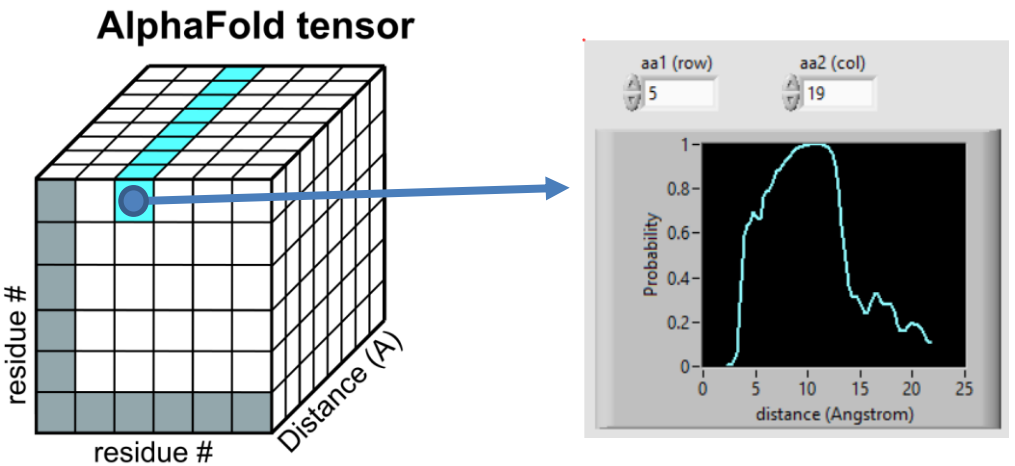
-**Input** = aa sequence (text string)
...
-Sequence alignment (database comparison)
-Prediction of distance and angle between aa pairs
...
-**Output** = 3D protein structure



4. Deviation from **Google's Deepmind AlphaFold tensor**

One of the outputs of AlphaFold is the so-called **distogram:**
Tensor of distance bins x aa x aa

→ **probability over distance** between pairs of aa
example below: aa at position 5 (Asparagine) vs aa at position 19 (Aspartic acid)



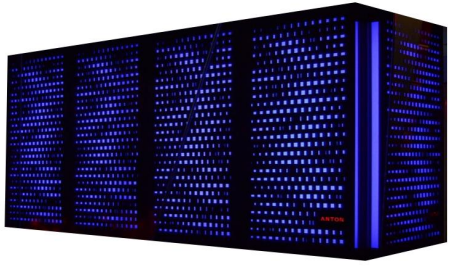
→ machine-learned 170k protein conformations

→ corresponds to a **quasi-chemical potential**

Identify the optimal CV

Training set:

Very long folding trajectories obtained by the most powerful supercomputer (Anton)
→ **200μs** of trajectories (Villin and Fip35 proteins)

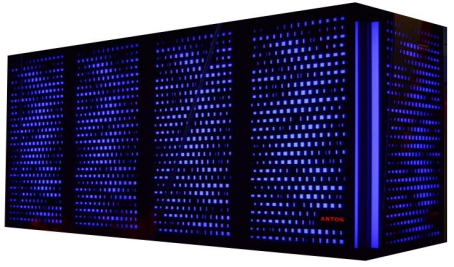


Anton

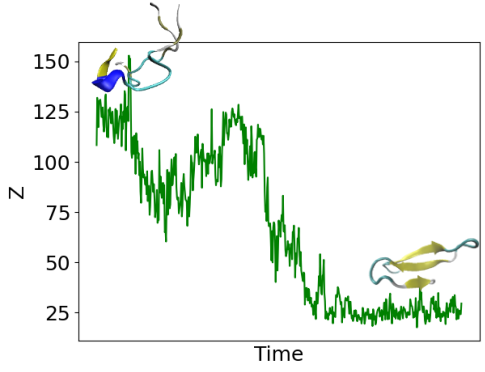
Identify the optimal CV

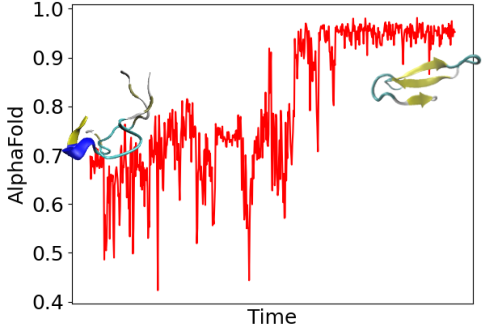
Training set:

