LOSCHMIDT LABORATORIES

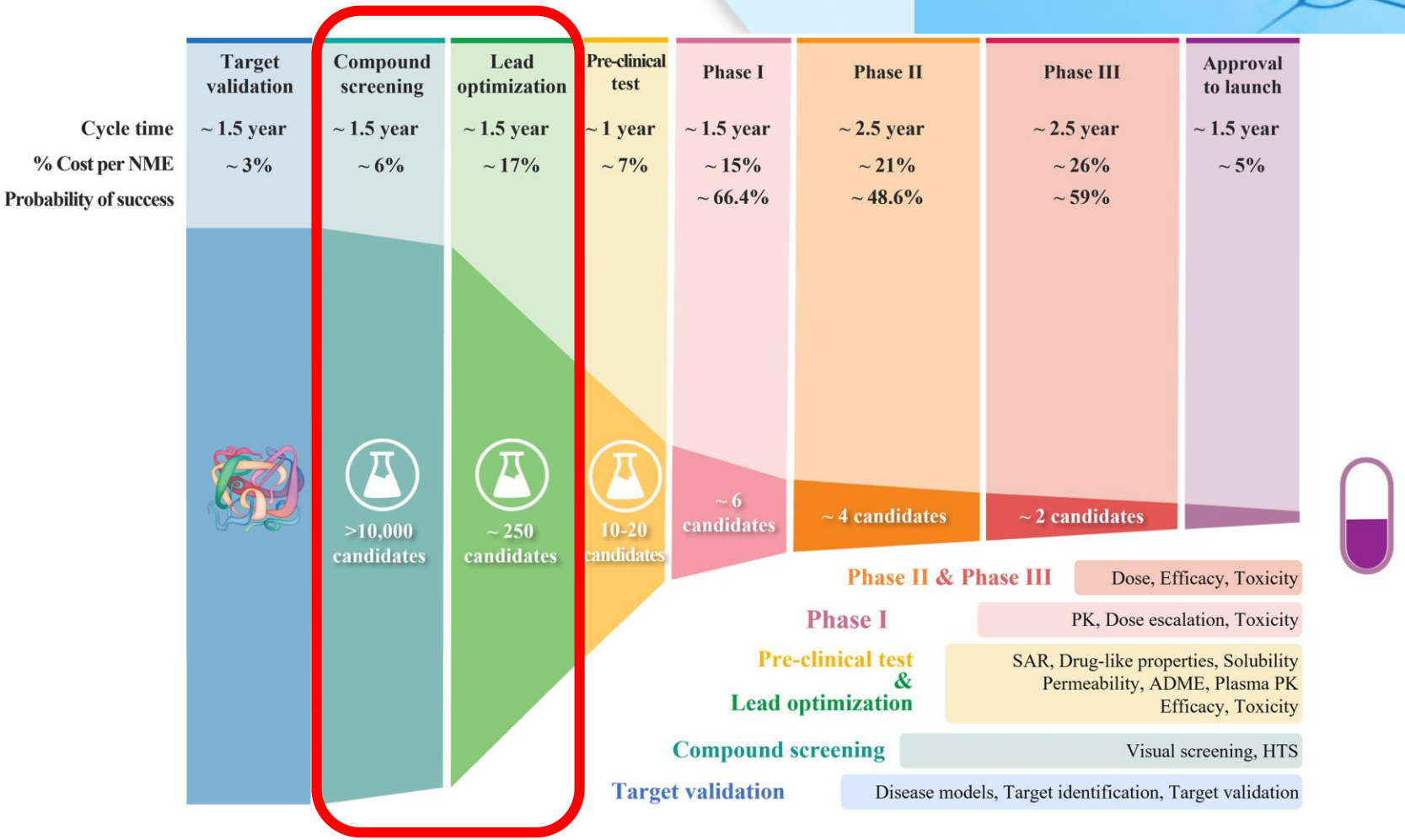
Research Data Day & EOSC National Tripartite Event

