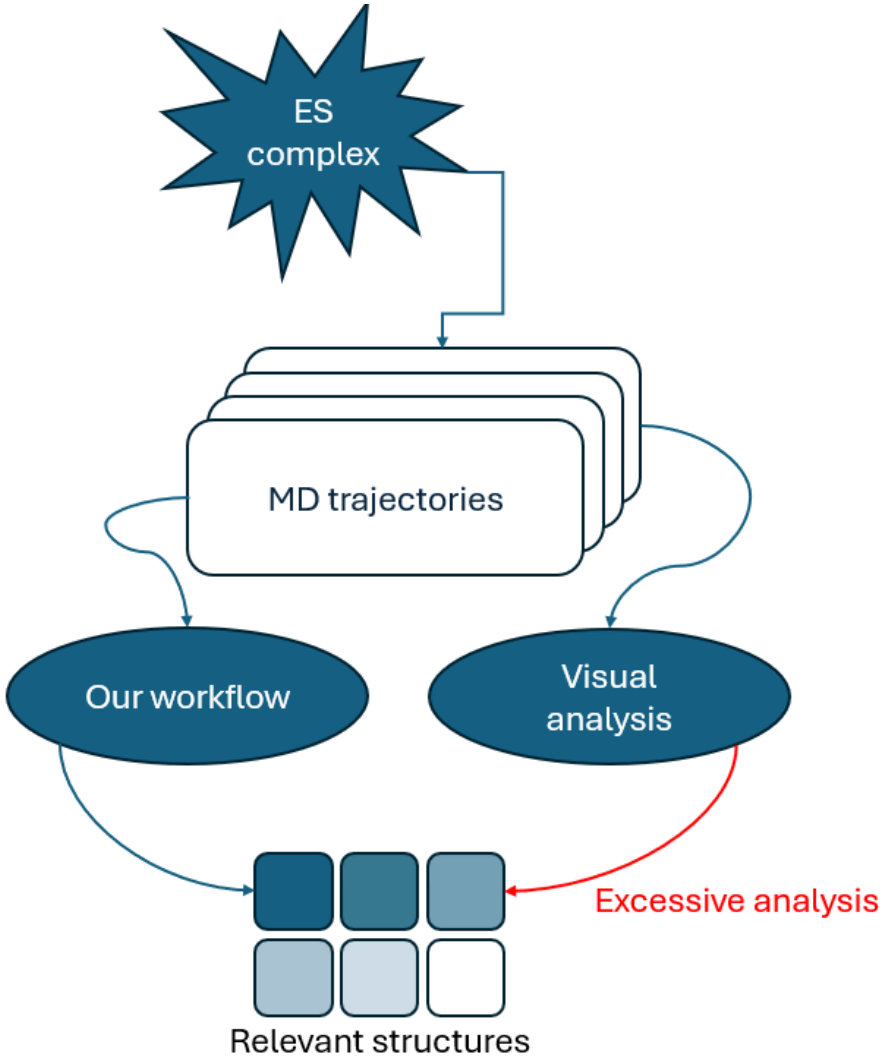
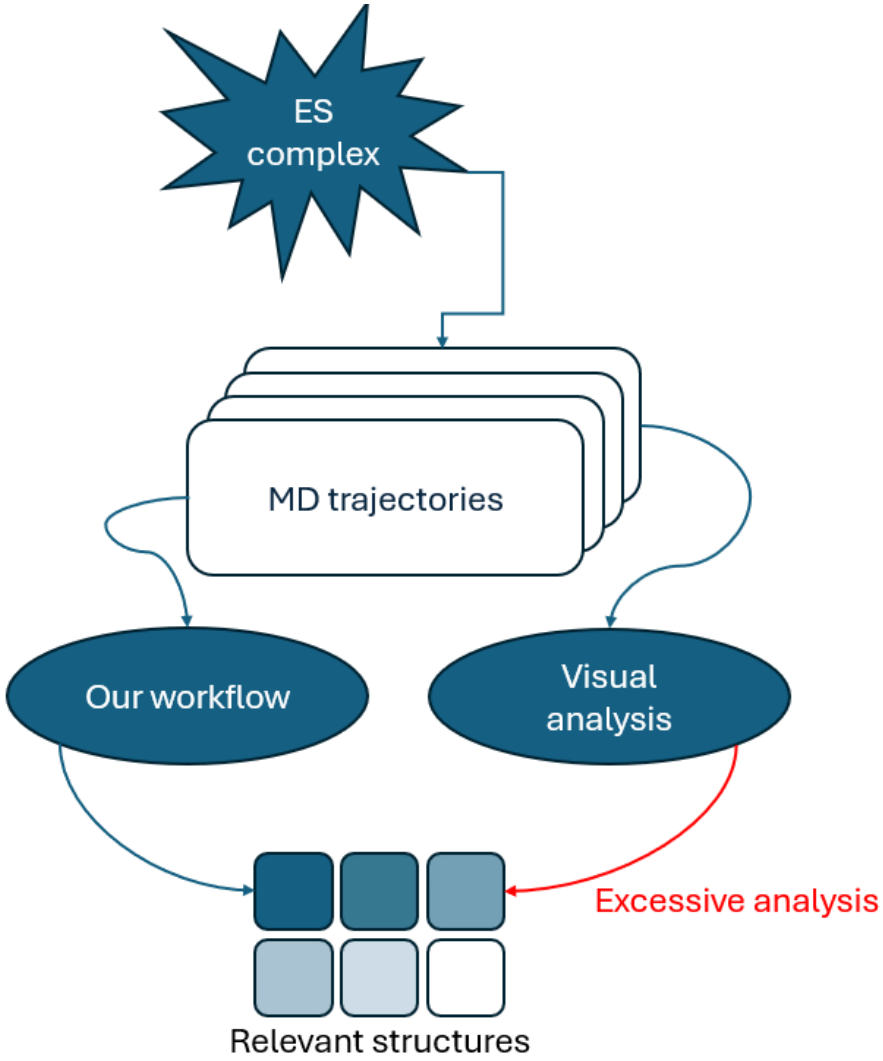
Dimensionality Reduction and Clustering Framework for Mapping Hydrolase Active Site Dynamics