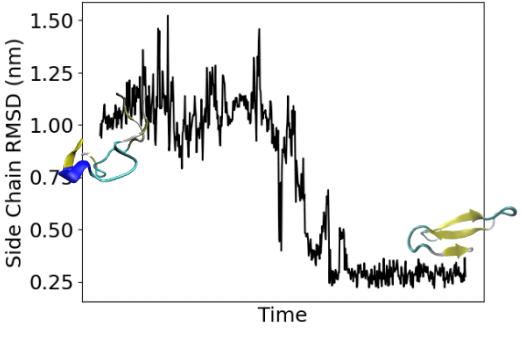
Very long folding trajectories obtained by the most powerful supercomputer (Anton)
→ **200μs** of trajectories (Villin and Fip35 proteins)

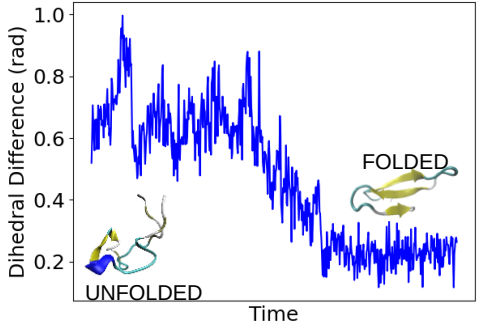


Anton









Identify the optimal CV

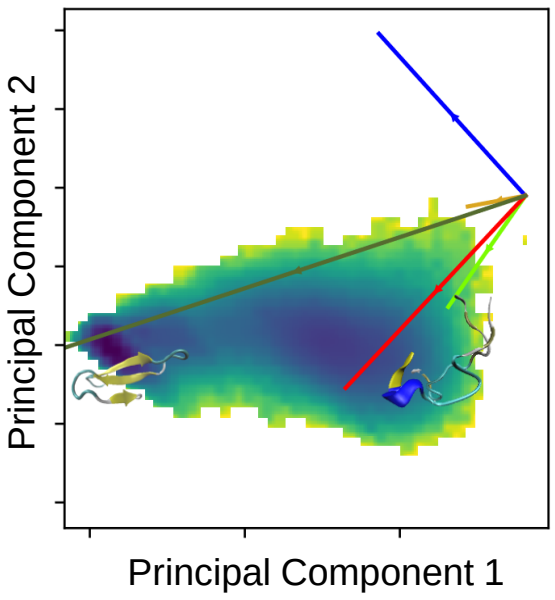
Training set:

Very long folding trajectories obtained by the most powerful supercomputer (Anton)
→ **200μs** of trajectories (Villin and Fip35 proteins)

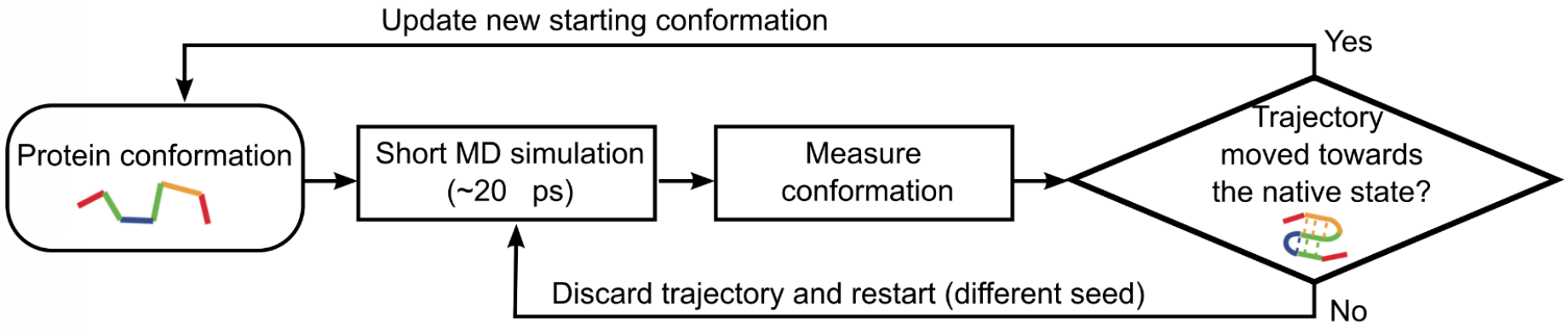
Training set:
396k rows
4 features

Machine learning (PCA)

