UGM Boston Oct. 4th 23

Chemaxon

The World's Largest Protein-Ligand Complex and Binding Affinity Dataset for Data Driven Methods in Drug Design

Dr. Arun Subramaniyan
Vice President Cloud & AI,
Strategy & Execution, Intel

UGM Boston
Oct. 4th 23
Chemaxon

AI^3: Creating The Largest Protein-Ligand Complex (PLC) Dataset to Accelerate Drug Discovery

- Amazon Web Services (AWS)
- International Institute of Information Technology Hyderabad (IIIT-H)
- Intel
- Insilico Medicine

Presenter:
Dr. Arun Subramaniyan
Vice President, Cloud & AI Strategy, Intel

Thanks to the team that made the magic happen

Alex Iankoulski
Yusong Wang
Stephen Litster
Srinivas Tadepalli

AWS

Rakesh Srivastava
Prathit Chatterjee
Divya Korlepara
Vasavi C.S.
Suyash Gupta
Vishal Kumar
Pradeep Kumar Pal
Aathira Nair

Shivangi Verma
Harshini Anand
Saswati Mallick
Kavita Thakran
Parth Kanani
Indhu Ramachandran
Divya Nayar
U. Deva Priyakumar

IITH

Faird El Chalouhi
Chetan Rao
Arun Karthi Subramaniyan
Akanksha R. Bilani
Varma S. Konala
Ramanathan Sethuraman

Intel

Vladimir Aladinskiy
Evgeny Kirilin
Georgiy Andreev
Arkadii Lin
Eugene Babin

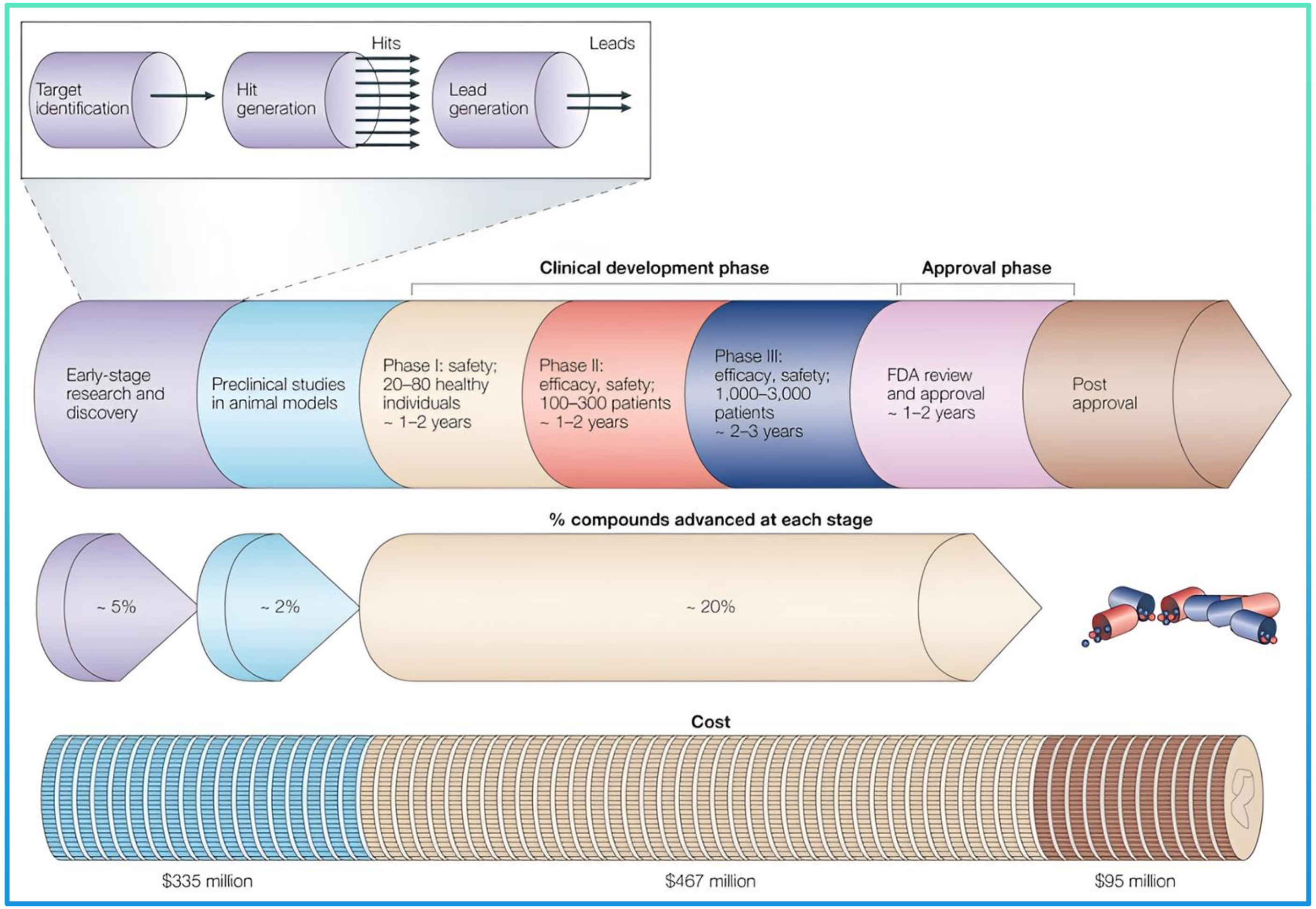
Insilico Medicine

AI^3

Introduction

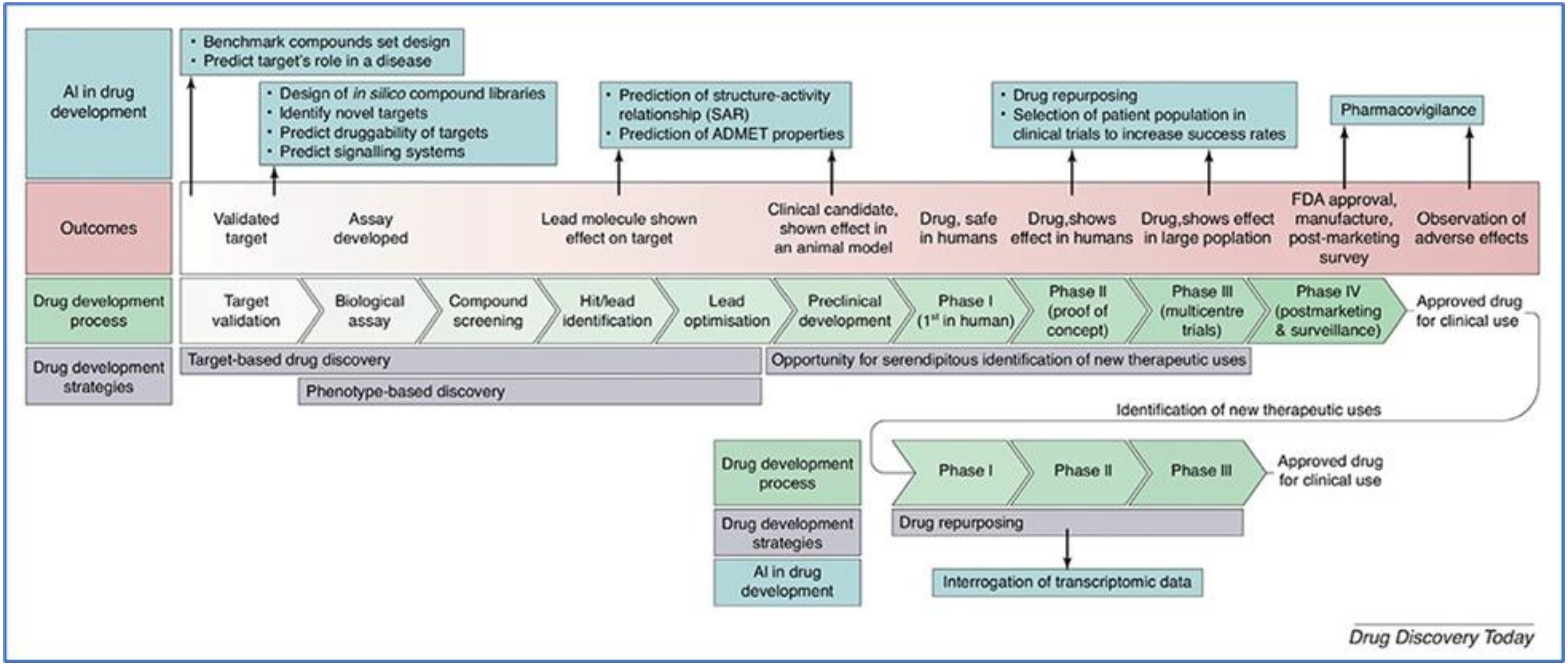
Inspiration

- Traditional drug discovery is costly, time-consuming, and has a low success rate.
- Computational techniques are crucial for revolutionizing drug discovery workflows.
- Recent advances in cloud computing and AI/ML could help accelerate drug discovery.