M.E. Kuznetsov, I.V. Polyakov

Lomonosov Moscow State University, Chemistry Department
Emanuel Institute of Biochemical Physics RAS

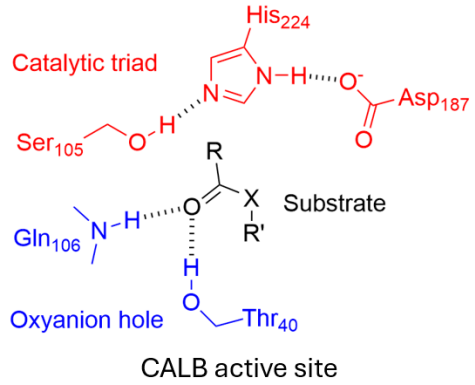
20.10.2026


ES complex
MD trajectories
Our workflow
Visual analysis
Excessive analysis
Relevant structures

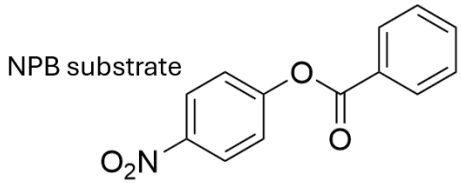


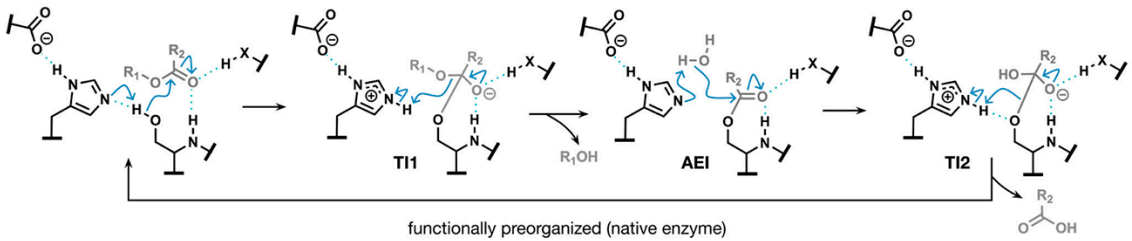
The proposed framework was tested on the *Candida antarctica lipase B* (CALB) complexed with the p-nitrophenyl benzoate (pNPB) substrate.





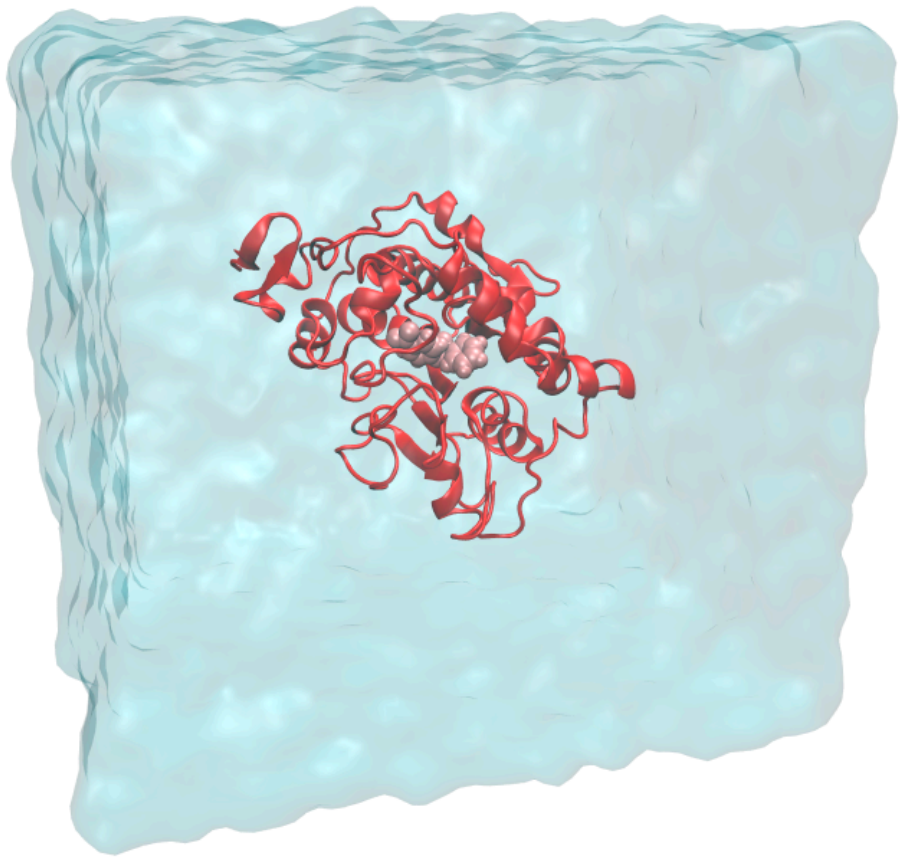
CALB active site





(10.1126/science.adu2454).

MD simulations



Structure

PDB ID: 6TP8, manually added NPB

Solvation, ionization → cell $72 \cdot 82 \cdot 75$ Å, 48k atoms

MD parameters

NPT-ensemble: 1 atm, 298 K.

Force fields CHARMM36m (protein), CGenFF (NPB), TIP3P (water).

Integration step: 2 fs (SHAKE/SETTLE).

Enhanced sampling by GAMD (200 ns) then 2 MD trajectories of 100 ns.

Trajectory analysis: python workflow, using MDAnalysis, scikit-learn. Feature Extraction → Dimensionality reduction → Clustering.

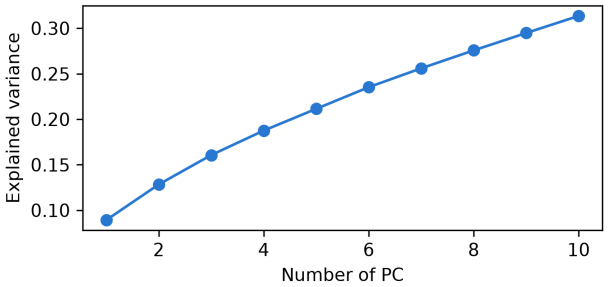
Feature Extraction

77 features

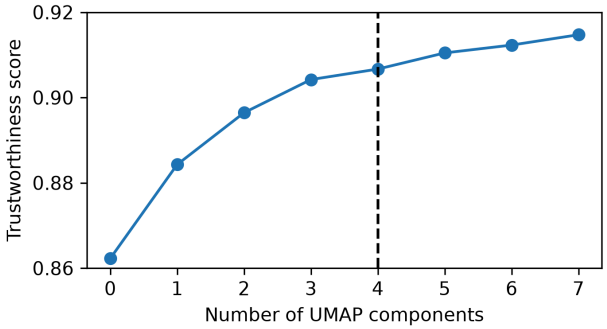
- Ligand features, e.g.
  - `rmsd`
  - `dist_serOG_carbonylC`
- Active site features, e.g
  - `pocket_backbone_rmsd`
  - `oxyanion_gln106N_ligO`
- Global protein features, e.g
  - `glob_protein_ca_rmsd`