Optimal CV identified



Run MD folding algorithm



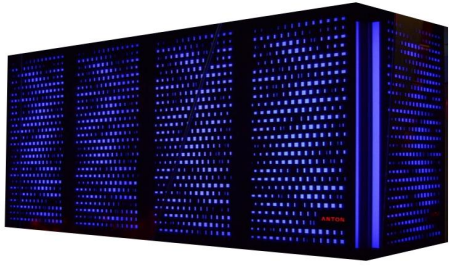
Towards the native state = **along the optimal CV**

NOTE: trajectories are unbiased

Results

Obtained the **folding trajectories** of 4 small proteins

My simulation time = 1 day per trajectory on a weak laptop
(Anton supercomputer would need weeks)

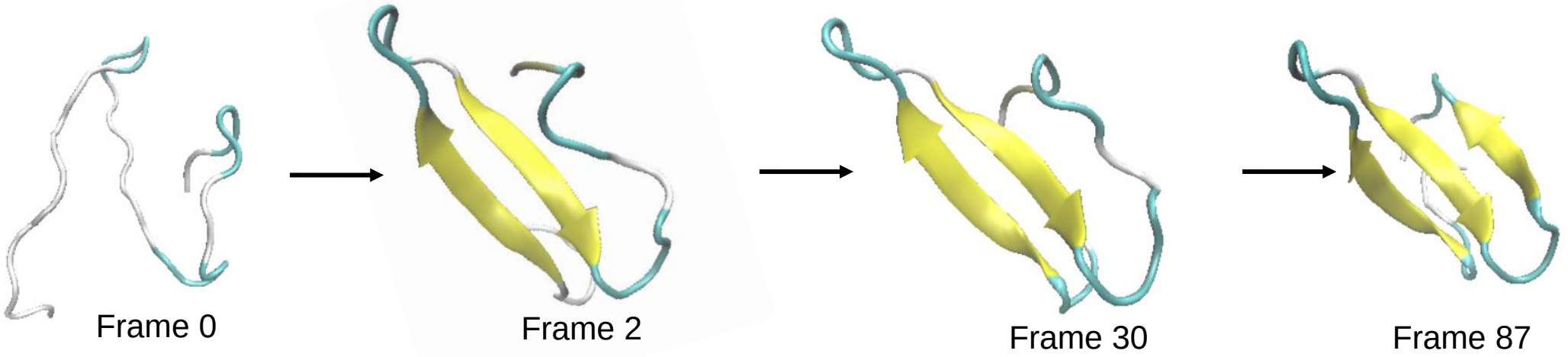


Brute force MD
~weeks

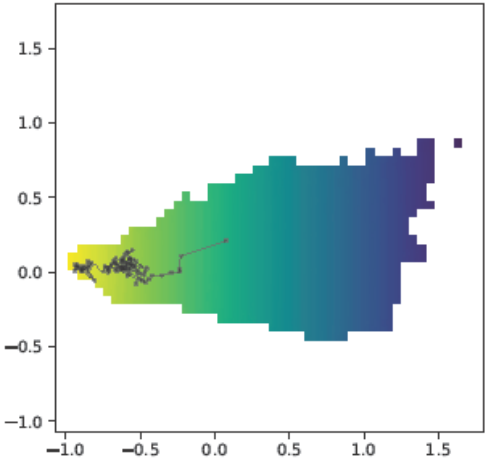
VS

Smart algorithm
~hours

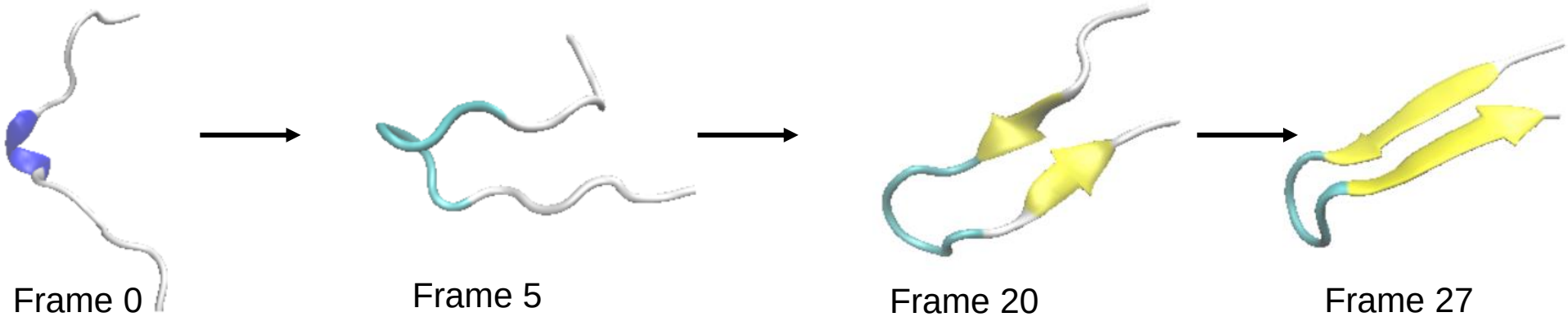
Results
Fip35



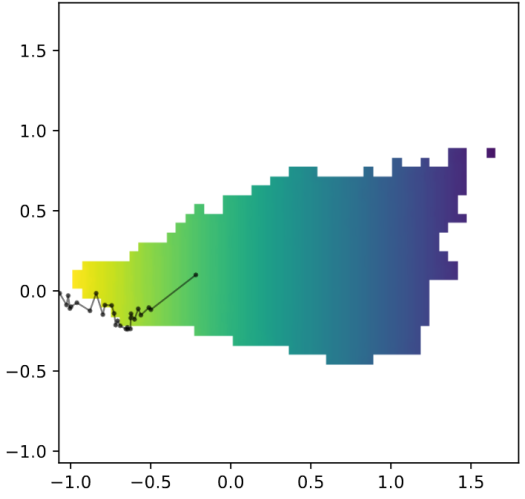
Fip35 (35 aa, 562 atoms): 16 trajectories, 1 folded