Quantum Mechanics-Enhanced Data Generation for ML-Driven Drug Discovery

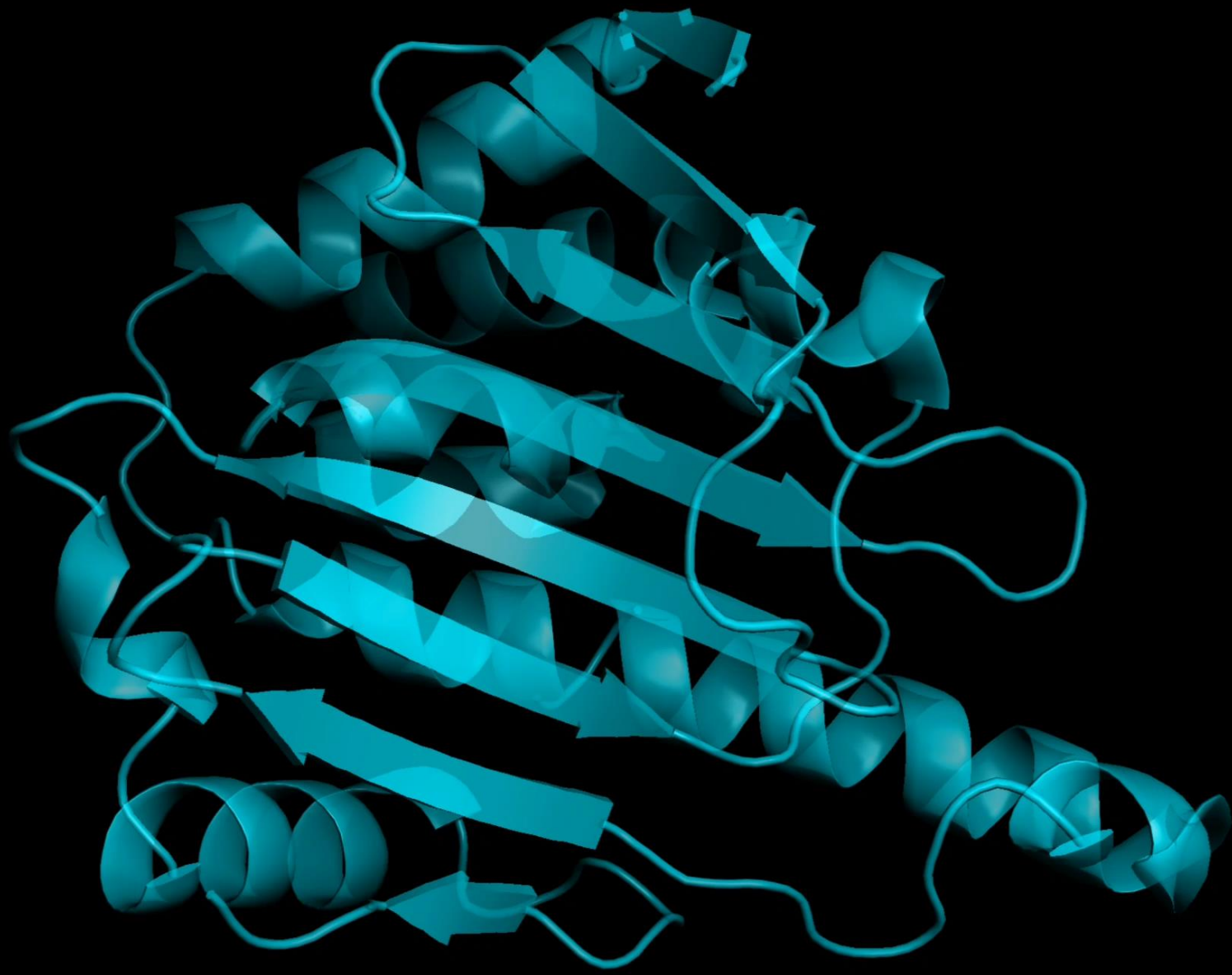
SALTUK M. EYRILMEZ, Ph.D

22.05.2025

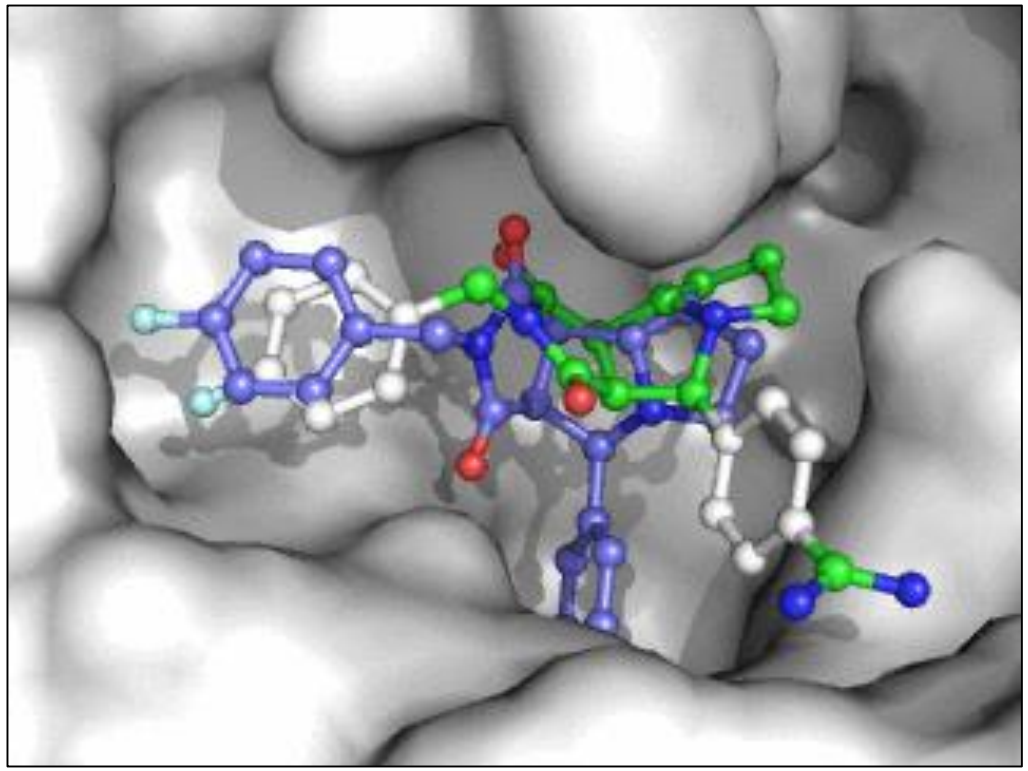
The Timeline of Drug Discovery



D. Sun, *et all*, *Acta Pharmaceutica Sinica B*, vol. 12, no. 7, pp. 3049-3062, 2022.



Molecular Docking



Accuracy:

↘ System size

- Binding site ↔ Search space

↘ Small molecule Complexity

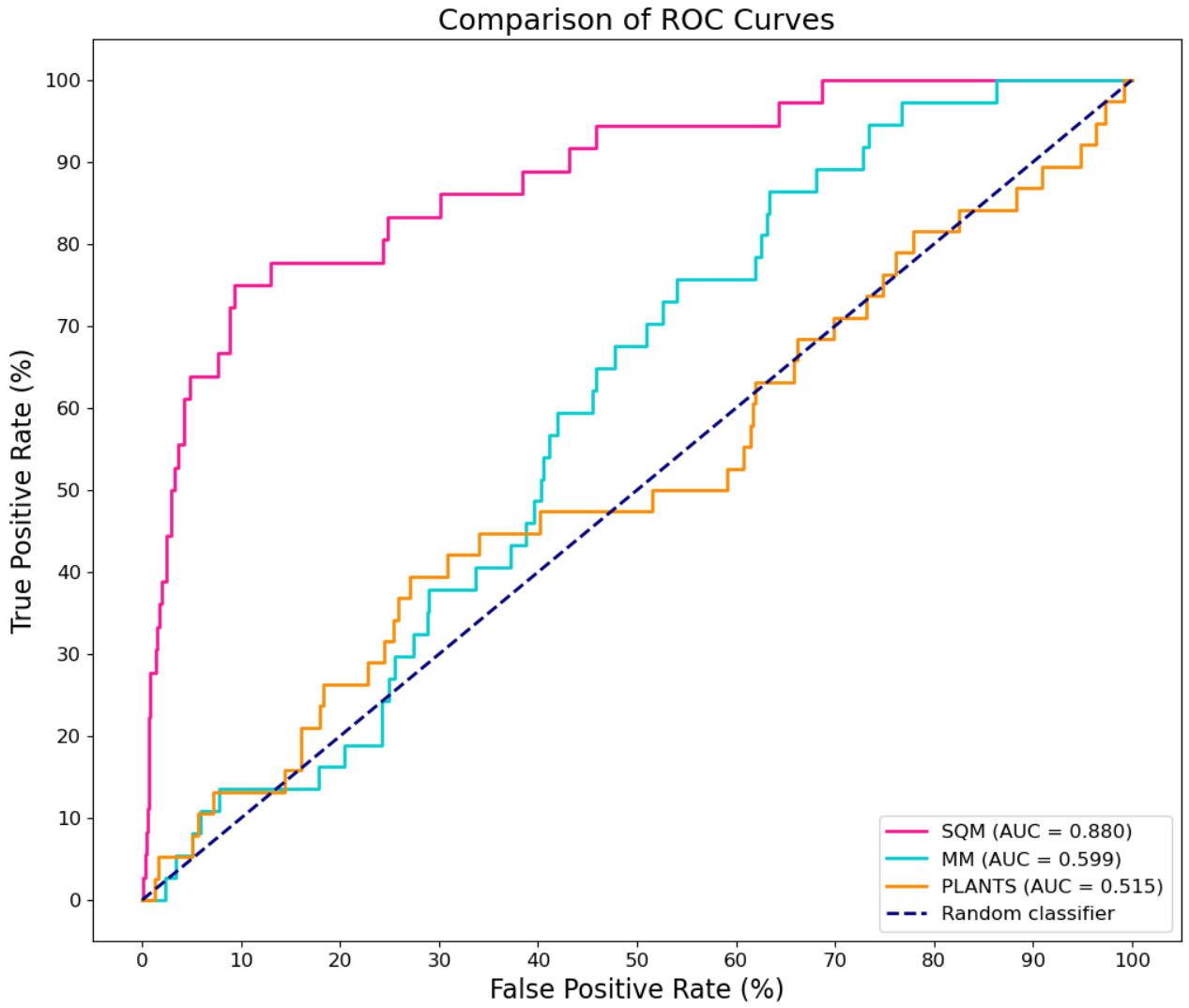
- Charges
- Rotatable Bonds

↘ Binding site complexity

- Protein residue compositions
- Flexible side-chains
- Flexible loops

Image source: Ligand docking with Smina | Cheminformania

RESULTS



- Benchmarking on challenging targets

Datasets:

- 40 active compounds VS. 1200 decoys

- QM methods significantly outperform classical methods.

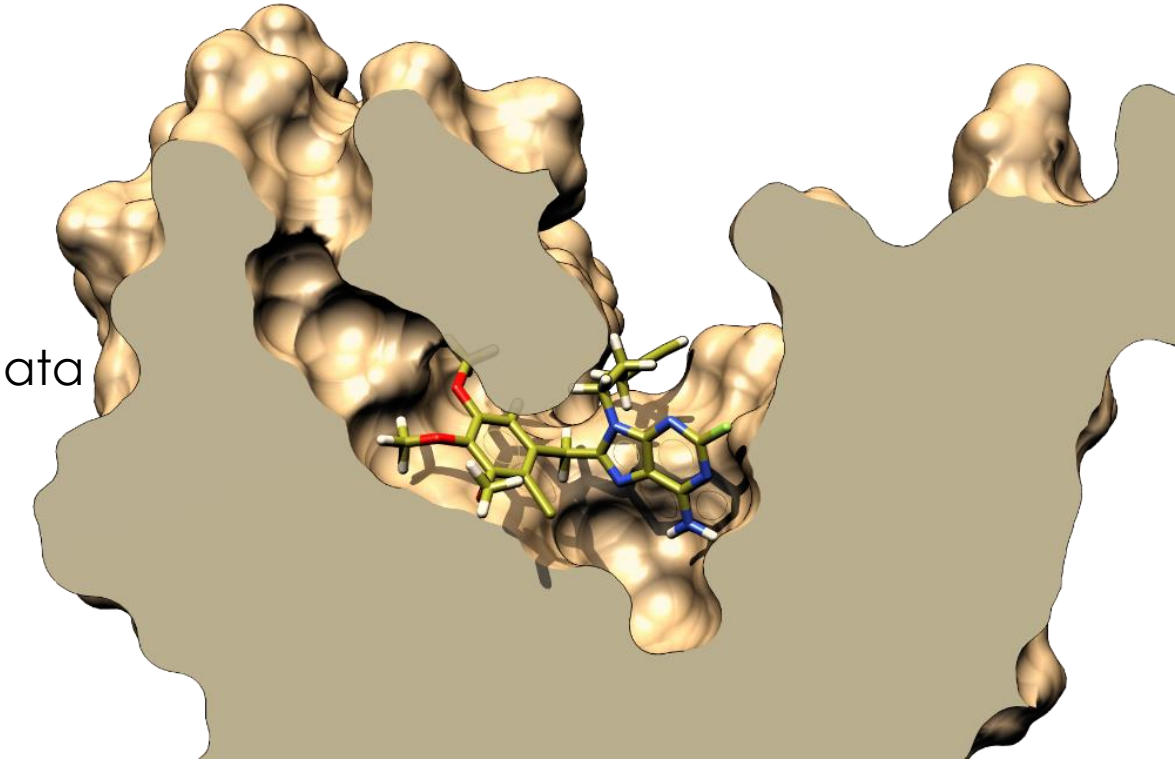
Data Quality/Quantity Challenge

- **Limited coverage:**
  - Estimated drug-like compounds: ~10^{33}-10^{60}
  - Current Coverage: ~10^{14}
- **Experimental Inconsistencies:**
  - Different labs, methods, conditions …
- **Reproducibility:**
  - Issues with the validation of published binding data

- **Bias toward positives:**
  - Successful binding data overrepresented
  - Failed interactions are rarely published

- **Impact on AI Models:**

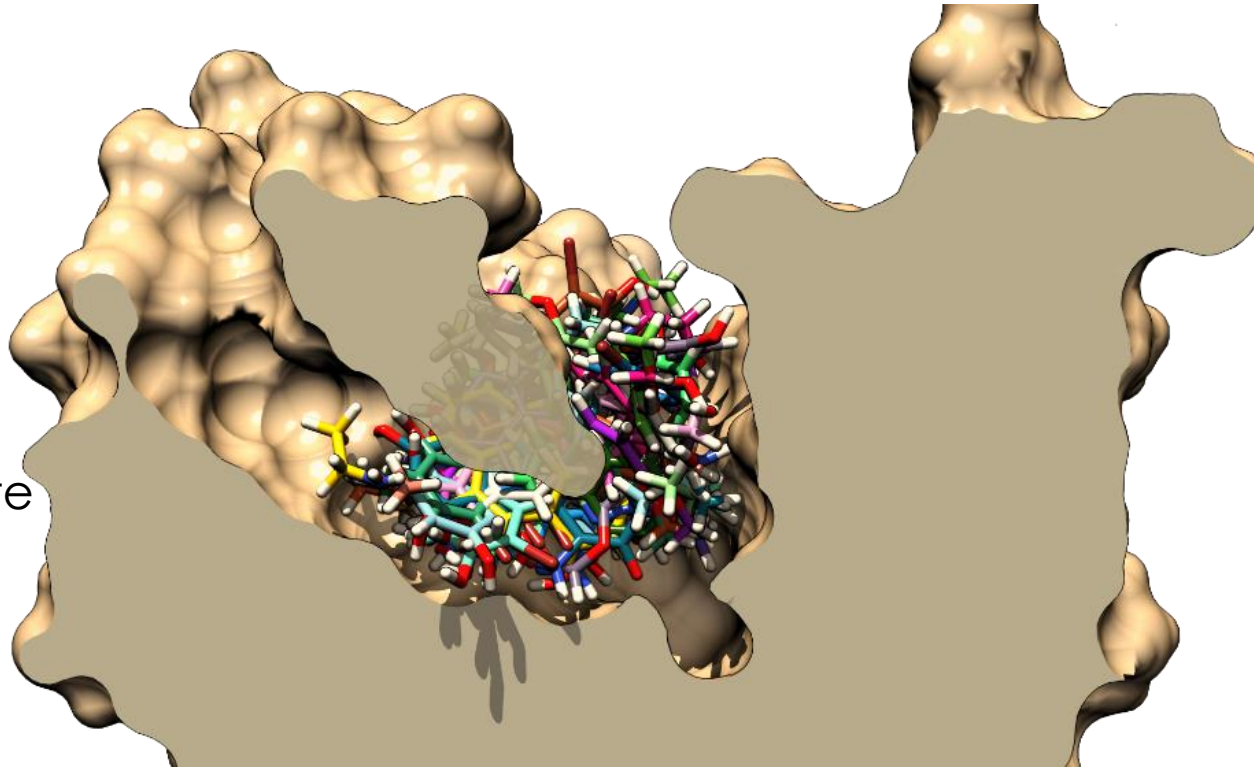
  - Models learn dataset artifacts rather than fundamental molecular principles
  - Cannot generalize effectively to novel chemical structures or to novel protein targets



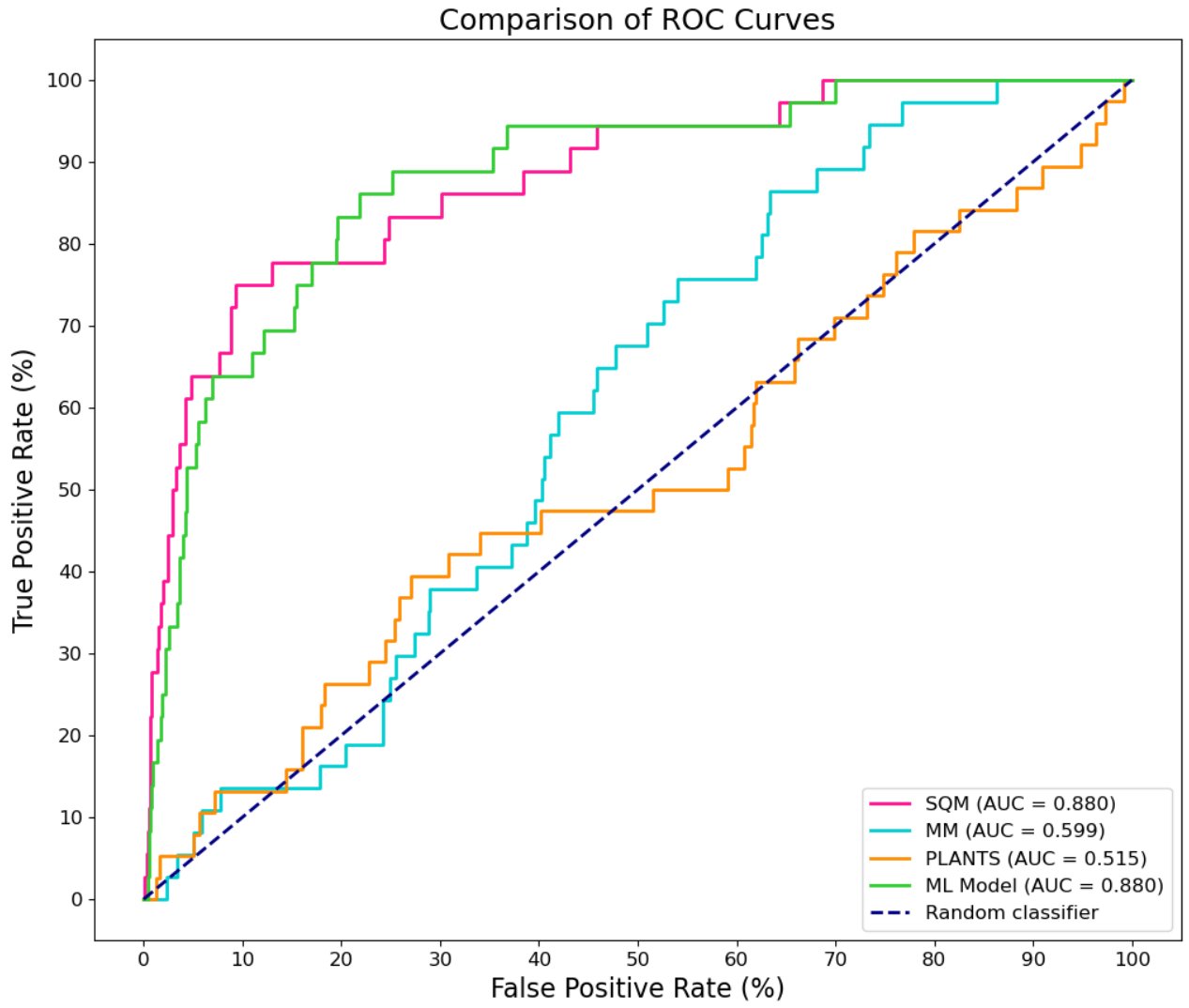
From Process to Prediction: Quantum Mechanics-Enhanced Solution

Our Methodology:

- Generating multiple binding states for training ML models
- Training exclusively on small molecule libraries with no overlap to test compounds
- Quantum mechanical calculations for accurate energy landscapes
- High-performance computing enabling quantum mechanical-level data generation and training



Results:
Model Performance and Validation



- **No overlap of test set and training set**
- Orders of magnitude faster:
  - ~300,000 times!
- Successfully learned how to "parallel park"

Disease-Specific Applications

Alzheimer's Disease Research

- APOE4 Protein
- Challenging Target: No certain Cavity

Brain Metabolomes
Function

Small Molecules
Correction

Cancer-Related Drug Discovery Research

- Breast Cancer
- Melanoma

Perspectives

- A library of ML models on key targets
- Target-specific hit identification (less or no side effects)
- Reduce the time and costs
- Tools for generating training sets for custom models.

- True generalization

LOSCHMIDT LABORATORIES

CAVER

CAVERDOCK

CAVER WE3

HOTSPOT WIZARD

CALFITTER

FIREPROT

FIREPROT ASR

AGGRE PROT

ENZYME MINER

T_MPROT

LOOP GRAFTER

SOLUPROT

Thank you for your attention!

CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE

- Josef Šivic
- Jiří Sedlář
- Jakub Kopko
- Ondřej Bouček

LOSCHMIDT LABORATORIES

MUNI | RECETOX

FNUSA ICRC
ST. ANNE'S UNIVERSITY HOSPITAL BRNO
INTERNATIONAL CLINICAL RESEARCH CENTER

Clara

cesnet
metacentrum

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

LUMI
www.lumi-supercomputer.eu