Dimensionality Reduction

- PCA → 10 PC describes <50% of variance



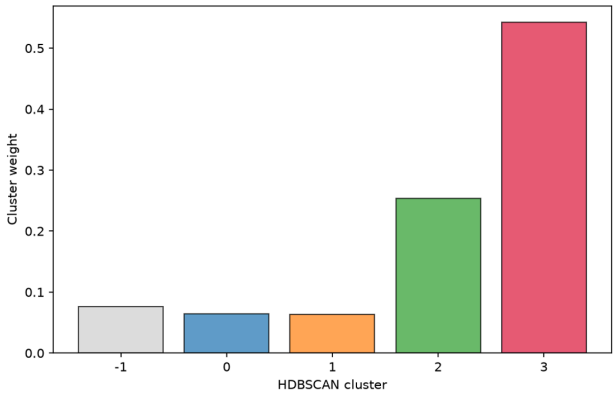
- UMAP → more effective. Number of components chosen by trustworthiness score



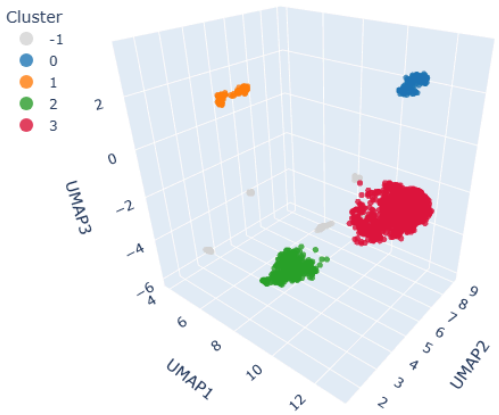
Clustering

Reduced data was clustered using HDBSCAN

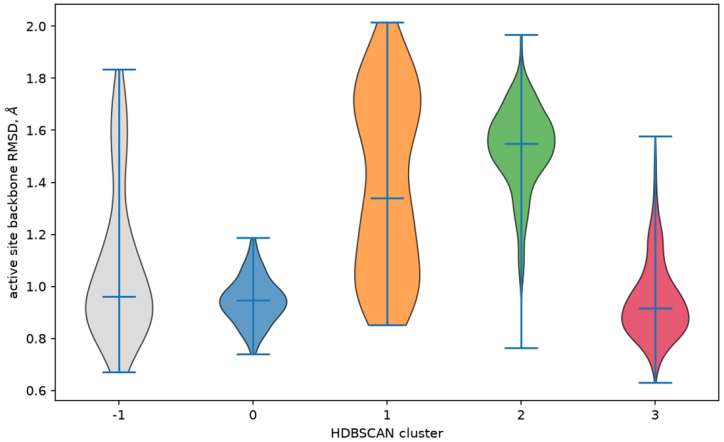
- `min_cluster_size = 100`
- 4 clusters
- 7.6% frames – noise



Clustering results



Clusters in $\langle \mathrm{UMAP1}, \mathrm{UMAP2}, \mathrm{UMAP3} \rangle$ space