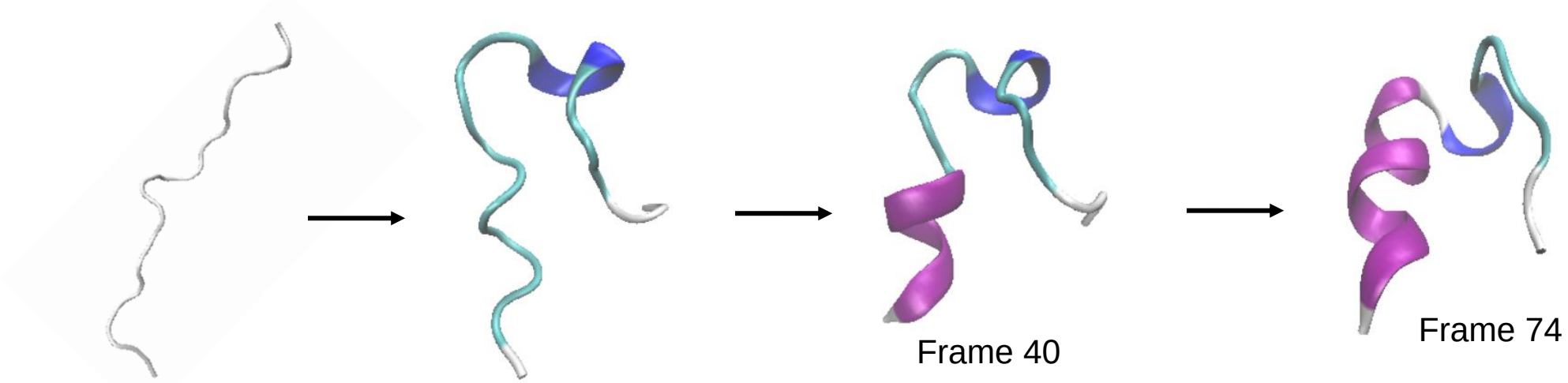
Results
Beta Hairpin



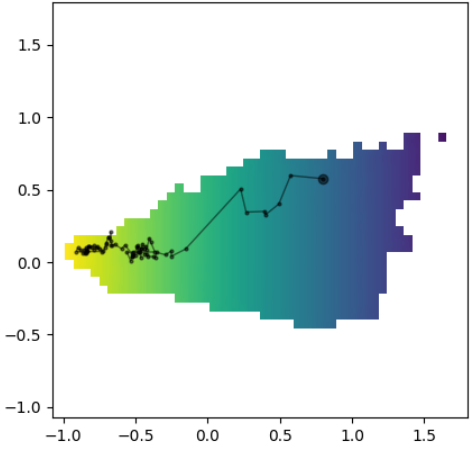
Beta Hairpin (16 aa, 247 atoms): 3 trajectories, 1 folded



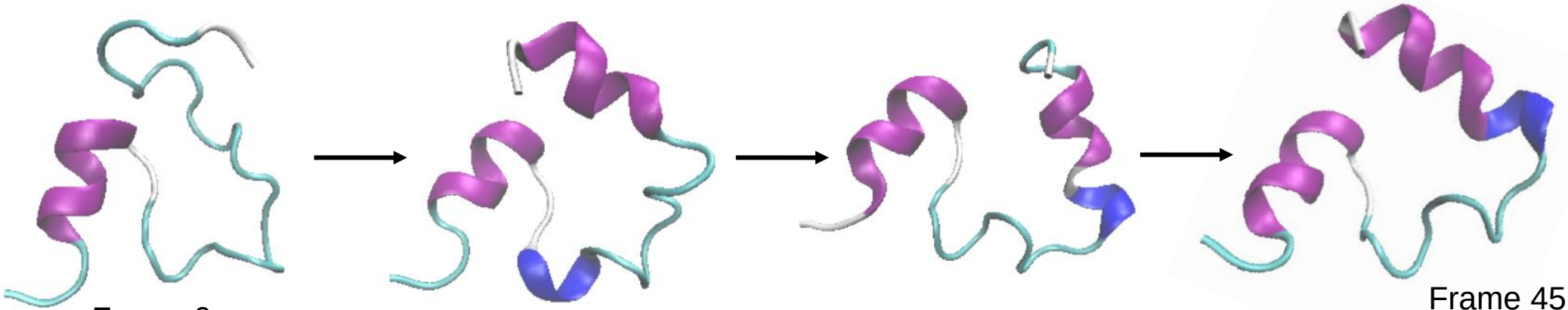
Results
TrpCage



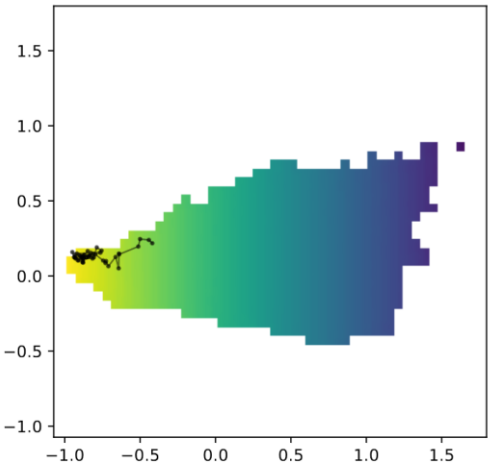
TrpCage (20 aa, 304 atoms): 12 trajs, 4 folded



