In-Silico Computation of Drug-Target Binding Affinity Prediction using Quantum for Cancer

Final Year Project

Guide - Dr. Mala T

Members - Diya Arshiya S (2021115033)
Preetham Vijayandhran (2021115077)
Samiukktha Dharmalingam (2021115089)
Sathyadharini Srinivasan (2021115097)

PROBLEM STATEMENT

- Drug discovery and development is a complex, time-consuming, and costly process.
- Drug binds to its target, a factor known as binding affinity, and predicting the likelihood of a drug-target interaction.
- In silico methods, offer a faster, cost-effective alternative .

The project aims to develop and implement:

- Computational models that accurately predict the binding affinity.
- By leveraging machine learning, deep learning techniques, quantum neural networks and bioinformatics databases, the project seeks to create a robust framework that can predict these interactions with high accuracy.

SOLUTION:

DTA:

- Strongly a drug binds to its target.
- It determines the effectiveness of the drug.
- Strong binding affinity usually correlates with a higher likelihood of the drug modulating the target's activity.

DTI:

- Involves estimating the likelihood that a given drug will interact with a specific biological target.
- DTI prediction is about whether or not an interaction occurs at all.

Relationship between DTA and DTI:

- DTA gives a quantitative measure of binding strength
- DTI provides a binary or probabilistic indication of whether an interaction occurs.
- DTA predictions can be used in conjunction with DTI models to not only predict whether an interaction occurs but also how strong it might be.

LITERATURE SURVEY:

Title of the Article	Journals/Conference Details and Year of Publication	Highlights of Approaches Followed	Challenges to be Addressed
Prediction of Drug-Target Affinity Using Attention Neural Network	Tang, X.; Lei, X.; Zhang, Y. *Int. J. Mol. Sci.* 2024	• Utilization of multiple datasets, such as KIBA and Davis, to ensure the model's generalizability across different data distributions. • Data augmentation techniques applied to artificially expand the dataset, improving model robustness.	• High computational demands due to complex model architecture. • Dependence on the quality and availability of DTA datasets.
Drug Target Interaction Prediction using Graph Convolution based Neural Fingerprinting	A. Joshy, G. C. Kasyap, P. D. Reddy, I. T. Anjusha and K. A. Abdul Nazeer, *2022 IEEE 19th India Council International Conference (INDICON)*, Kochi, India	• NFPCNN integrates Neural Fingerprinting and 1D CNNs for accurate DTI prediction, focusing on drug repurposing for orphan diseases. • Demonstrates strong performance on the KIBA dataset with low MSE, outperforming models like SimBoost and DeepDTA.	• Balancing computational efficiency with model complexity in NFPCNN. • Ensuring data quality and diversity for reliable DTI predictions.

SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines	He, T., Heidemeyer, M., Ban, F. *et al* *J Cheminform* 9, 24 (2017)	• SimBoost improves continuous DTA prediction over binary classification, offering nuanced insights. • Outperforms KronRLS and Matrix Factorization on datasets like Davis, Metz, and KIBA.	• Handling missing data and distinguishing it from true negative interactions. • Managing the computational complexity of feature construction and GBM training
CutQC: using small Quantum computers for large Quantum circuit evaluations	Wei Tang, Teague Tomesh, Martin Suchara, Jeffrey Larson, and Margaret Martonosi. Association for Computing Machinery, New York, NY, USA	• CutQC enables evaluation of circuits up to 100 qubits. • Extends quantum device capabilities using hybrid computing.	• Tackling noise and low qubit counts in NISQ devices. • Managing hybrid quantum-classical integration complexity.
Quantum Machine Learning: A tutorial	José D. Martín-Guerrero, Lucas Lamata. Neurocomputing,Volume 470,2022,Pages 457-461,ISSN 0925-2312	• QML offers speedups by enhancing ML with quantum resources. • Provides a clear classification of QML approaches and applications.	• Sensitivity to parameters like Gaussian kernel length scale. • Bridging theory and practical implementation in QML.

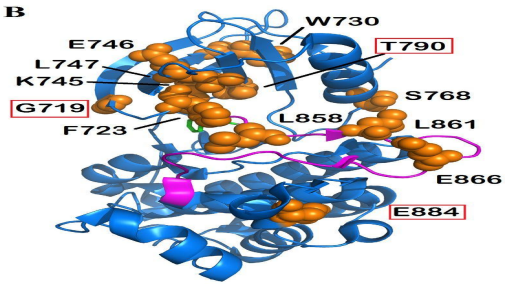
DEFINITIONS:

DRUG:

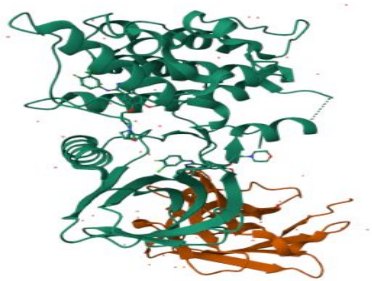
Drug is a molecule or compound that is designed to interact with a specific biological target, such as a protein, enzyme, receptor, or gene, to produce a desired therapeutic effect.

PROTEIN:

Protein is a biological molecule that serves as a potential target for drug action. Proteins are made up of amino acids and play crucial roles in the body's cellular processes, including signaling, metabolism, and structural support.



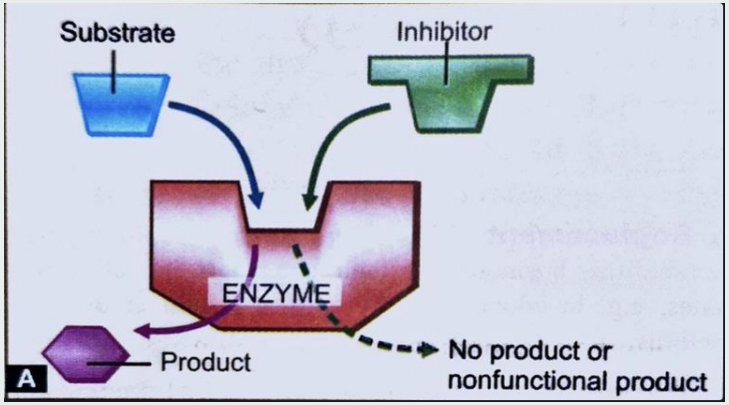
Epidermal Growth Factor Receptor (Protein)



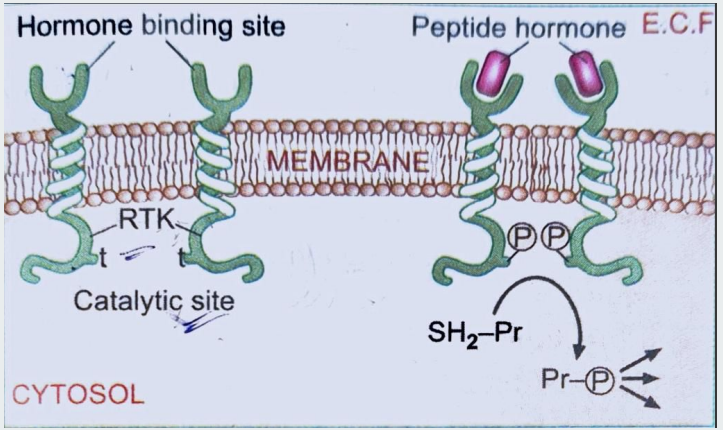
Gefitinib(Drug-Tyrosine Kinase Inhibitor)

DRUG TARGET BINDING:

1) Target recognition
2) Initial contact
    a) Electrostatic interaction
    b) Helps orient the drug properly towards binding site
3) Induced fit and conformational change
    a) Conformational change of target may take place to accommodate the drug better.
4) Formation of non-covalent interaction
    a) To stabilize the drug-target complex
5) Formation of drug-target complex
    a) Dynamic equilibrium between bonded and unbonded forms.



Drug - Target Binding



Tyrosine Kinase Receptor

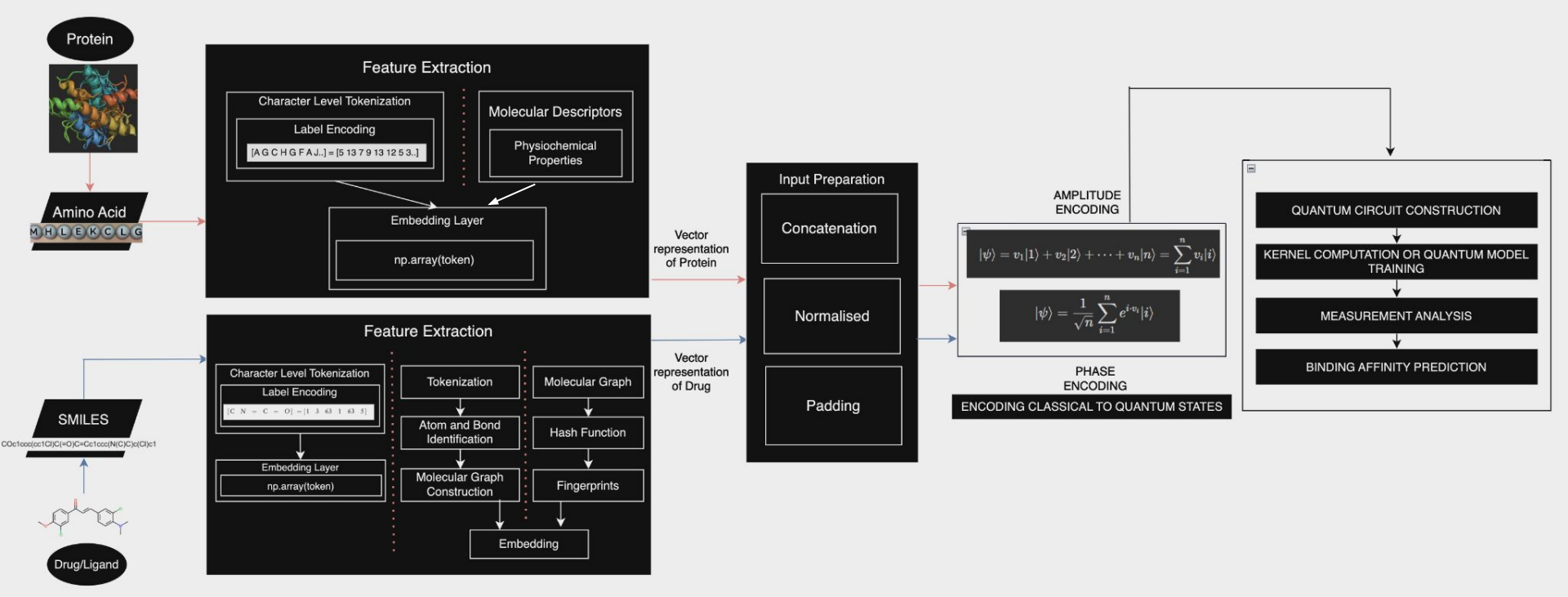
DATASET AND ATTRIBUTES:

- **KIBA:** Chemblid (drug_id), SMILES (molecular drug representation), Target_seq (protein-amino acid sequence), Target_id (Protein id), KIBA_score (ki,kd,IC50)
- **DAVIS:** Compound_id, SMILES, Protein_seq, K_d(dissociation constant for the drug-target pair), P_Kd (negative logarithm of the K_d value (pK_d = -log(K_d), convenient measure of binding affinity as it scales the values to a small range)
- **PLAS 5K:** PDBid, binding_affinity (kcal/mol), binding_affinity_sd (kcal/mol), electrostatic (kcal/mol), electrostatic_sd (kcal/mol), polar_solvation (kcal/mol)....

Main Attributes:

- SMILES
- Amino acid sequence
- pKd
- Ki
- IC50

ARCHITECTURE DIAGRAM:



TOKENIZATION AND FEATURE EXTRACTION:

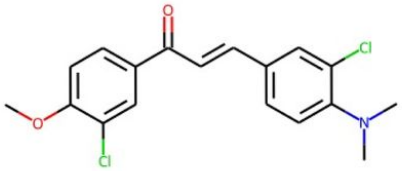
Steps to translate the raw data into a format that the model can process and learn from.

For DRUG/Ligand:

a. SMILES Tokenization

- **SMILES Representation**: A SMILES string is a text-based representation of a molecule, where atoms are represented by symbols (e.g., C for carbon) and bonds by specific characters (e.g., = for double bonds).
- **Tokenization**:
  - **Character-Level Tokenization**: Each character in the SMILES string is treated as a token (e.g., "CCO" → ["C", "C", "O"]).
  - **Substructure-Level Tokenization**: Sometimes, common chemical substructures (e.g., "C=O") are tokenized as single units.
- **Embedding**: After tokenization, each token is mapped to a vector embedding, either through pre-trained embeddings or through learning embeddings during model training.

PHASE 1: Tokenization of SMILES (Drug/Ligand):



SMILES

COc1ccc(cc1Cl)C(=O)C=Cc1ccc(N(C)C)c(Cl)c1

Tokenized SMILES

COc1ccc(cc1Cl)C(=O)C=Cc1ccc(N(C)C)c(Cl)c1

Token IDs

```
[8220, 66, 16, 535, 66, 7, 535, 16, 2601, 8, 34, 7, 28,
46, 8, 34, 28, 34, 66, 16, 535, 66, 7, 45, 7, 34, 8,
34, 8, 66, 7, 2601, 8, 66, 16]
```

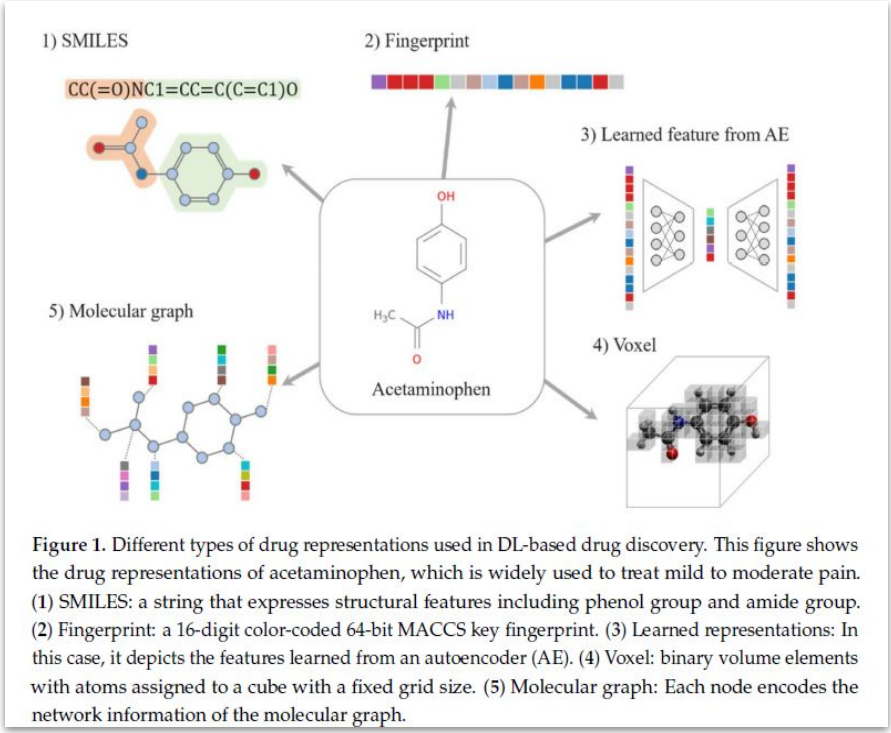


Figure 1. Different types of drug representations used in DL-based drug discovery. This figure shows the drug representations of acetaminophen, which is widely used to treat mild to moderate pain. (1) SMILES: a string that expresses structural features including phenol group and amide group. (2) Fingerprint: a 16-digit color-coded 64-bit MACCS key fingerprint. (3) Learned representations: In this case, it depicts the features learned from an autoencoder (AE). (4) Voxel: binary volume elements with atoms assigned to a cube with a fixed grid size. (5) Molecular graph: Each node encodes the network information of the molecular graph.

b. Molecular Graphs (for GNNs)

- **Graph Representation**: A molecule is represented as a graph, where atoms are nodes and bonds are edges.
- **Node Features**: Each node (atom) is assigned a feature vector that can include properties like atomic number, atom type, hybridization state, and charge.
- **Edge Features**: Bonds between atoms are encoded with features such as bond type (single, double, etc.) and whether the bond is aromatic.
- **Graph Neural Networks (GNNs)**: These models directly process the molecular graph to extract features, propagating information through the network to learn node (atom) and edge (bond) embeddings.

c. Molecular Fingerprints

- **Description**: Molecular fingerprints are fixed-length binary vectors that encode the presence or absence of specific substructures in a molecule.
- **Feature Extraction**: The SMILES string is converted into a molecular graph, and a hashing function is used to generate the fingerprint, which is then fed into the model.

For Protein/Target

a. Amino Acid Sequence Tokenization

- **Sequence Representation**: The primary structure of a protein is a sequence of amino acids, each represented by a single-letter code (e.g., "MKWVTFISLLFLFSSAYSR" for the first part of the human albumin sequence).
- **Tokenization**:
  - **Character-Level Tokenization**: Each amino acid is treated as a token (e.g., "MKWV" → ["M", "K", "W", "V"]).
  - **K-mer Tokenization**: Instead of single amino acids, groups of k consecutive amino acids (k-mers) are tokenized (e.g., "MKWV" → ["MK", "KW", "WV"] for k=2).
- **Embedding**: Tokens are mapped to vector embeddings.

b. Protein Descriptors

- **Physicochemical Properties**: Features can be extracted based on the physicochemical properties of the amino acids (e.g., hydrophobicity, charge).

.

PHASE 2: Quantum Enhanced Binding Affinity Prediction:

When predicting and classifying binding affinity using datasets such as **Davis** and **KIBA**, which are widely used in drug-target interaction studies, you typically rely on quantitative structure-activity relationship (QSAR) models, machine learning, and deep learning techniques.

Binding Affinity Prediction using Davis and KIBA Datasets

The binding affinity prediction task involves predicting a continuous value, such as Kd from the Davis dataset or the KIBA score, for drug-target pairs. Quantum-enhanced machine learning methods can be applied to improve prediction accuracy.

- **Feature Representation**:
  - ➔ Represent drugs and proteins using molecular descriptors (e.g., SMILES for drugs and protein sequences).
  - ➔ These features can be encoded into quantum states for quantum machine learning models, such as quantum kernels or quantum-enhanced neural networks. Amplitude encoding is used for this.

$$|\psi\rangle = \frac{1}{\|\mathbf{x}\|} \sum_{i=1}^{n} x_i |i\rangle,$$

Use of Quantum -

The core idea is to use quantum-enhanced machine learning models to predict the binding affinity between drugs and their target proteins, as measured in the Davis and KIBA datasets.

A. Quantum Kernel-Based Models

One approach is to use **quantum support vector regression (QSVR)** or **quantum kernel methods** to predict binding affinity.

- **Quantum Kernel**: Compute the kernel between two feature vectors (e.g., drug and target embeddings) encoded as quantum states. For example:

$$K(\mathbf{x}_i, \mathbf{x}_j) = |\langle \phi(\mathbf{x}_i) | \phi(\mathbf{x}_j) \rangle|^2$$

where ϕ(xi), is the quantum state representation of the drug-target features.

- **Quantum Support Vector Regression**: Using the quantum kernel, the binding affinity y (e.g., Kd from the Davis dataset) and we need to formulate a objective function.

B. Quantum Neural Networks (QNNs)

Quantum neural networks (QNNs) can be used for regression tasks, predicting continuous values such as the binding affinity.

- **Quantum Layers**: Incorporate quantum circuits as layers in a neural network. The network takes drug-target feature vectors as input and passes them through quantum layers that perform transformations based on quantum gates.
- **Output**: The final layer of the QNN outputs a continuous prediction, such as the Kd value from the Davis dataset or the KIBA score.

Major libraries:QISKIT LEARN by IBM , Paddle APIs.

DOMAIN KNOWLEDGE:

MODEL	APPROACH	ARCHITECTURE	ADVANTAGES	DISADVANTAGES
DeepDTA (2019)	Deep learning from raw drug (SMILES) and protein (sequence) data	CNN for SMILES and protein sequences, concatenated for affinity prediction	Eliminates need for manually crafted features	Ignores structural/3D information of drugs and proteins
MT-DTA (2019)	Multi-task learning for affinity and interaction classification	CNNs for SMILES and protein sequences; dual branches for affinity score and interaction classification	Multi-task learning improves generalization	Lacks structural/3D data integration
MLSDTA (2019)	Multi-level similarity based on sequence, structural, and interaction similarities	Uses multiple similarity matrices (e.g., sequence, molecular fingerprint) with machine learning (e.g., kernel regression)	Incorporates various similarity measures	Requires precomputed similarity matrices, limiting scalability
AttentionDTA (2020)	Attention mechanism to focus on important sequence regions	CNNs for SMILES and protein sequences, with attention layer	Interpretable due to attention mechanism	Relies only on raw sequence data; no structural/3D information
NFPCN (2021)	Neural fingerprint convolutional network for structural drug and target information	GNN for molecular graphs (drugs), CNN for protein sequences, combined for prediction	Incorporates structural information from molecular graphs	Still lacks detailed 3D target structure data

Are constrained by their reliance on classical approximations of molecular interactions!!

1. Simplified Molecular Representations:

- Classical models often rely on simplified representations of molecules, such as molecular fingerprints, sequences, or 2D/3D structures.
- These representations do not capture quantum effects, such as electron distribution, molecular orbitals, or quantum tunneling, which are essential to understanding how drugs and proteins interact at a fundamental level.

2. Limited Modeling of Electron Behavior:

- Classical models cannot directly simulate how electrons, which are central to chemical bonding and interactions, behave in molecules. Electrons operate based on quantum mechanics (wave-particle duality, superposition, etc.), but classical models only use approximate methods like molecular mechanics or force fields to predict interactions.

3. Inability to Capture Quantum Phenomena:

- Quantum phenomena, such as **superposition** (a molecule can exist in multiple states simultaneously) and **entanglement** (correlation between particles across space), are crucial for accurately modeling molecular interactions.
- Classical models cannot capture these phenomena due to their reliance on deterministic equations from classical physics (e.g., Newtonian mechanics, Boltzmann statistics).

4. Approximation Methods (e.g., Force Fields):

- Classical models often use **force fields** (approximation techniques) to estimate the energy of molecular systems. These force fields are parameterized based on empirical data, but they lack the precision required to model the behavior of electrons and atoms at the quantum scale.
- This leads to approximate predictions of binding affinities, reaction rates, and other molecular properties.

Why Quantum?

Quantum machine learning offers the potential to model molecular interactions at the quantum level, leading to more accurate and efficient predictions of drug-target binding affinity.

Quantum Model	Approach	Reason for Transition
Quantum Kernel Models	Compute complex similarity measures between drugs and targets in quantum space	**Quantum effects in molecular interactions** (electron orbitals, superposition) require quantum methods for accurate modeling.
Quantum Neural Networks (QNNs)	Process quantum-encoded data using quantum states	**Exponential growth in problem size** in molecular systems can be handled efficiently using quantum superposition and entanglement.
Quantum Self-Supervised Models	Capture richer relationships for self-supervised learning in DTA	**Quantum embeddings** allow for a more nuanced representation of drug-target interactions beyond classical descriptors.
Quantum Simulators	Quantum computers simulate molecular systems using Schrödinger's equation	**Better simulation of complex systems** like drug-target interactions at the quantum mechanical level.