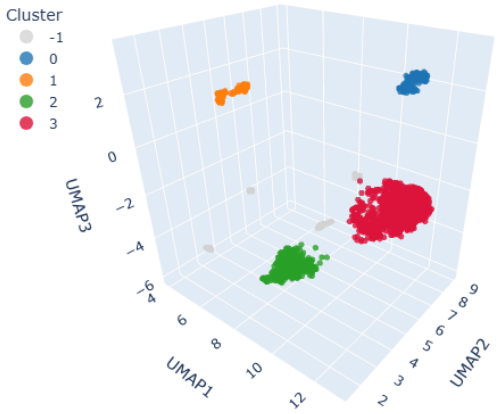


Active site backbone RMSD distribution

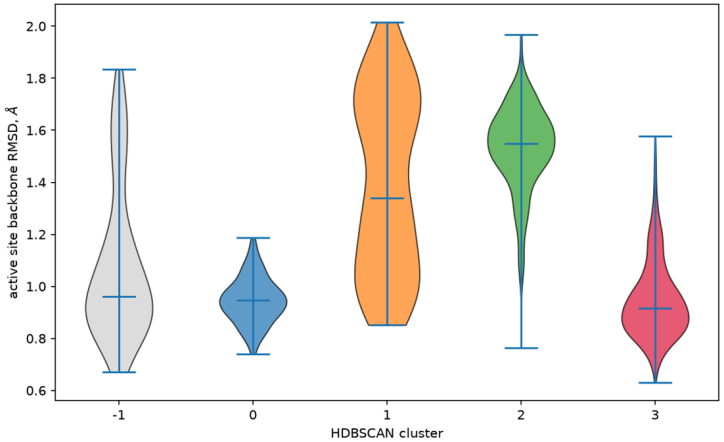
UMAP dimensionality reduction and HDBSCAN clustering allows to efficiently analyze MD-trajectories. The most noticeable difference between the clusters – RMSD of active site backbone atoms.

In terms of other individual features there are no noticeable differences.

Clustering results



Clusters in $\langle \mathrm{UMAP1}, \mathrm{UMAP2}, \mathrm{UMAP3} \rangle$ space

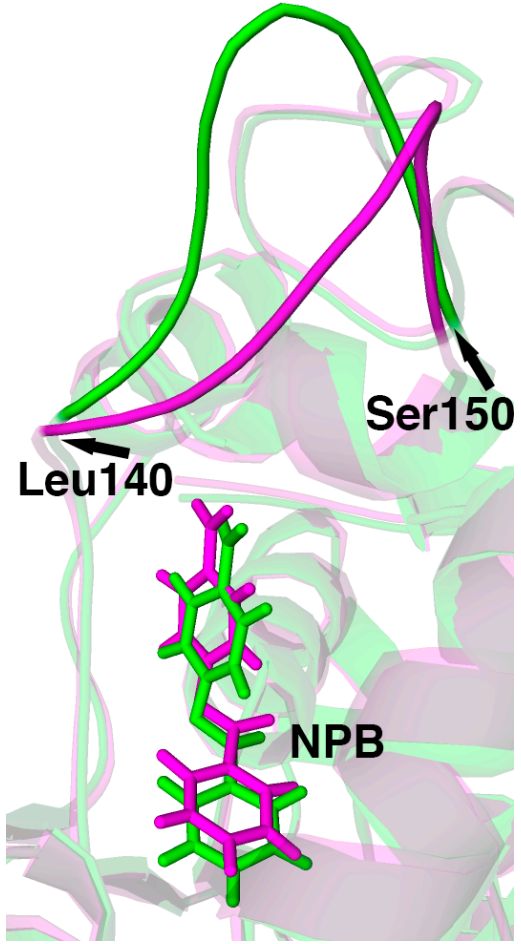


Active site backbone RMSD distribution

UMAP dimensionality reduction and HDBSCAN clustering allows to efficiently analyze MD-trajectories. The most noticeable difference between the clusters – RMSD of active site backbone atoms.

In terms of other individual features there are no noticeable differences.

Centroids differs in the position of the ligand and the conformation of the loop T140-S150.



Structures in the centroids of clusters

Conclusion

- The proposed pipeline offers a semi-automated, and suitable workflow for processing MD trajectories of serine hydrolase ES complexes.
- By efficiently filtering through the flexible ES landscape and identifying stable, highly populated conformations, developed tool is useful to select representative starting points for subsequent high-level QM/MM simulations and covalent inhibitor design
- Active site geometry depends on the mobility of neighbouring protein elements. The clusters differ in how the ligand pocket and the oxyanion hole are arranged, and this comes from shifts in nearby helices and loops, not only from substrate motion.
- The conformational landscape of the ES complex is essentially non-linear. PCA retains less than 50% of the variance even at 10 components, so analysing such systems requires non-linear dimensionality reduction such as UMAP.

This research was funded by the Russian Science Foundation (grant number 25-23-00965). The research was carried out using the equipment of the shared research facilities of high-performance computing resources at Lomonosov Moscow State University including Istok computing system (Agreement 075-15-2025-541)