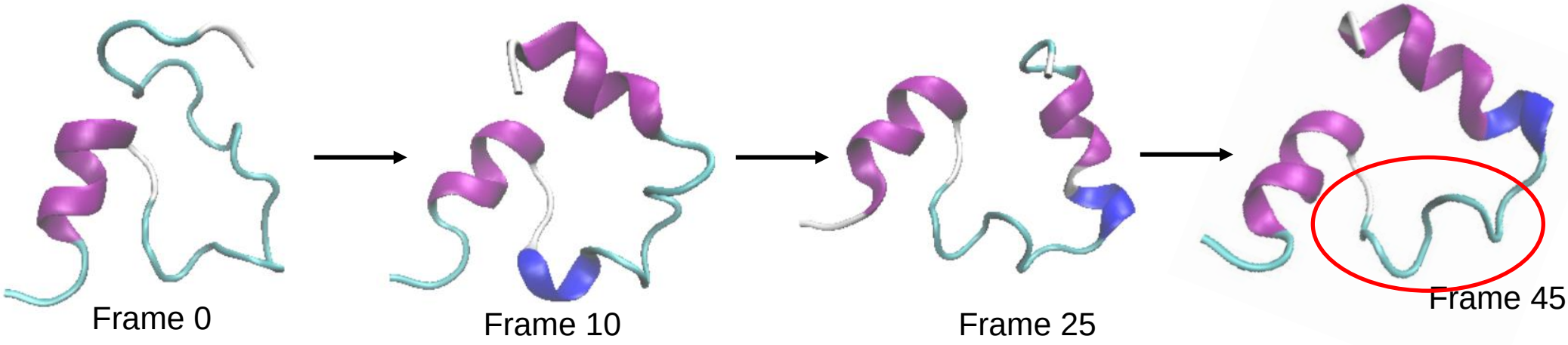
Results
Villin



Villin (35 aa, 583 atoms): 16 trajs, 4 “almost” folded

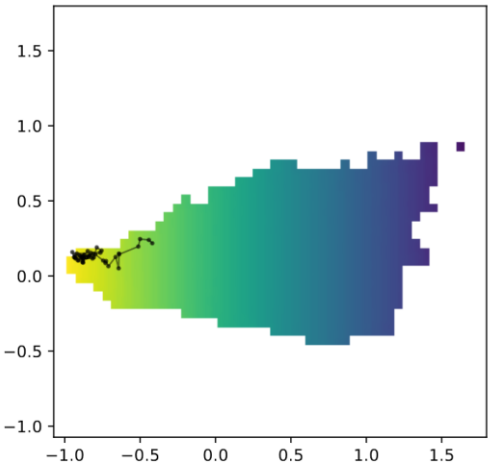


Results
Villin



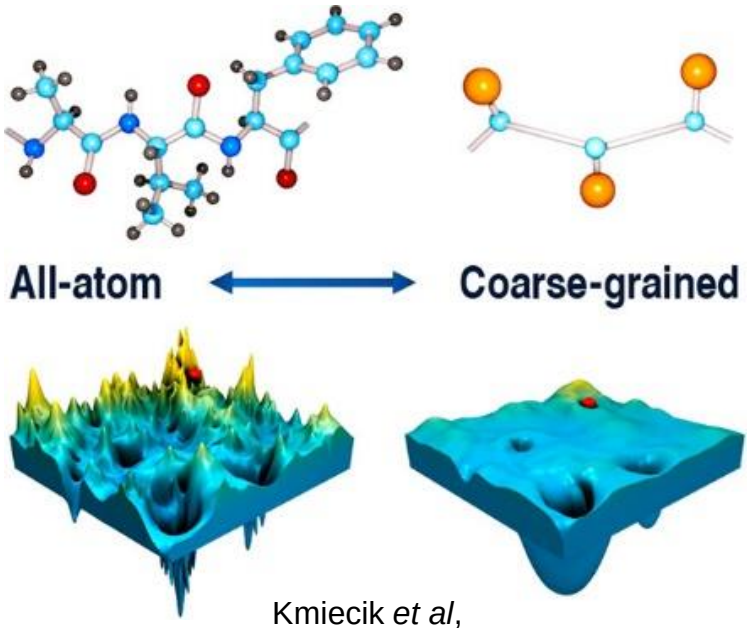
Villin (35 aa, 583 atoms): 16 trajs, 4 “almost” folded

→ High energy barrier for the last helix?