Hits
Leads
Target identification
Hit generation
Lead generation
Clinical development phase
Approval phase
Early-stage research and discovery
Preclinical studies in animal models
Phase I: safety; 20–80 healthy individuals ~ 1–2 years
Phase II: efficacy, safety; 100–300 patients ~ 1–2 years
Phase III: efficacy, safety; 1,000–3,000 patients ~ 2–3 years
FDA review and approval ~ 1–2 years
Post approval
% compounds advanced at each stage
~ 5%
~ 2%
~ 20%
Cost
$335 million
$467 million
$95 million

Traditional drug discovery workflow (dLab, 2023)

Potential Applications of AI in Drug Discovery Process

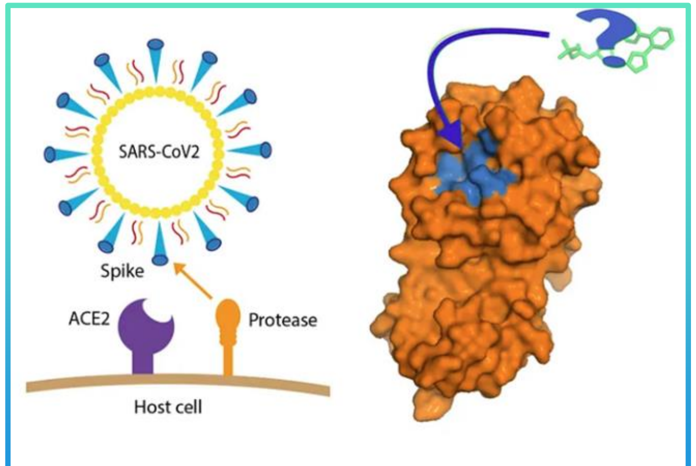

AI in drug development
Outcomes
Drug development process
Drug development strategies
• Benchmark compounds set design
• Predict target's role in a disease
• Design of *in silico* compound libraries
• Identify novel targets
• Predict druggability of targets
• Predict signalling systems
• Prediction of structure-activity relationship (SAR)
• Prediction of ADMET properties
• Drug repurposing
• Selection of patient population in clinical trials to increase success rates
Pharmacovigilance
Validated target
Assay developed
Lead molecule shown effect on target
Clinical candidate, shown effect in an animal model
Drug, safe in humans
Drug,shows effect in humans
Drug,shows effect in large poplation
FDA approval, manufacture, post-marketing survey
Observation of adverse effects
Target validation
Biological assay
Compound screening
Hit/lead identification
Lead optimisation
Preclinical development
Phase I (1st in human)
Phase II (proof of concept)
Phase III (multicentre trials)
Phase IV (postmarketing & surveillance)
Approved drug for clinical use
Target-based drug discovery
Phenotype-based discovery
Opportunity for serendipitous identification of new therapeutic uses
Identification of new therapeutic uses
Drug development process
Drug development strategies
AI in drug development
Phase I
Phase II
Phase III
Approved drug for clinical use
Drug repurposing
Interrogation of transcriptomic data
Drug Discovery Today

In-Jin, 2019, https://tcpharm.org/DOIx.php?id=10.12793/tcp.2019.27.3.87

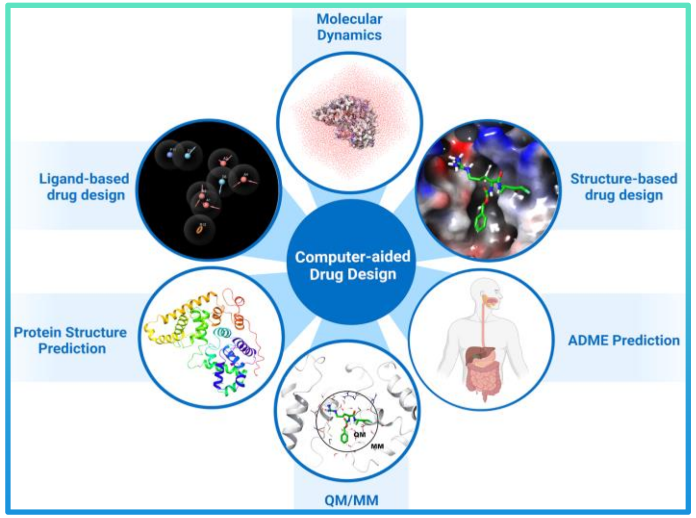
Importance of PLCs in Computer-aided Drug Design

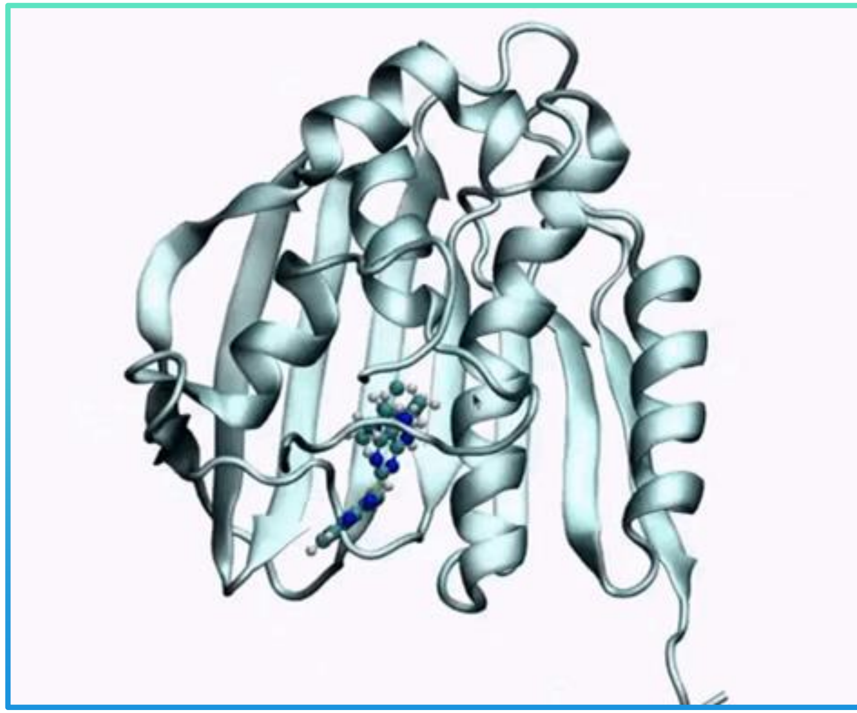
Protein-ligand Complexes (PLCs) & their role in Drug Design

- Proteins are essential biological molecules with diverse structures and functions.
- Ligands, or small molecules, can alter or assist protein structure and function by binding to proteins.
- Drug design often involves tuning druggable molecules to interact energetically with protein binding sites.
- Predicting binding affinity in PLCs is challenging but crucial for drug design.
- In-silico methods reduce production costs and enable the study of inaccessible molecular interactions



Left — SARS-Cov-2 capsid approaches a host cell; right — a spatial structure of the protease, active site (blue), candidate inhibitor (green).


Molecular Dynamics
Ligand-based drug design
Structure-based drug design
Computer-aided Drug Design
Protein Structure Prediction
ADME Prediction
QM
MM
QM/MM



A Guide to In Silico Drug Design, Chang, 2022

Popular Existing Datasets & Limitations

PDBbind (2004 onwards, 23,496 PLC entries; http://www.pdbbind.org.cn/)

DUDE (2012 onwards, 22,886 active compounds; https://dude.docking.org/)

ONIONnet (2019 onwards; https://pubs.acs.org/doi/10.1021/acsomega.9b01997)

BindingDB (2007 onwards; https://www.bindingdb.org/rwd/bind/index.jsp)

AffinDB (2006 onwards; https://academic.oup.com/nar/article/34/suppl_1/D522/1132614)

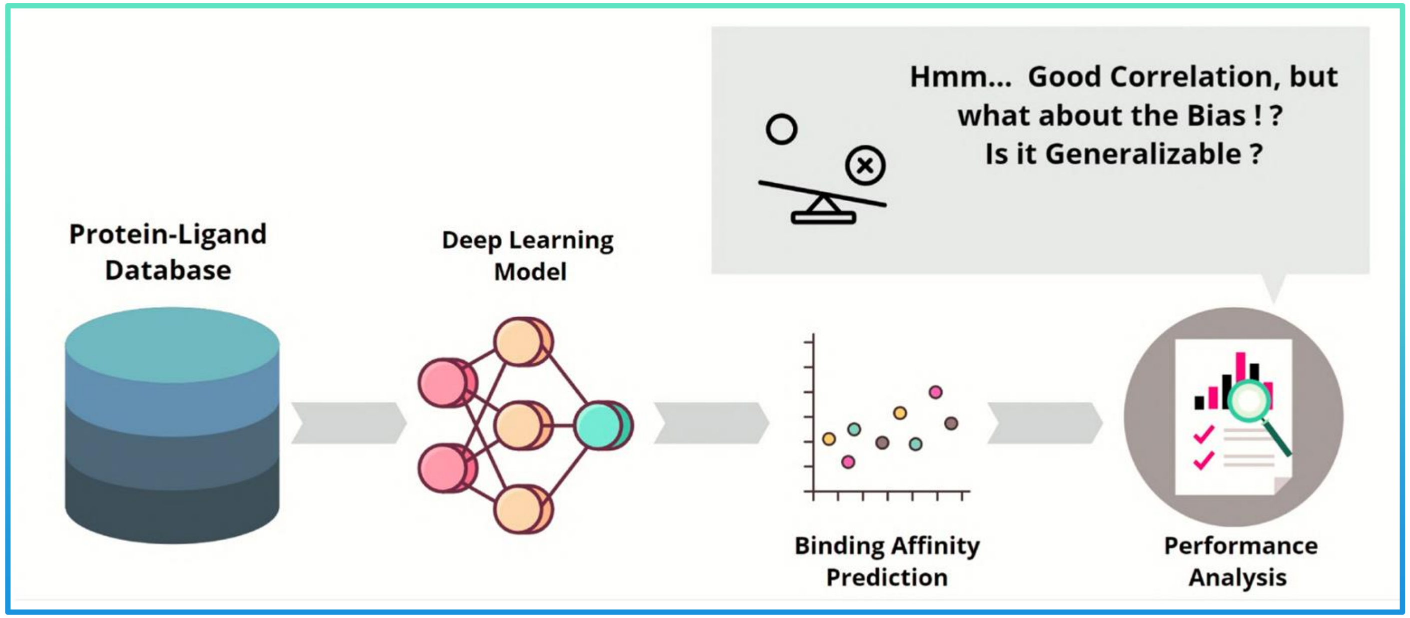
Binding MOAD (2005 onwards, 41409 structures; http://www.bindingmoad.org/)

Limitations

- Poor target and ligand diversity
- Low transferability to broad drug targets
- Lack of high-energy data (both structural & thermodynamic)
- Low volume
- High variability in validation quality – experimental errors from different labs & time

Motivation behind calculating PLC Binding Affinity

- Machine learning based scoring functions, for predicting binding affinity, have acceptable evaluation scores.
- Yet, they fail to perform similarly in virtual screenings.
- Hidden biases plausibly originate from data obtained from different experimental protocols.
- Inspiration to create a homogeneously computed Binding Free Energy of PLCs.



Latent Biases in Machine Learning Models for Predicting Binding Affinities Using Popular Data Sets, Kanakala, 2023

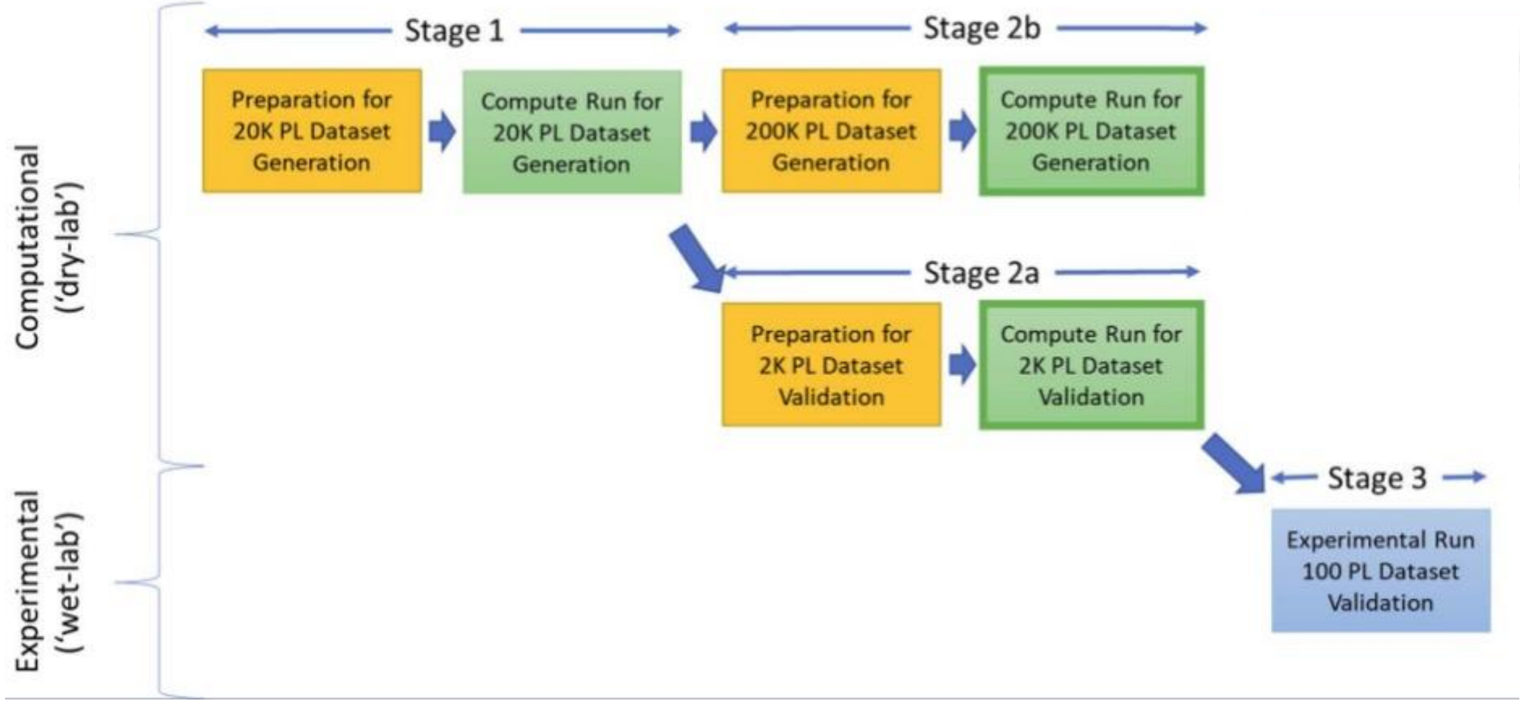
AI^3: Generate and Validate World’s Largest PLC dataset

- The World's Largest Open PLC Dataset (AI^3: OPLD) initiative aims to address corresponding dataset limitations.

- Collaboration between AWS, IIIT-H, INTEL,
  and Insilico Medicine.

- Phase 1 AI^3 dataset, consists of ~20,000 PLCs and corresponding binding affinity.

- Goal: Create a dataset with ~220,000 entries, including the negative examples.

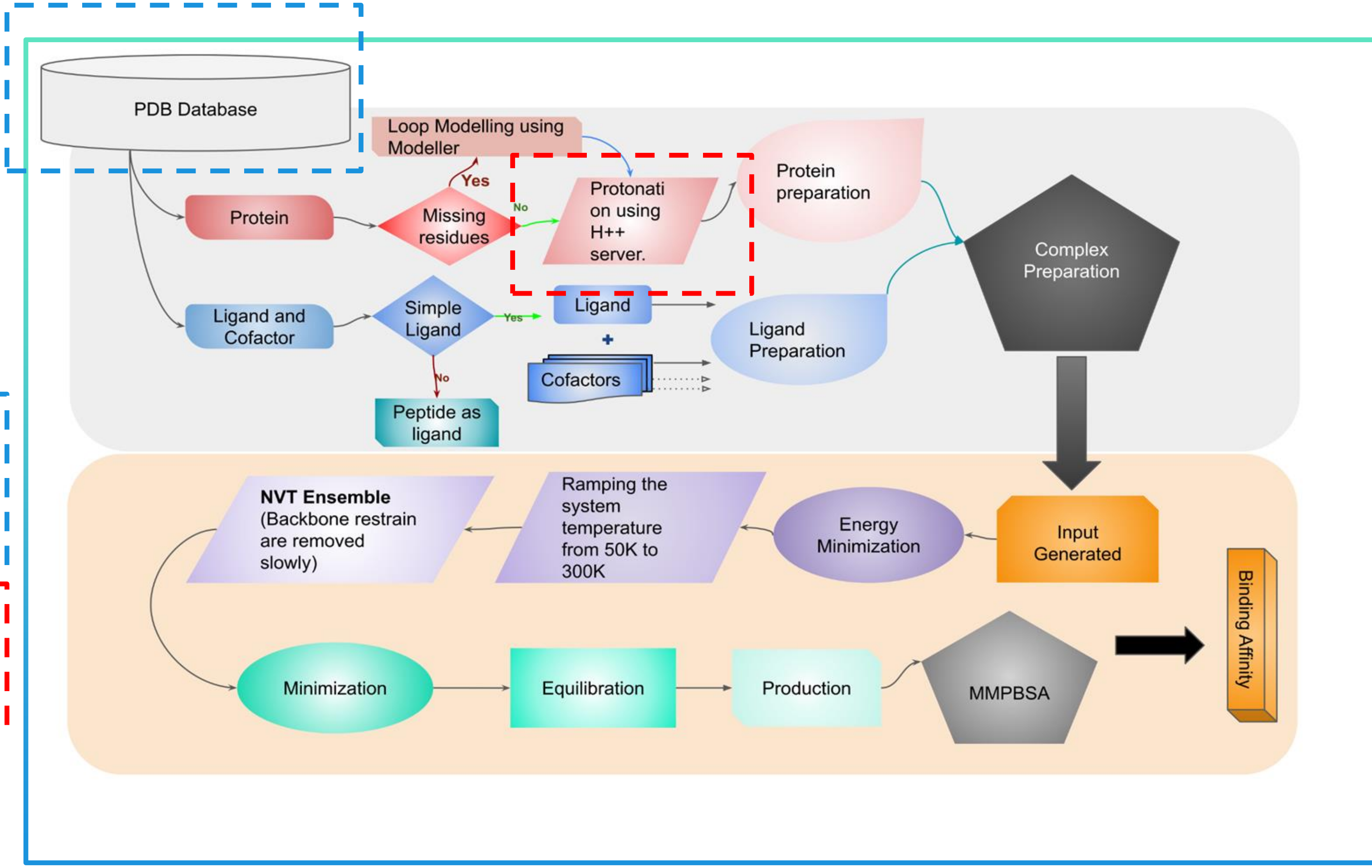

Computational ('dry-lab')
Experimental ('wet-lab')
Stage 1
Preparation for 20K PL Dataset Generation
Compute Run for 20K PL Dataset Generation
Stage 2b
Preparation for 200K PL Dataset Generation
Compute Run for 200K PL Dataset Generation
Stage 2a
Preparation for 2K PL Dataset Validation
Compute Run for 2K PL Dataset Validation
Stage 3
Experimental Run 100 PL Dataset Validation

Dataset Preparation

- The AI^3 (AWS-IIITH-Intel-Insilico) dataset is being prepared in two stages: Stage 1 (20 K bound PLCs) and Stage 2 (200 K unbound or partially bound PLCs).

- Protein structures are downloaded from RCSB PDB, and missing residues are modeled, modeler package UCSF Chimera.

- Protonation states are determined, from H++ server, at pH 7.4.

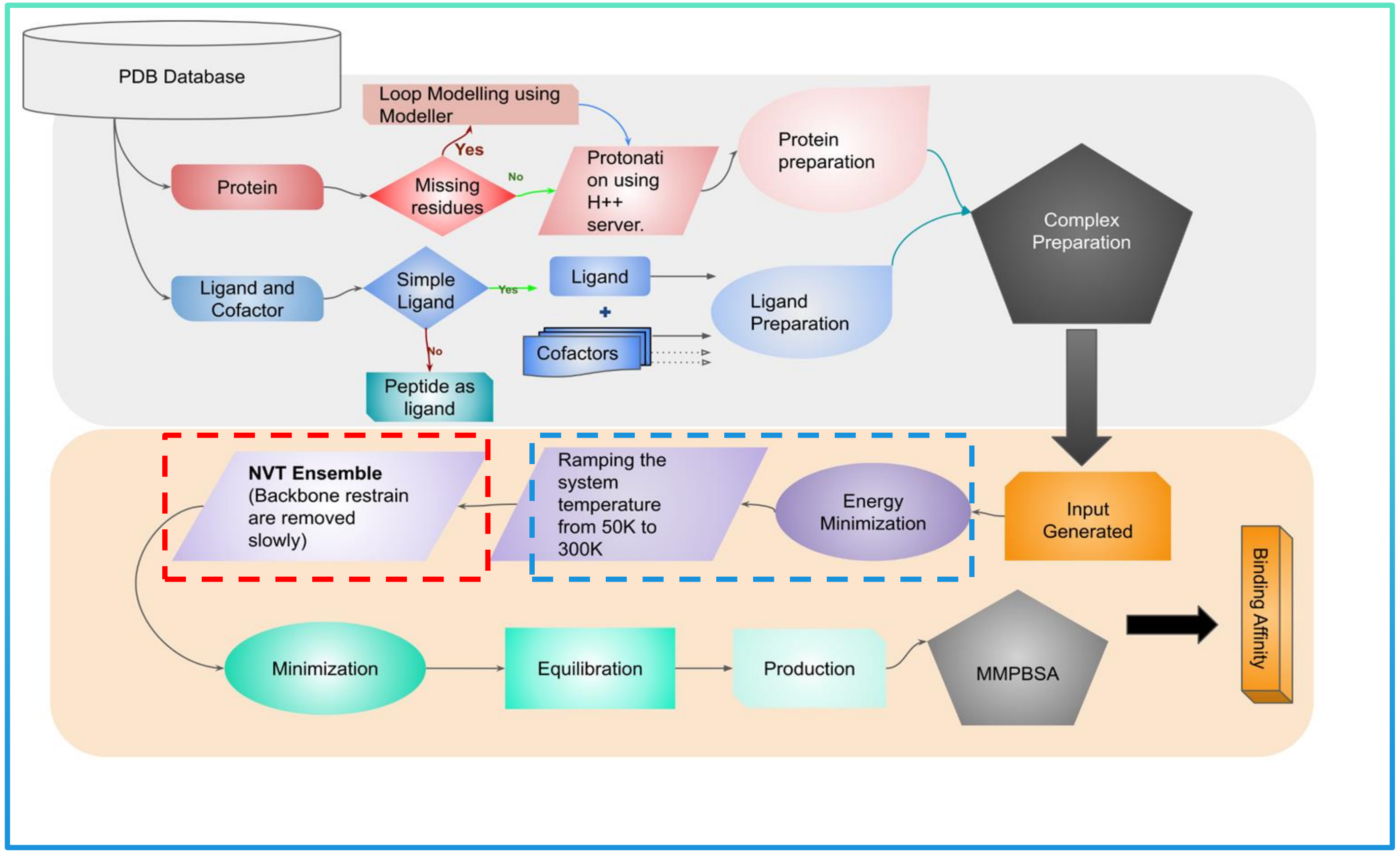


MD simulations protocols

- Amber ff1414 SB force field, TIP3P water model, GROMACS simulation package.

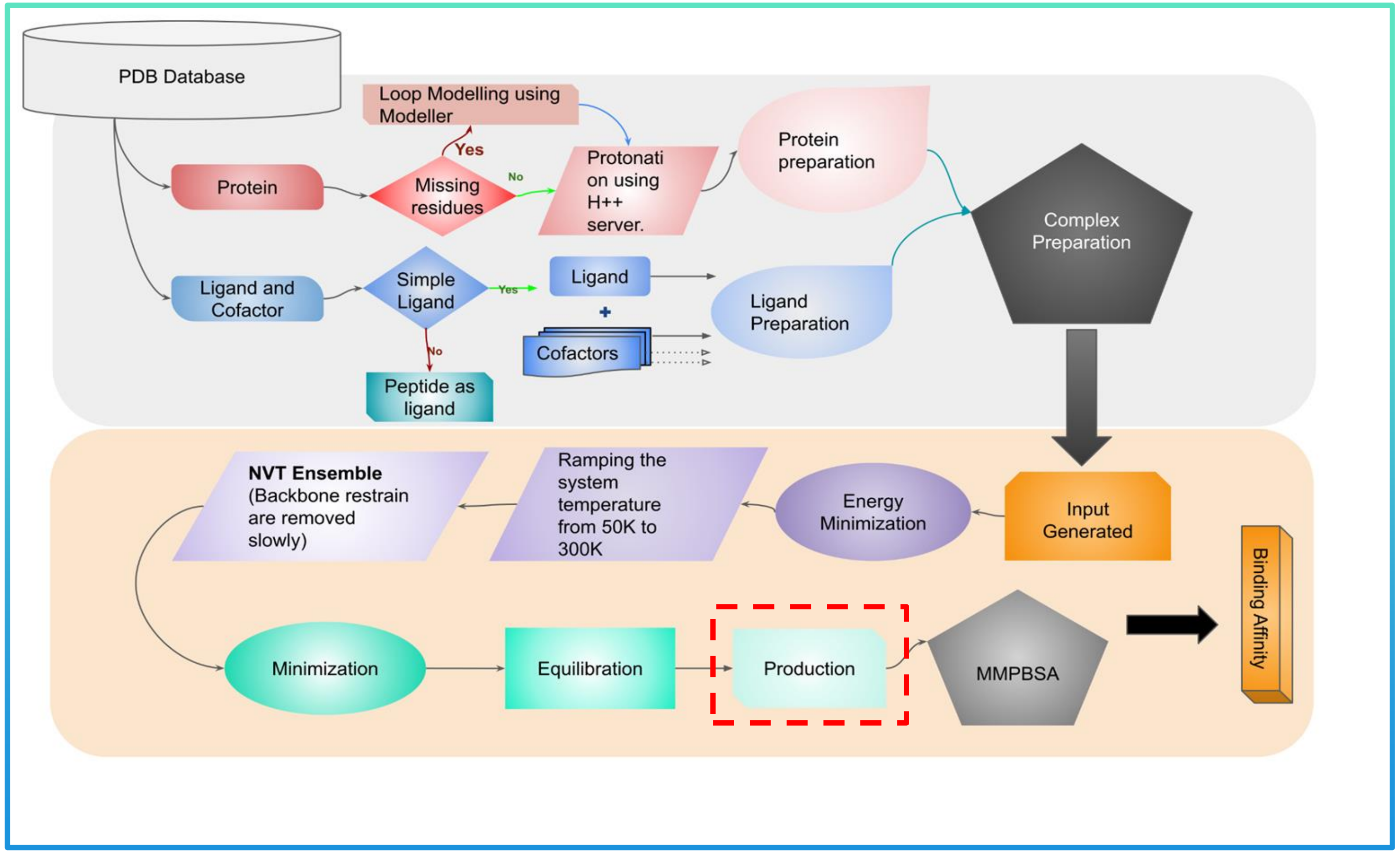
- Solvated PLC systems were subject to 2000 steps of steepest descent energy minimization and heating to 300 K.

- Backbone restrained were removed in a following NVT ensemble simulations, for 400 ps time length.



MD simulations protocols

- Multiple short independent simulations are conducted to reduce uncertainty in predicting binding affinities.
- Each independent production MD runs are carried out under NPT conditions at 300K for 6 ns.
- Simulation frames are saved at regular intervals for analysis. The last 4 ns of data are used for binding free energy calculations.

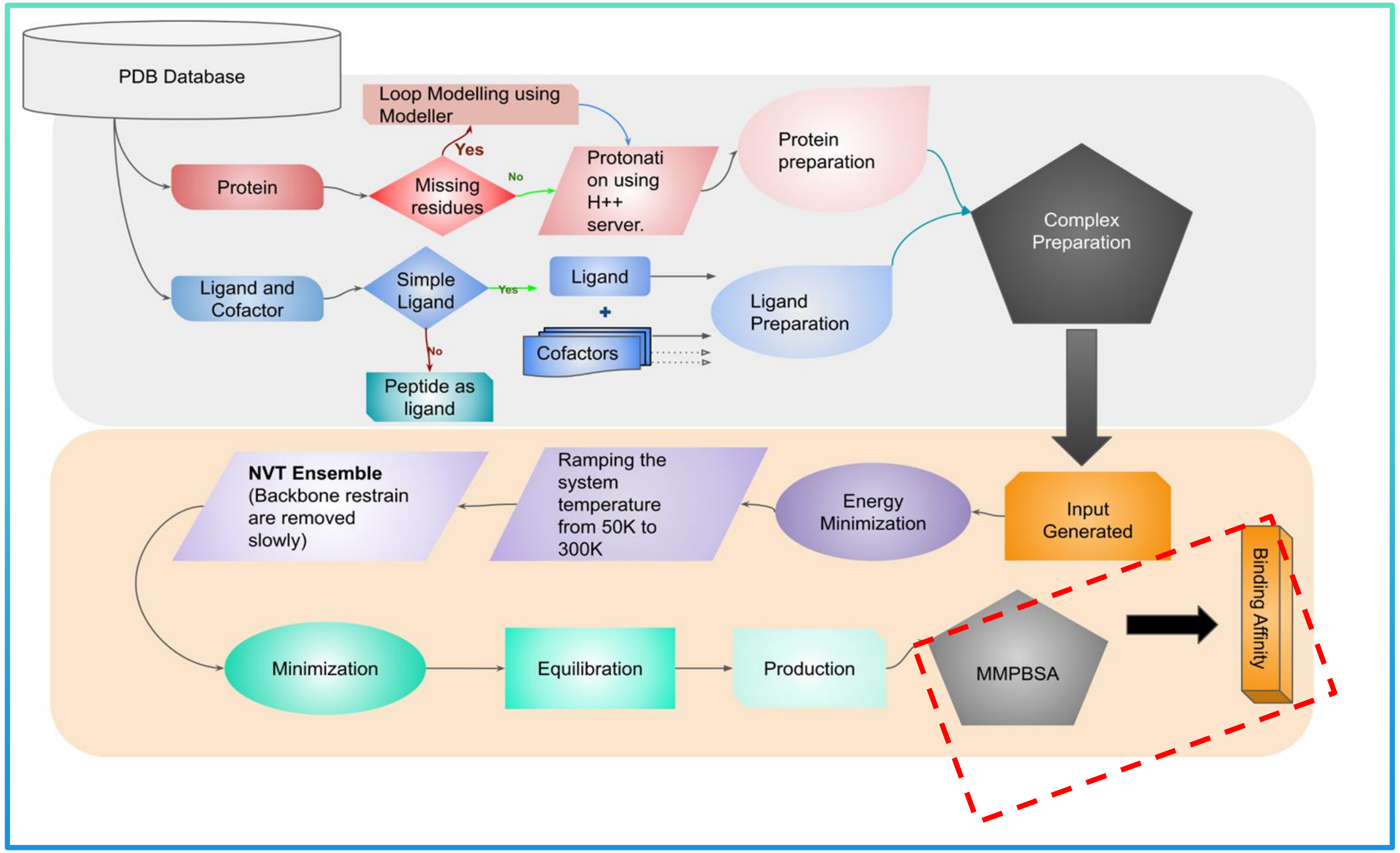


Binding Free Energy Estimation

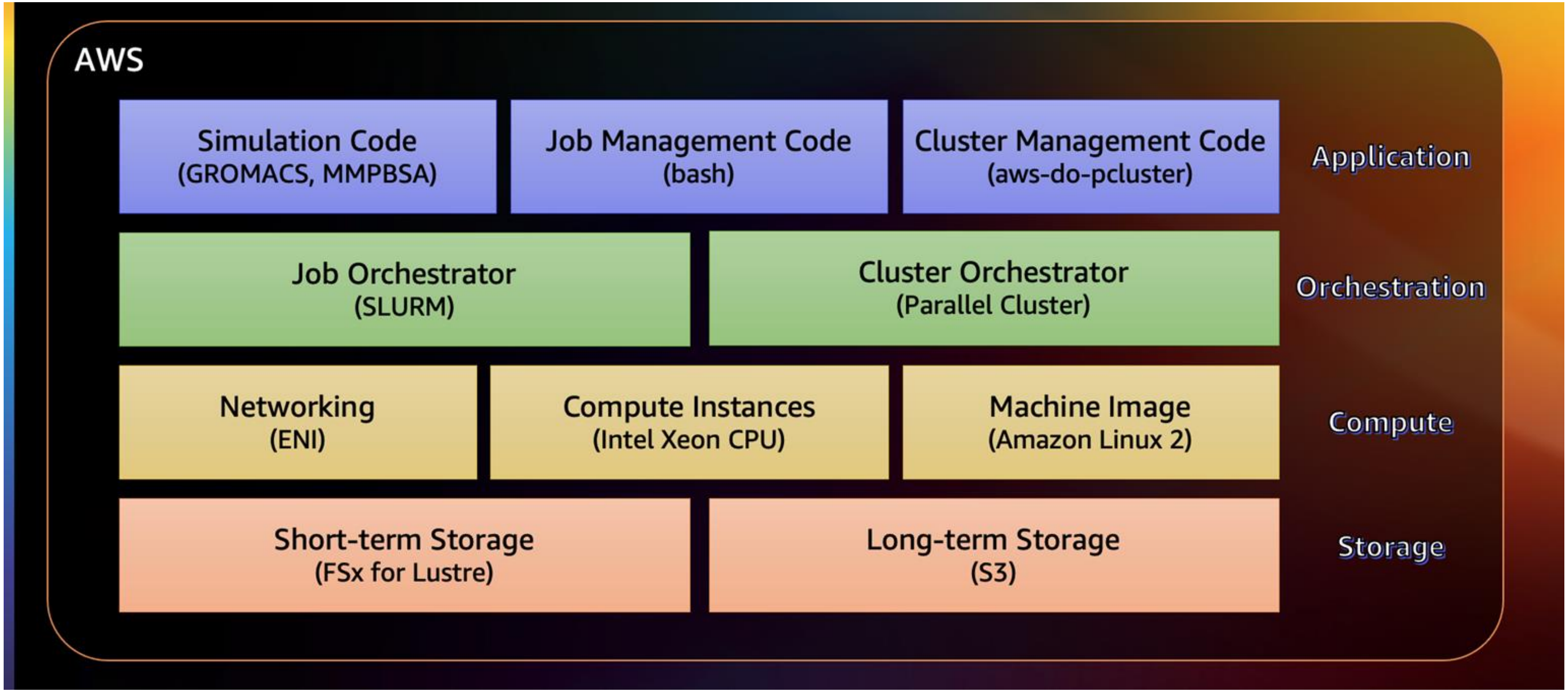
- The binding free energy is estimated using the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) approach.

- This method treats the solvent environment as a dielectric continuum. Polar and nonpolar solvation components are estimated.

- A single trajectory protocol is used to estimate binding affinities, ensuring robustness and accuracy.



Technology Stack and Architecture



AWS

Simulation Code (GROMACS, MMPBSA)

Job Management Code (bash)

Cluster Management Code (aws-do-pcluster)

Application

Job Orchestrator (SLURM)

Cluster Orchestrator (Parallel Cluster)

Orchestration

Networking (ENI)

Compute Instances (Intel Xeon CPU)

Machine Image (Amazon Linux 2)

Compute

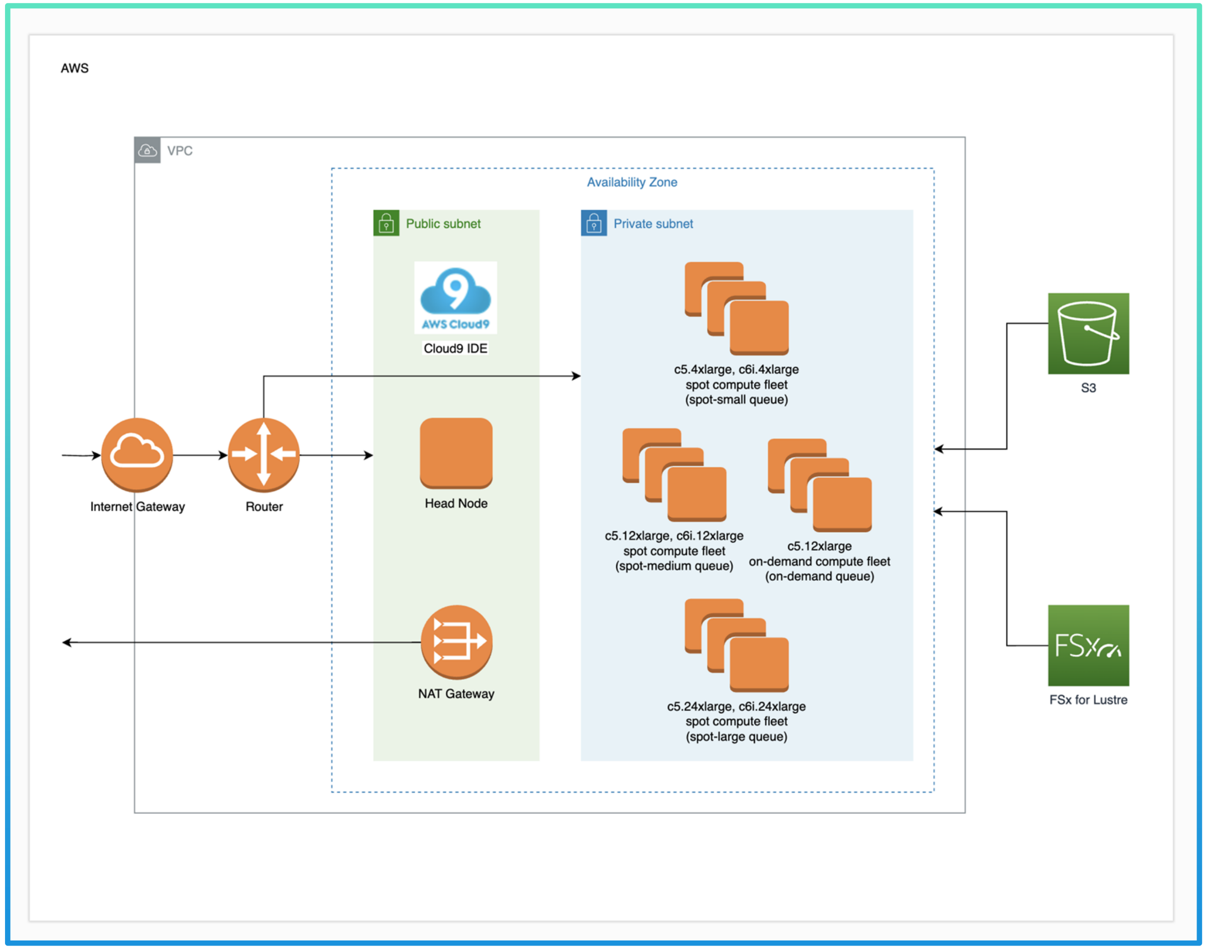
Short-term Storage (FSx for Lustre)

Long-term Storage (S3)

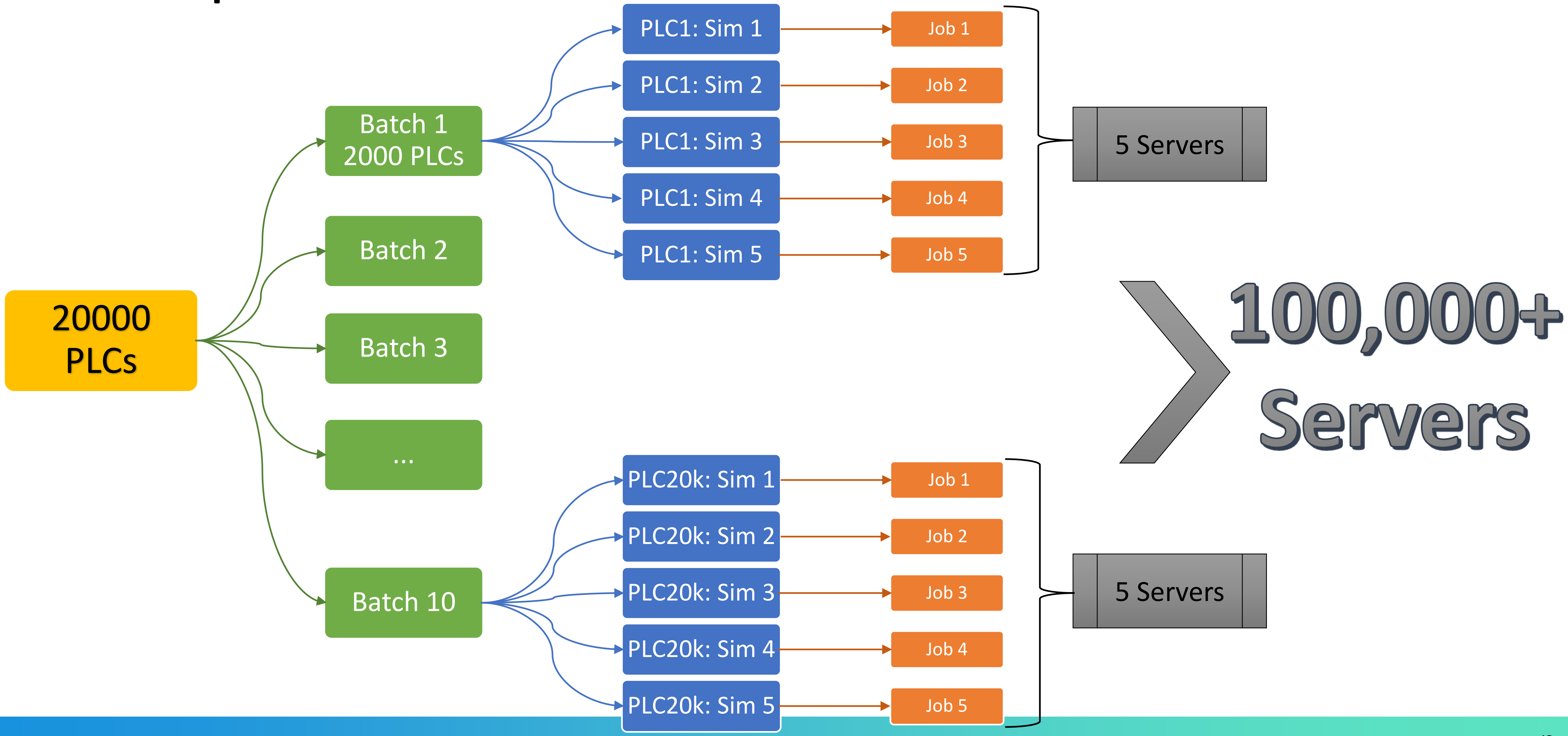
Storage

Execution and Deployment Architecture

Lower Bound (Number of Atoms)	Higher Bound (Number of Atoms)	Number of CPU Cores	Time to Complete Job
0	50,000	8	2-6 hrs
50,001	100,000	24	6-8 hrs
100,001	500,000	40	8-12 hrs



Leveraging AWS’s “Planetary-scale” Computing Footprint



Total Compute utilized … for Stage 1

100,000+ Servers

2.3M+ CPU cores

100M+ CPU core-hours

20000 PLCs

3 Weeks

Calculated Observables

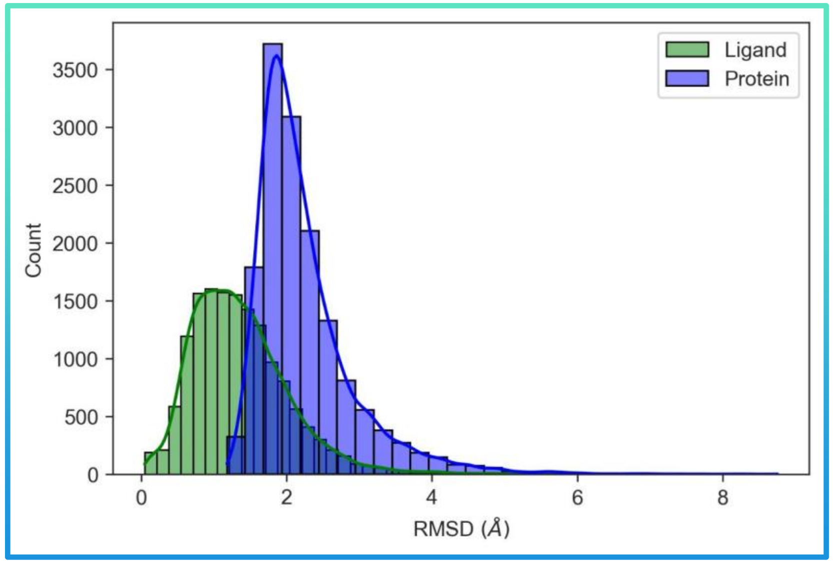
Stage 1: PLAS-20k as a Baseline Validated Dataset

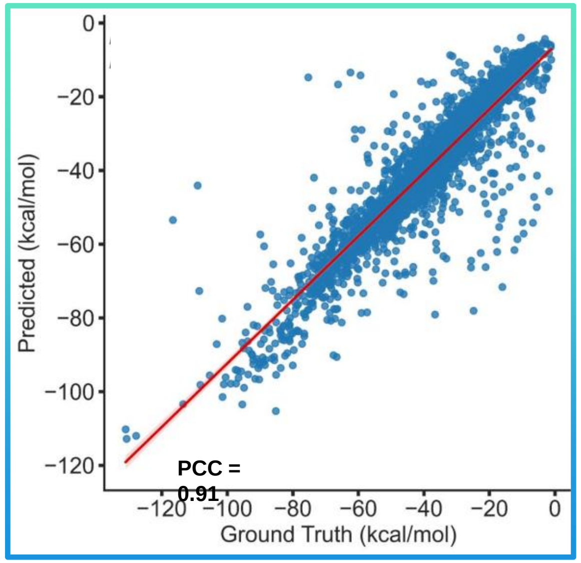
Training Deep-Learning Models with PLAS-20K

- **Objective**: Accurate prediction of Protein-Ligand (PL) complex binding affinity.
- **Method**: Utilizing PLAS-20k dataset and deep learning model OnionNet.

- **Results**: PLAS-20k achieves PCC of 0.91 (Strong correlation).
- RMSE of 8.15 kcal/mol (Accurate predictions).

- **Significance**: PLAS-20k dataset is a powerful training resource.
- PLAS-20k demonstrates the potential of deep learning in binding affinity prediction.

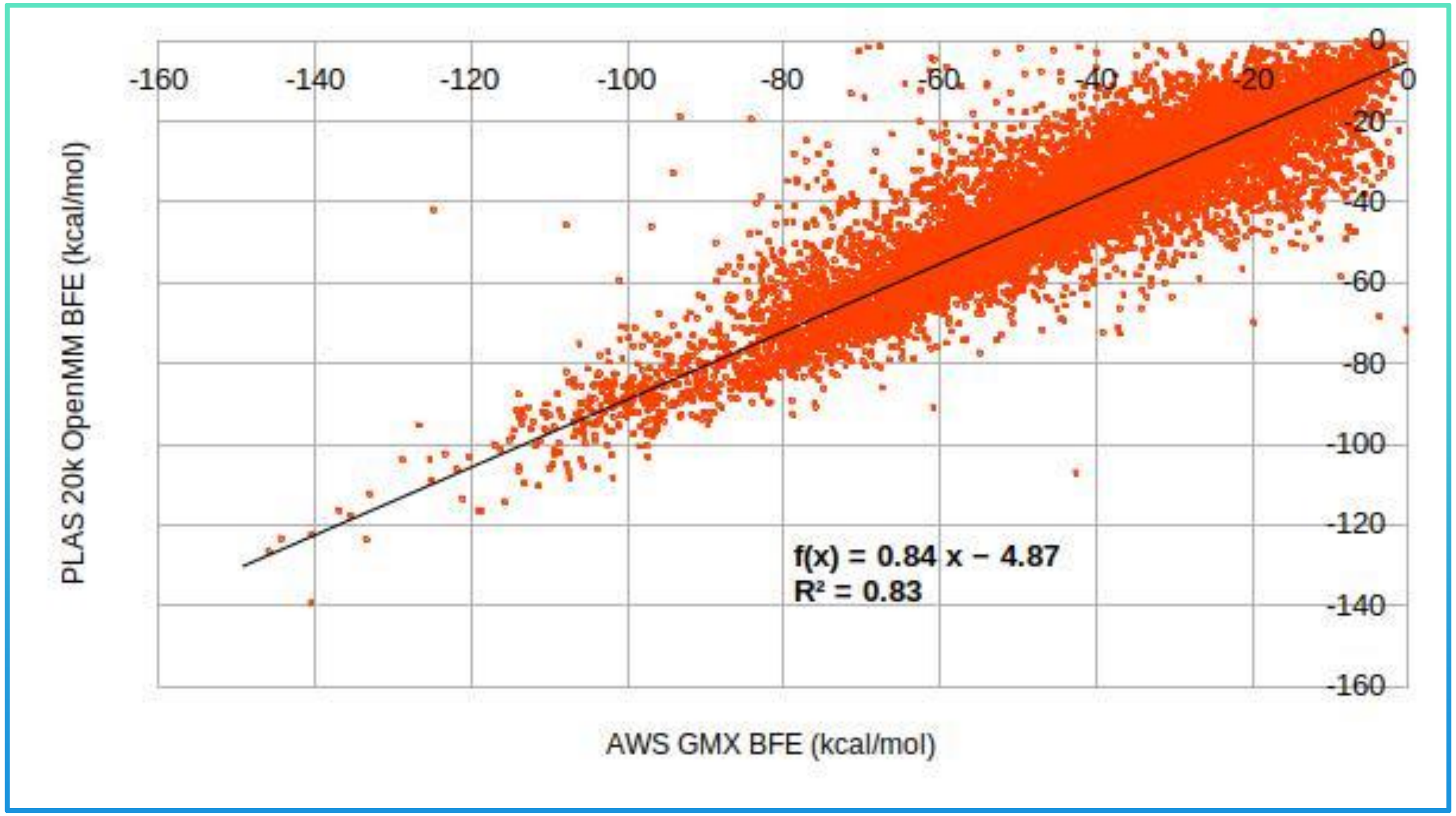




https://doi.org/10.26434/chemrxiv-2023-mg07d

Stage 1: Results (PLAS-20k versus AI^3)

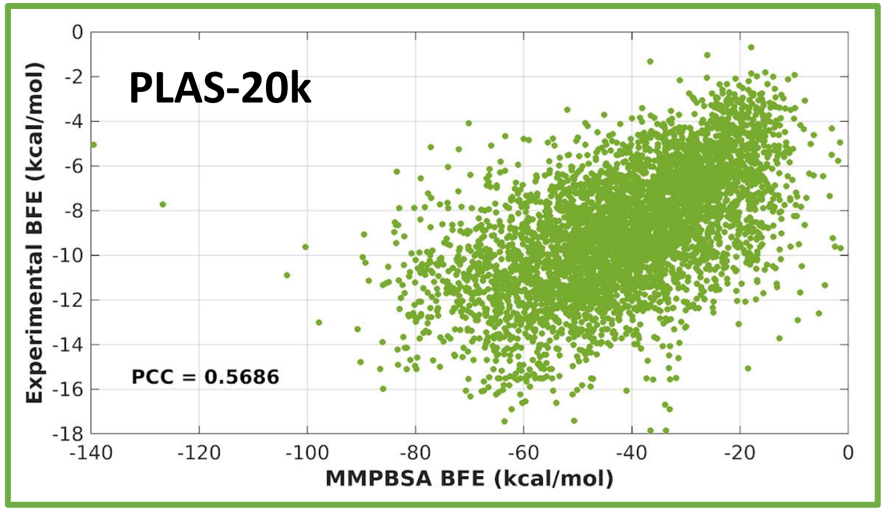
PLAS-20k

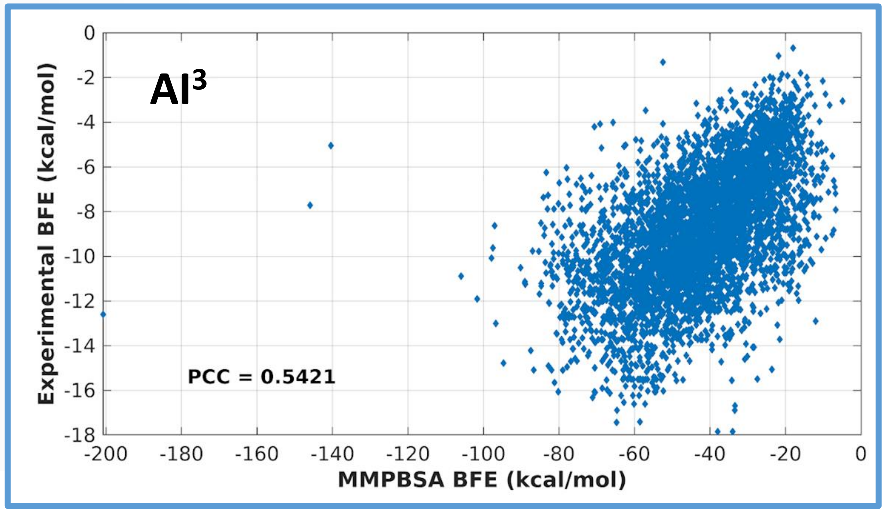


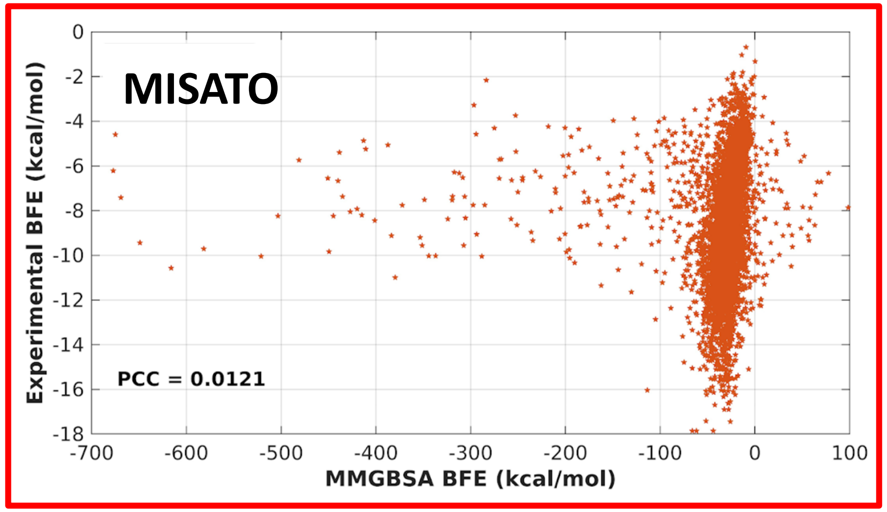
AI^3

Stage 1: Comparison with Other Datasets

Dataset	PCC
PLAS-20k	0.5686
AI^3	0.5421
MISATO	0.0121







AI^3: Stage 1

Consortium to accelerate Protein-Ligand Datasets

AI^3

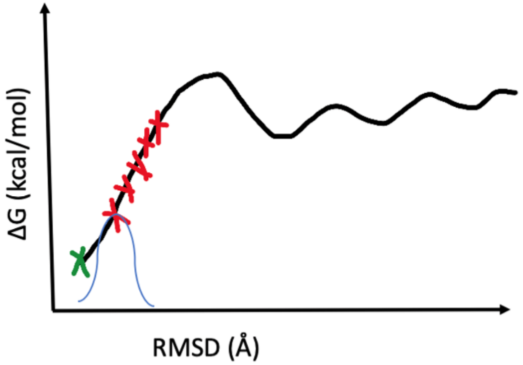
Releasing Stage 1 of the World’s largest Protein binding

4.6 TB

Releasing Code for Augmenting the PLC Dataset on Cloud Infrastructure

Github Repo (coming soon)

Future Plans for the AI^3 Dataset



Stage 2b: Higher Energy PLCs

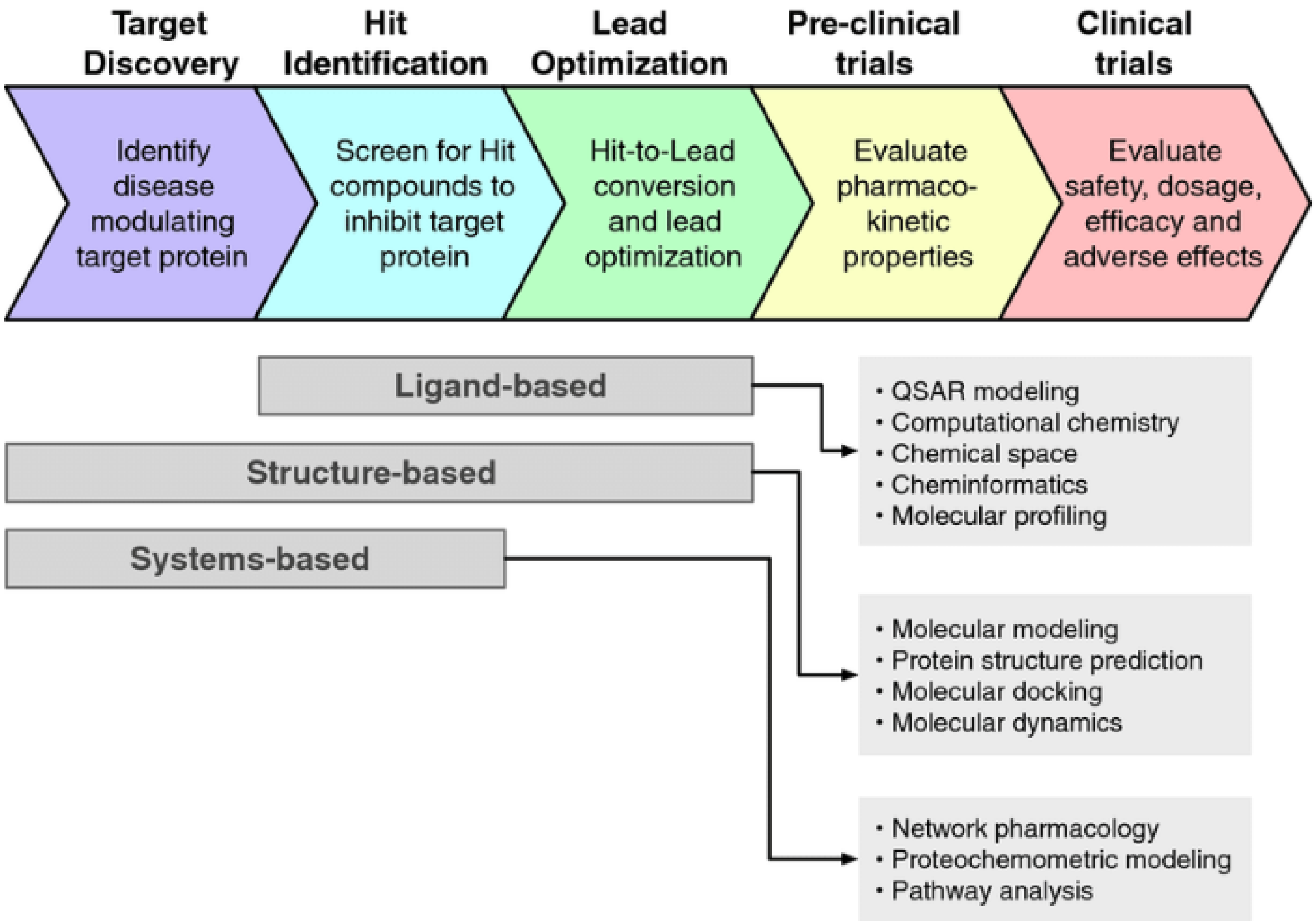
Generate ~220,000 PLC datasets.

Evaluate ~100 kinase inhibitors experimentally in the same lab conditions, and validate computational results.

Join us on the AI^3 journey to build the largest PLC Datasets & Accelerate Drug Discovery

Thank you

Backup



Target Discovery: Identify disease modulating target protein

Hit Identification: Screen for Hit compounds to inhibit target protein

Lead Optimization: Hit-to-Lead conversion and lead optimization

Pre-clinical trials: Evaluate pharmaco-kinetic properties

Clinical trials: Evaluate safety, dosage, efficacy and adverse effects

Ligand-based
- QSAR modeling
- Computational chemistry
- Chemical space
- Cheminformatics
- Molecular profiling

Structure-based
- Molecular modeling
- Protein structure prediction
- Molecular docking
- Molecular dynamics

Systems-based
- Network pharmacology
- Proteochemometric modeling
- Pathway analysis

Schematic summary of the drug discovery process overlayed with corresponding computational approaches

Computer aided drug design methodology (1990s) (Nalini, 2020)

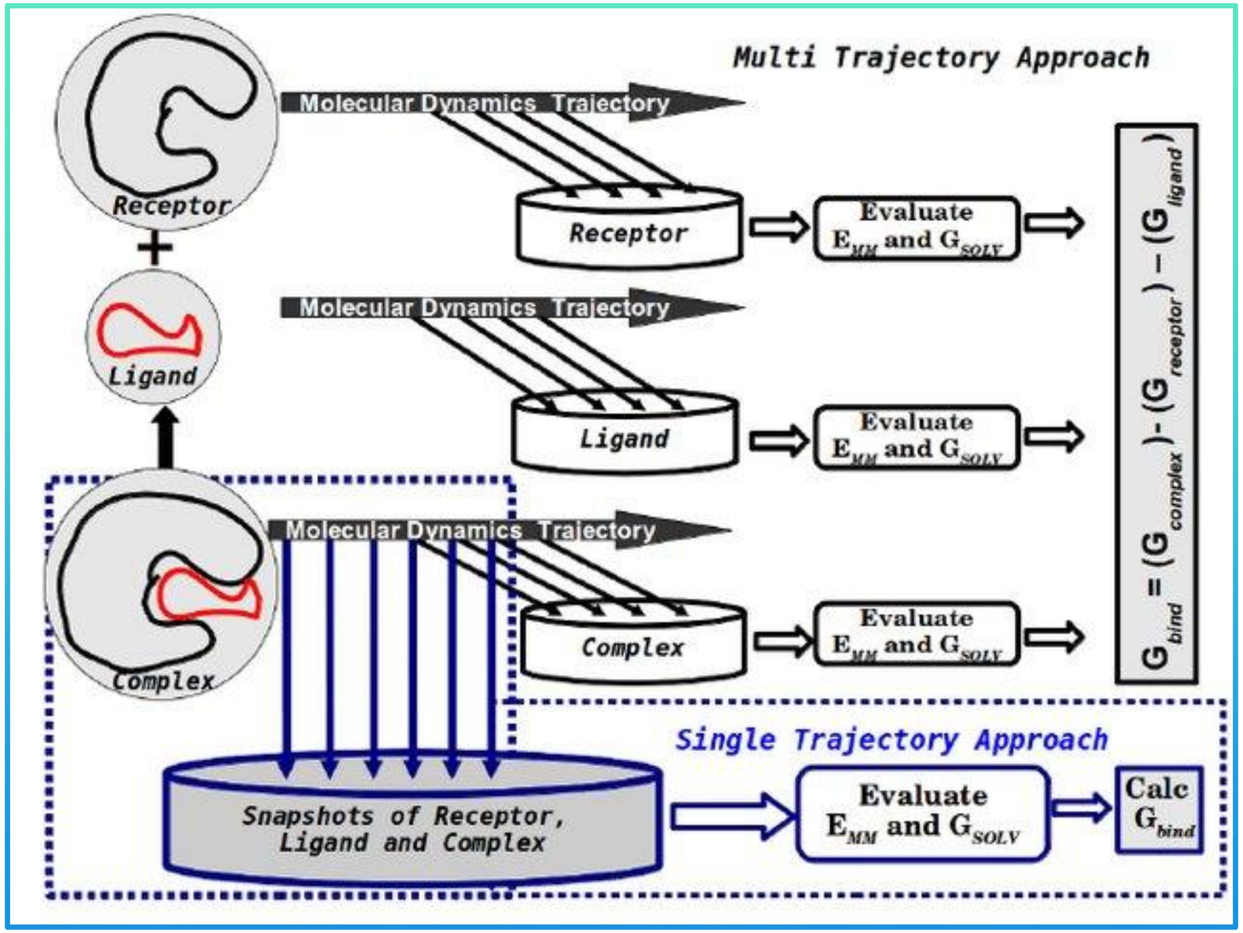
Molecular Dynamics (MD) Simulations

- Molecular dynamics simulations consider protein conformational rearrangements during binding.
- Techniques like MM-PBSA and MM-GBSA calculate binding free energy.
- Post-processing methods, including thermodynamic integration and free-energy perturbation (FEP), contribute to binding free energy determination.

$$\Delta G_{MM-PBSA} = \Delta E_{MM} + \Delta G_{Sol}$$

$$\Delta E_{MM} = \Delta E_{ele} + \Delta E_{vdw}$$

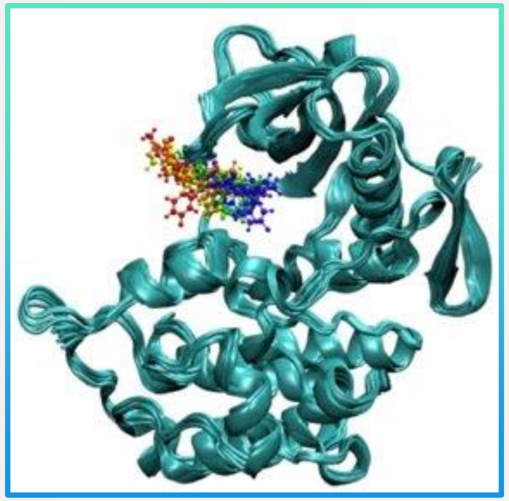
$$\Delta G_{Sol} = \Delta G_{pol} + \Delta G_{np}$$



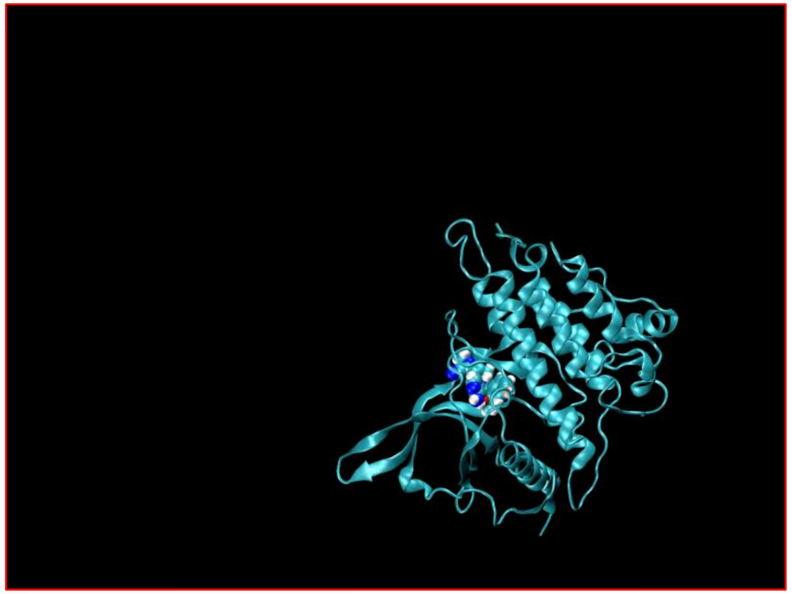
Multi Trajectory Approach
Receptor
+
Ligand
Complex
Molecular Dynamics Trajectory
Molecular Dynamics Trajectory
Molecular Dynamics Trajectory
Receptor
Ligand
Complex
Evaluate E_{MM} and G_{SOLV}
Evaluate E_{MM} and G_{SOLV}
Evaluate E_{MM} and G_{SOLV}
$G_{bind} = (G_{complex}) - (G_{receptor}) - (G_{ligand})$
Single Trajectory Approach
Snapshots of Receptor, Ligand and Complex
Evaluate E_{MM} and G_{SOLV}
Calc G_{bind}

In Silico Engineering of Proteins That Recognize Small Molecules, Mishra, 2012

Stage 2b: Higher Energy PLCs (under development)



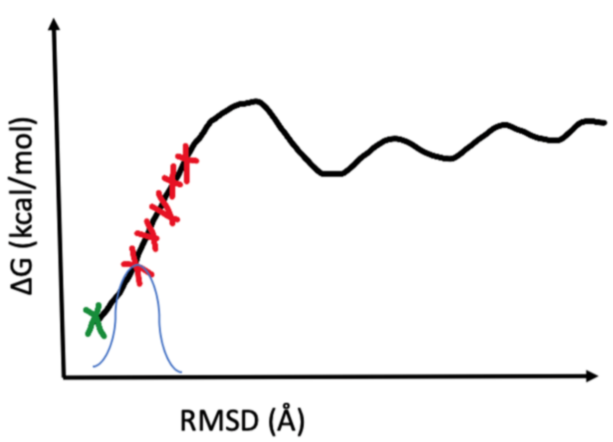
For each PLC, 10 partially bound (higher energy) structures are generated through steered MD simulation



Partially bound or Unbound Protein-Ligand (P-L) Binding Affinity Dataset

▶

Application: Negative examples in machine-learning model training for computational drug discovery.


ΔG (kcal/mol)
RMSD (Å)