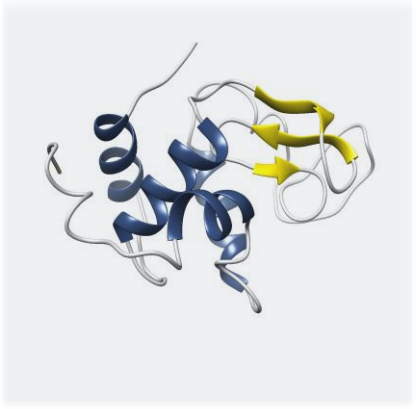
Future works

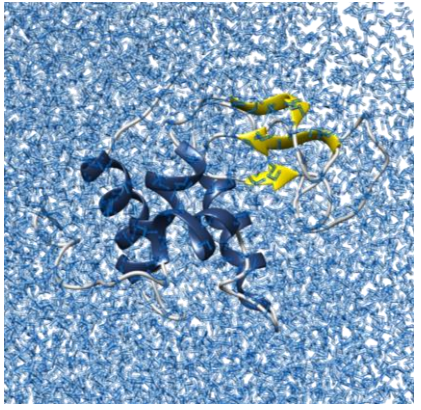
Coarse graining



Kmiecik *et al*,
Chem. Rev. 2016,

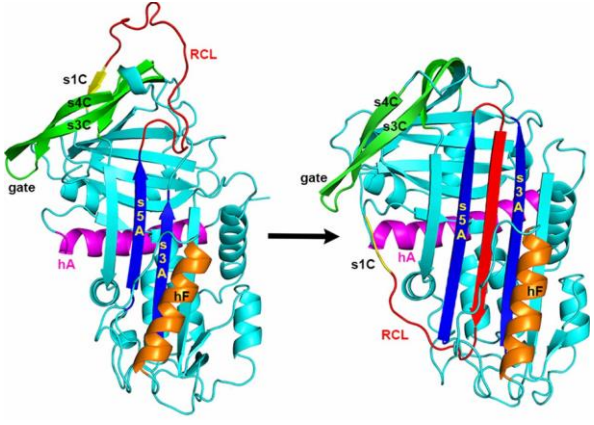
Explicit solvent



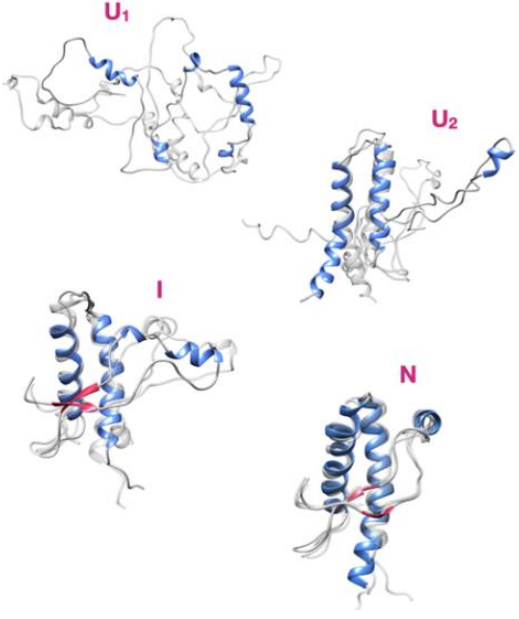


Method applications

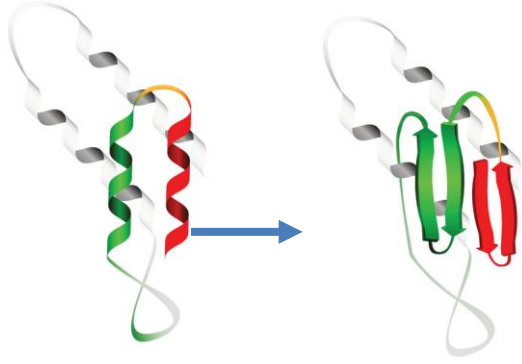
Conformational transitions and point mutations



Identify intermediate conformations



Misfolding



unibz

Thanks for the attention!

SFSCON

Prof. Diego Calvanese
Smart Data Factory
University of Bolzano

Prof. Emiliano Biasini
Collaborator
University of Trento

CIBIO
UNIVERSITÀ DEGLI STUDI DI TRENTO

Prof. Pietro Faccioli
Master‘s supervisor
University of Milan

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA