Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

AI-Based Platform for Drug-Pathogen Molecular Interaction Analysis: A Full-Stack Adaptive Framework with Random Forest Affinity Prediction and Physicochemical Profiling

Dr. Varsha S Jadhav
Information Science and Engineering
SDM College of Engineering and Technology
Dharwad, India

Shankar Patil
Information Science and Engineering
SDM College of Engineering and Technology
Dharwad, India

Keshav Terdal
Information Science and Engineering
SDM College of Engineering and Technology
Dharwad, India

Manikanth R Hebballi
Information Science and Engineering
SDM College of Engineering and Technology
Dharwad, India

Vaibhav Kalyanshetti
Information Science and Engineering
SDM College of Engineering and Technology
Dharwad,India

***Abstract* - Growing concerns over antimicrobial resistance (AMR) have created an urgent need to rethink how candidate drugs are identified in early development stages. Existing physics-driven methods such as molecular docking and highthroughput screening (HTS) suffer from high computational overhead and poor scalability, rendering them unsuitable for keeping pace with rapidly evolving pathogenic mutations. This work introduces a modular, full-stack computational system that tightly couples machine learning-based affinity estimation, onthe-fly molecular characterisation, and live interaction scoring within one unified pipeline. At the core of the system lies a Random Forest (RF) ensemble optimised for drug–target binding affinity regression. The primary novelty is the *Adaptive Affinity Module*, which constructs 1044-dimensional input vectors by concatenating 512-bit ECFP4 Morgan Fingerprints encoding small-molecule ligands with 532-element integer-mapped protein sequence representations. A penalty-weighted scoring engine built on RDKit evaluates molecular stability to deterministically identify high-quality binding candidates. For shortlisted complexes, five-dimensional ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) profiles are generated in real time, yielding actionable drug-likeness estimates. The *Mutation Laboratory* has been substantially upgraded into a *Physicochemical and Evolutionary Intelligence Engine*, incorporating BLOSUM62-guided evolutionary impact scoring, lightweight 3D burial analysis, and a context-aware physicochemical refinement formula that adjusts the predicted $\Delta\Delta G$ beyond raw ML output. Testing on 500 curated drug–protein records drawn from BindingDB and ChEMBL produced an RMSE of 0.62 and R^2 of 0.89, with predictions generated in approximately 1.2 seconds per query.**

***Index Terms*—Antimicrobial Resistance (AMR), Random Forest, Morgan Fingerprints (ECFP4), ADMET Profiling, Drug– Target Interaction (DTI), Binding Affinity, SMILES, RDKit, BLOSUM62, Mutation Resistance Simulation, Evolutionary Impact, Residue Burial Analysis, Full-Stack AI, Flask, React**

I. INTRODUCTION

A. The Global Burden of Antimicrobial Resistance

Among the foremost threats to global public health in the modern era, antimicrobial resistance (AMR) stands out for its potential scale and urgency. Projections from the World Health Organization suggest that without coordinated intervention, fatalities attributable to drug-resistant pathogens may exceed 10 million per year by 2050, overtaking oncological diseases as a primary cause of death [1]. Murray *et al.* [4] reported through a comprehensive 2022 study that AMR accounted for roughly 1.27 million direct deaths worldwide in 2019, while Laxminarayan *et al.* [3] underscored the necessity of globally coordinated countermeasures. The proliferation of multidrug-resistant (MDR) organisms—among them Methicillin-Resistant *Staphylococcus aureus* (MRSA), carbapenem-resistant *Enterobacteriaceae* (CRE), and extensively drug-resistant *Mycobacterium tuberculosis* (XDR-TB)—continues to erode the effectiveness of available antibiotics [2]. Well-documented biochemical resistance strategies, such as enzymatic drug inactivation, efflux-mediated expulsion, and alteration of binding targets, collectively challenge the continued utility of current treatment regimens [42], [43].

Conventional drug development workflows encompass target identification, compound screening, lead refinement, and multi-phase clinical evaluation, with typical timelines of 10– 15 years and expenditures reaching USD 2.6 billion per successful approval [5]. Research by Paul *et al.* [6] pinpointed high late-stage attrition as the dominant cost driver, motivating investment

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

in more reliable upstream predictive technologies. Structure-based computational tools—including AutoDock Vina [7], Glide [9], and GROMACS [46]—have partially accelerated early phases yet impose significant resource demands, often requiring dedicated computing clusters and extended runtimes per compound [51]. Although structure-guided virtual screening strategies [52] informed by the Protein Data Bank [40] have expanded screening capacity, their dependence on experimentally resolved 3D structures remains a fundamental bottleneck.

B. Limitations of Existing Approaches

Three recurring weaknesses characterise contemporary DTI prediction and screening frameworks:

- Throughput and Scalability Constraints: Structure based docking algorithms exhaustively traverse 3D conformational landscapes of protein active sites. Processing a single compound with AutoDock Vina may consume 20–60 minutes, making proteome-scale screening infeasible in standard laboratory settings.
- Fragmented Output Reporting: The majority of DTI tools output only a binding affinity estimate (ΔG or K_d), leaving out the broader physicochemical context—such as Lipinski compliance, metabolic vulnerability, and safety flags—that clinical translation demands.
- Limited Transparency and Mutation Modelling: Many deep learning DTI systems (e.g., DeepDTA, GraphDTA) function as black boxes, providing no insight into which molecular features govern affinity predictions. Furthermore, these models lack mechanisms to assess how point mutations or single-nucleotide polymorphisms (SNPs) within pathogen targets translate to drug resistance, and none incorporate evolutionary or structural context into their mutation analysis.

C. Proposed Solution and Contributions

The proposed platform directly tackles each of these shortcomings through an integrated AI-driven approach. The principal contributions of this work include:

1) A novel 1044-dimensional molecular representation formed by combining ECFP4 ligand fingerprints with ordinally encoded protein sequences.
2) A tuned Random Forest ensemble delivering $R^2 = 0.89$ and RMSE = 0.62 with near-instantaneous inference (∼1.2 s per query).
3) An embedded five-axis ADMET scoring module providing real-time pharmacokinetic assessment within the prediction loop.
4) An upgraded *Physicochemical and Evolutionary Intelligence Engine* within the Mutation Laboratory, incorporating BLOSUM62 evolutionary scoring, 3D residue burial analysis, and a physicochemical refinement formula for biologically grounded $\Delta\Delta G$ estimation.
5) A holographic 3D visualisation interface dynamically linked to the mutation calculator for spatial context analysis.
6) A production-ready full-stack system secured via JWT tokens and PBKDF2-HMAC-SHA256 password hashing.

II. LITERATURE SURVEY

Computational methods for drug–target interaction (DTI) prediction have undergone a substantial evolution, transitioning from knowledge-driven pharmacophore rules to sophisticated data-centric paradigms leveraging machine learning and deep neural networks. Table I presents a structured comparison of notable methods spanning 2018 to 2025.

A. Sequence-Based Deep Learning

The DeepDTA framework [10] introduced the use of character-level SMILES and amino acid sequence embeddings fed into dual-branch 1D convolutional networks, achieving a concordance index (CI) of 0.878 on the Davis kinase benchmark [31]. However, its reliance on n-gram character features limits chemical interpretability. DeepConv-DTI [14] augmented this architecture by employing multi-scale convolutions across protein sub-sequences. TransformerCPI [11] adopted self-attention mechanisms [19] to improve classification performance, though its GPU memory demands escalate severely for sequences exceeding 1000 residues.

B. Graph Neural Network Approaches

In GraphDTA [15], drug molecules are modelled as attributed graphs where atomic nodes exchange neighbourhood features via message-passing [18], enabling topologicallyaware chemical encoding. Complementary work by Tsubaki *et al.* [47] demonstrated end-to-end graph-sequence learning for compound–protein interaction, while NeoDTI [48] extended this to heterogeneous biological networks. Despite their representational richness compared to fingerprint vectors, graphbased pipelines incur non-trivial construction and batching overheads that conflict with real-time inference requirements.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

TABLE I
COMPARATIVE LITERATURE SURVEY OF DRUG–TARGET INTERACTION PREDICTION METHODS (2018–2025)

Reference	Year	Method	Dataset	Key Metric	Limitation
DeepDTA [10]	2018	CNN (SMILES + Seq)	Davis, KIBA	= 0.878	No ADMET; slow training
GraphDTA [15]	2021	GNN + CNN	Davis, KIBA	MSE = 0.229	Graph construction overhead
TransformerCPI [11]	2020	Transformer	Human, *C. elegans*	AUC = 0.97	GPU-intensive; no regression
SCTDTI [12]	2022	Siamese CNN	BindingDB	RMSE = 0.71	No mutation analysis
AttentionDTA [13]	2020	Attention + BiGRU	Davis	MSE = 0.230	Binary interaction only
MolBERT [20]	2020	BERT (SMILES)	ChEMBL	AUC = 0.91	No protein encoding
DGraphDTI [16]	2020	Dual Graph	DrugBank	AUC = 0.963	No affinity regression
RF-Score [23]	2010	Random Forest	PDBbind	$R^2 = 0.77$	3D structure required
Proposed	2025	RF + Evo. Intel.	BindingDB+ChEMBL	R^2=0.89, RMSE=0.62	2D base; no full folding

C. Classical Machine Learning with Fingerprints

The RF-Score study [23] established that Random Forest models [24] trained on crystallographic interaction fingerprints [30] can attain $R^2 = 0.77$ on PDBbind, while Svetnik *et al.* [25] confirmed the aptness of ensemble trees for highdimensional QSAR feature spaces. The present work builds upon this tradition using Morgan/ECFP4 fingerprints [26], [27] as ligand descriptors, extending the 2D-only paradigm to achieve $R^2 = 0.89$ by fusing sequence-level protein features, thereby bypassing the 3D structure requirement entirely.

D. ADMET and Drug-Likeness Integration

Standalone tools such as SwissADME [36] and pkCSM [37] offer pharmacokinetic and toxicity profiling but operate in isolation from binding affinity estimators. Classical filter criteria including Lipinski's rule-of-five [35] and Veber's bioavailability guidelines [38] continue to serve as industry-standard oral drug-likeness benchmarks. While pre-trained molecular language models like ChemBERTa [21] have begun integrating ADMET-oriented learning objectives, no prior system has demonstrated their simultaneous delivery with affinity scoring and evolutionary mutation intelligence in a live, full-stack deployment.

E. Evolutionary Substitution Matrices in Drug Resistance

The BLOSUM62 matrix [22], derived from conserved protein blocks across diverse species, encodes the log-odds probability of observing one amino acid substituted by another in naturally occurring homologous sequences. Unlike PAM matrices which extrapolate from evolutionary models, BLOSUM62 is empirically derived and has been validated as the gold standard for local alignment scoring. Its integration into mutation resistance analysis is novel in the DTI context: negative BLOSUM62 scores indicate evolutionarily improbable substitutions that are more likely to destabilise protein fold and alter binding pocket geometry, making them high-impact candidates for drug resistance markers.

III. SYSTEM ARCHITECTURE

The proposed system adopts a three-tier design comprising a React/Vite user interface layer, a Flask REST service layer, and a combined Python ML + RDKit computation layer supported by a SQLAlchemy/SQLite storage backend.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

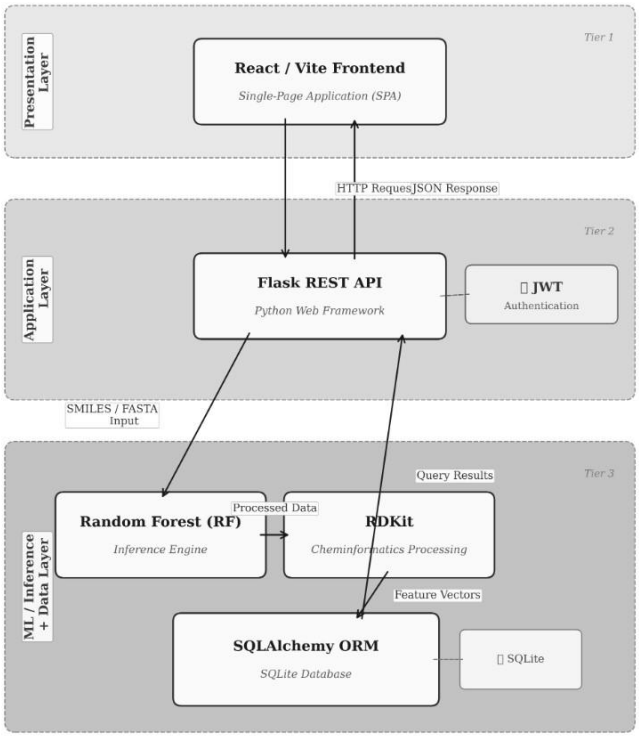

Fig. 1. High-level three-tier system architecture. The React/Vite frontend communicates with the Flask REST backend via JWT-authenticated HTTP requests. The backend dispatches SMILES and FASTA payloads to the RF inference engine and RDKit cheminformatics layer, persisting results to a SQLAlchemy ORM.

A. Frontend Architecture (React + Vite)

The client interface is built with React 18 and bundled via Vite, which enables Hot Module Replacement (HMR) for rapid iterative development. The UI is organised into the following functional components:

- Input Panel: Receives SMILES (ligand) and FASTA (protein) entries with real-time format checking through built-in regex validators.
- Results Dashboard: Dynamically renders a pentagonal ADMET radar chart via the Tremor library, alongside binding affinity indicators and a molecular stability gauge that refresh upon each completed inference.
- Mutation Lab Interface: Provides an upgraded interface accepting residue position, substitution amino acid, and optional PDB structure upload; renders the $\Delta\Delta G$ refinement breakdown as an interactive panel with evolutionary and structural sub-scores.
- Holographic 3D Viewer: A 3Dmol.js-powered ribbon diagram with real-time residue selection, dynamically linked to the mutation calculator for spatial context analysis of the target site.
- History Panel: Aggregates prior prediction records from both browser-local storage and the server persistence layer to support longitudinal analysis.

B. Backend Architecture (Flask REST API)

The server layer exposes five RESTful endpoints to client applications:

- POST /predict — Receives SMILES and FASTA inputs; returns computed ΔG, ADMET axes, and stability score.
- POST /mutate — Accepts residue index, substitution character, and optional PDB data; returns the refined $\Delta\Delta G$ with BLOSUM62 and burial sub-scores.
- POST /structure — Accepts PDB payload and residue index; returns burial score and neighbourhood count.
- GET /history — Delivers paginated logs of past prediction events.
- POST /auth/login — Authenticates users and issues a signed JWT access token.

C. Security and Cryptographic Overlay

The platform employs a two-layer security model: stateless session management via JSON Web Tokens (JWT) signed with 256-bit HMAC-SHA256, and credential protection through PBKDF2-HMAC-SHA256 key derivation at 480,000 iterations—in line with NIST SP 800-132 guidelines. All network communication is encrypted using HTTPS over TLS 1.3.

IV. WORKFLOW AND MECHANISM

A. Step 1 — Input Ingestion and Validation

The pipeline accepts two distinct molecular descriptors: a SMILES string encoding the candidate small-molecule

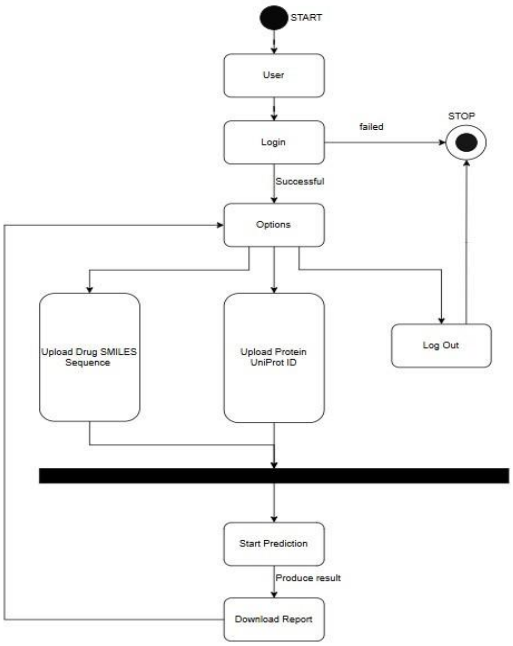

Fig. 2. End-to-end prediction workflow. SMILES and FASTA inputs are independently featurised, fused into a 1044-D vector, passed through the RF ensemble for ΔG regression, followed by RDKit-based ADMET scoring and optional mutation simulation.

compound, and an amino acid sequence in FASTA format representing the pathogenic protein target. Structural validity of the SMILES input is confirmed using RDKit's MolFromSmiles() parser, while the protein sequence undergoes alphabet enforcement via a custom FASTA validator restricted to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

the canonical 20-residue set. Inputs failing either check are immediately rejected with informative error messages before entering the ML inference stage.

B. Step 2 — Molecular Featurisation

A validated SMILES entry is transformed into an ECFP4 circular fingerprint (diameter 4, equivalent to Morgan radius r = 2) using RDKit's AllChem.GetMorganFingerprintAsBitVect() with n_{bits} = 512. The resulting binary vector $\mathbf{f}_L \in \{0,1\}^{512}$ encodes the local atomic neighbourhood up to two bond hops from every atom.

Each amino acid in the protein sequence is mapped to an integer via an ordinal scheme (A→1 through Y→20, with X→0 for non-standard residues). The encoded sequence is either padded with zeros or clipped to a uniform length of 532, yielding $\mathbf{f}_P \in Z^{532}$.

Both vectors are joined by concatenation:

$$\mathbf{x} = [\mathbf{f}_L \parallel \mathbf{f}_P] \in R^{1044} \tag{1}$$

C. Step 3 — Random Forest Inference

The concatenated 1044-dimensional descriptor $\mathbf{x}$ is forwarded to the pre-trained RF ensemble. Individual trees $h_t(\mathbf{x})$ independently regress a pK_d value, and the final ensemble estimate is obtained by averaging across all T = 200 decision trees:

$$\hat{pK}_d = \frac{1}{T}\sum_{t=1}^{T} h_t(\mathbf{x}) \tag{2}$$

D. Step 4 — Binding Free Energy Calculation

The ensemble-predicted pK_d is converted to a Gibbs binding free energy ΔG (kcal/mol) through a standard thermodynamic transformation evaluated at physiological body temperature (T = 310.15K):

$$\Delta G = -R \cdot T \cdot \ln\left(10^{-\hat{pK}_d}\right) \tag{3}$$

where R = 1.987 × 10^{-3} kcal $mol^{-1}K^{-1}$. Increasingly negative values of ΔG reflect progressively stronger predicted binding interactions.

E. Step 5 — Molecular Stability Scoring

Five RDKit-derived physicochemical descriptors are evaluated: molecular weight (MW), lipophilicity (LogP), hydrogen bond donor count (HBD), hydrogen bond acceptor count (HBA), and rotatable bond count (RBC). These feed a penaltybased stability scoring function to produce $S \in [0,100]$:

$$S = 100 - (P_{\text{MW}} + P_{\text{LogP}} + P_{\text{Bond}}) \tag{4}$$

where penalty terms are defined as:

$$P_{\text{MW}} = \max\left(0,\ \frac{\text{MW} - 500}{10}\right) \tag{5}$$

$$P_{\text{LogP}} = \max(0,\ (\text{LogP} - 5) \times 3) \tag{6}$$

$$P_{\text{Bond}} = \max(0,\ (\text{RBC} - 10) \times 2) \tag{7}$$

Compounds scoring S > 80 are labelled *Stable*; those in [60,80] are *Moderate*; those below 60 are *Unstable*.

F. Step 6 — ADMET Profiling

All shortlisted complexes receive scores across five pharmacokinetic dimensions: Absorption (derived from LogP and MW estimates), Distribution (computed from H-bond profile and topological polar surface area, TPSA), Metabolism (CYP450 susceptibility approximated via RBC), Excretion (renal clearance proxy based on MW), and Toxicity (modelled as an inverse drug-likeness index).

G. Step 7 — Physicochemical and Evolutionary Intelligence Engine (Upgraded Mutation Laboratory)

The Mutation Laboratory has been fundamentally redesigned from a basic sequence-editing tool into a multilayered *Physicochemical and Evolutionary Intelligence Engine*. The upgraded pipeline executes four tightly coupled analytical stages described in the following subsections and summarised in Algorithm 1.

V. PHYSICOCHEMICAL AND EVOLUTIONARY INTELLIGENCE ENGINE

A. Stage 1 — BLOSUM62 Evolutionary Impact Scoring

The first analytical stage queries the BLOSUM62 substitution matrix [22] to evaluate the biological plausibility of a proposed amino acid exchange. BLOSUM62 encodes empirically derived log-odds scores for every pair of amino acid substitutions observed in conserved protein blocks across diverse species, making it the benchmark for quantifying evolutionary conservation.

For a substitution from wild-type residue a to mutant residue b, the system retrieves the log-odds score:

$$\sigma = \text{BLOSUM62}(a,b) \tag{8}$$

A negative score (σ < 0) indicates an evolutionarily improbable exchange—one rarely observed in nature because it disrupts conserved physicochemical properties of the residue. Substitutions with σ < 0 are flagged as High Evolutionary Impact, and an explicit stability penalty is incurred:

$$P_{\text{evo}} = |\sigma| \times 5.0 \qquad \text{if } \sigma < 0 \tag{9}$$

This penalty reflects the well-established biophysical principle that mutations disrupting evolutionarily conserved positions are more likely to perturb the protein fold and remodel the binding pocket geometry. Conversely, substitutions with $\sigma \geq 0$ are labelled Low Evolutionary Impact, and no stability penalty is applied. The normalised evolutionary impact score used downstream is defined as:

$$I_{\text{evo}} = \frac{|\sigma|}{\sigma_{\max}} \tag{10}$$

where $\sigma_{\max}$ = 11 is the maximum positive BLOSUM62 entry (self-substitution of Tryptophan).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

B. Stage 2 — Lightweight 3D Residue Burial Analysis

The second stage evaluates the topographical context of the mutation site within the protein's three-dimensional structure. When a PDB coordinate file is available—either uploaded by the user or retrieved from the Protein Data Bank [40]— the engine parses the Cα atom coordinates and computes a *burial score* for the target residue i:

$$B_i = \left|\{j \neq i : \|\mathbf{r}_i - \mathbf{r}_j\| \leq 8.0\,\text{Å}\}\right| \tag{11}$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the Cα position vectors of residues i and j respectively, and B_i counts the number of neighbouring residues within the 8.0A radius shell.˚

Residues are classified based on their burial score:

$$\text{Context}(i) = \begin{cases} \textit{Buried (Structural Core)} & \text{if } B_i > 8 \\ \textit{Surface (Exposed)} & \text{if } B_i \leq 8 \end{cases} \tag{12}$$

A higher structural multiplier is assigned to buried residues because mutations in the protein core are substantially more likely to propagate conformational distortion toward the binding pocket than equivalent substitutions on surface-exposed loops:

$$M_{\text{struct}} = \begin{cases} 1.5 & \text{if Context}(i) = \textit{Buried} \\ 1.0 & \text{if Context}(i) = \textit{Surface} \end{cases} \tag{13}$$

If no PDB data is supplied, the system defaults to $M_{\text{struct}} = 1.0$ and proceeds with sequence-only analysis.

C. Stage 3 — Physicochemical Refinement of ΔΔG

The raw ML-predicted $\Delta\Delta G$ is refined by combining the evolutionary and structural signals into a single *Refinement Multiplier*:

$$R_{\text{total}} = M_{\text{struct}} \times (1.0 + I_{\text{evo}}) \tag{14}$$

The final physicochemically refined binding affinity shift is:

$$\Delta\Delta G_{\text{refined}} = \Delta\Delta G_{\text{ML}} \times R_{\text{total}} + P_{\text{evo}} \tag{15}$$

This formulation ensures that the platform responds to chemical reality even in cases where the ML model returns a near-zero or attenuated delta for a single amino acid substitution—a known limitation of sequence-only regressors when applied to subtle point mutations. The additive penalty P_{evo} from Equation (9) further amplifies the resistance signal for evolutionarily rare substitutions.

D. Stage 4 — Resistance Classification

The refined $\Delta\Delta G_{\text{refined}}$ is used as the definitive resistance metric:

- $\Delta\Delta G_{\text{refined}} > +1.0$kcal/mol: flagged as a Resistance Marker.
- $\Delta\Delta G_{\text{refined}} < -0.5$kcal/mol: flagged as a Sensitising Mutation.
- Otherwise: classified as Neutral.

E. Holographic 3D Visualisation Interface

The upgraded Mutation Laboratory is paired with a holographic 3D visualisation module rendered via the 3Dmol.js library. The viewer displays the full protein ribbon diagram and supports:

- Real-Time Residue Selection: Users click directly on a target residue (e.g., Position 32) in the 3D ribbon to populate the mutation calculator automatically, eliminating manual index entry.
- Active Structural Link: The viewer is dynamically coupled to the mutation engine—upon submitting a substitution, the selected residue is highlighted in red (mutant) against blue (wild-type), and the burial neighbourhood sphere is rendered as a translucent 8.0A shell.˚
- Binding Site Proximity Overlay: The spatial distance between the mutated residue and the ligand centroid is computed and displayed, enabling the researcher to visually assess whether the mutation site lies within the primary binding pocket or a distal regulatory region.

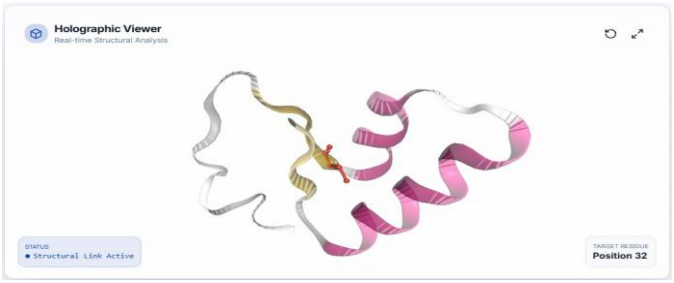

Fig. 3. Holographic 3D visualisation interface of the upgraded Mutation Laboratory. The ribbon diagram displays the full protein structure with Position 32 selected (highlighted sphere). The 8.0A burial neighbourhood shell is˚ rendered in translucent blue. The active structural link dynamically populates the BLOSUM62 and burial scores in the adjacent mutation calculator panel upon residue selection.

Algorithm 1 Physicochemical & Evolutionary Intelligence Engine

Require: Wild-type residue a, mutant residue b, position i, raw $\Delta\Delta G_{\text{ML}}$, optional PDB data
Ensure: Refined $\Delta\Delta G_{\text{refined}}$, resistance label
$\sigma \leftarrow$ BLOSUM62(a,b)
$I_{\text{evo}} \leftarrow |\sigma|/\sigma_{\max}$
if $\sigma < 0$ then
    $P_{\text{evo}} \leftarrow |\sigma| \times 5.0$
    Flag mutation as High Evolutionary Impact
else
    $P_{\text{evo}} \leftarrow 0$
end if
if PDB data available then
    $B_i \leftarrow$ count neighbours within 8.0A˚
    $M_{\text{struct}} \leftarrow 1.5$ if $B_i > 8$ else 1.0
else
    $M_{\text{struct}} \leftarrow 1.0$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

14: end if
15: $R_{\text{total}} \leftarrow M_{\text{struct}} \times (1.0 + I_{\text{evo}})$
16: $\Delta\Delta G_{\text{refined}} \leftarrow \Delta\Delta G_{\text{ML}} \times R_{\text{total}} + P_{\text{evo}}$
17: if $\Delta\Delta G_{\text{refined}} > +1.0$ then
18: return $\Delta\Delta G_{\text{refined}}$, Resistance Marker
19: else if $\Delta\Delta G_{\text{refined}} < -0.5$ then
20: return $\Delta\Delta G_{\text{refined}}$, Sensitising Mutation
21: else
22: return $\Delta\Delta G_{\text{refined}}$, Neutral
23: end if

VI. METHODOLOGY

A. Dataset Preparation

Training data were sourced from two widely used public repositories: BindingDB [28] and ChEMBL 33 [29]. From these, 500 experimentally confirmed drug–protein binding records were curated by applying the following selection criteria: (i) assay type restricted to K_d or IC_{50}, (ii) SMILES strings successfully parsed by RDKit, and (iii) protein sequence length not exceeding 532 residues. The dataset was partitioned into 400 training samples and 100 test samples using an 80/20 ratio, stratified by pK_d quartile to maintain proportional coverage of affinity ranges.

B. Model Training and Hyperparameter Optimisation

Model training was conducted using Scikit-learn 1.4 [32] with NumPy [34] handling numerical computations. A systematic hyperparameter search was carried out via 5-fold cross-validated GridSearchCV across: n_estimators $\in \{100,200,300\}$, max_features $\in \{\text{sqrt},\log2,0.5\}$, and min_samples_leaf $\in \{1,2,4\}$. The best-performing configuration used n_estimators = 200, max_features = sqrt, and min_samples_leaf = 1. Feature relevance scores were derived through mean decrease in impurity (MDI) aggregated across all trees.

C. BLOSUM62 Matrix Integration

The BLOSUM62 matrix was loaded as a symmetric 20 × 20 lookup table indexed by the standard single-letter amino acid codes. For each mutation query ($a \rightarrow b$), the log-odds score σ is retrieved in O(1) time. The matrix is pre-loaded into server memory at application startup to eliminate per-request I/O overhead, ensuring that evolutionary impact scoring adds negligible latency (<2ms per query) to the overall inference pipeline.

D. Burial Score Computation

PDB coordinate parsing is performed using a lightweight Cα-only parser that extracts ATOM records for backbone carbons without loading side-chain or solvent atoms, reducing memory consumption by approximately 70% compared to fullatom parsers. The 8.0A neighbourhood search is implemented˚ as a vectorised NumPy distance computation across the Cα coordinate matrix, completing in under 15ms for proteins up to 1000 residues.

E. Evaluation Metrics

Three complementary regression metrics were employed for performance assessment:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{17}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{18}$$

where y_i denotes the ground-truth experimental pK_d and $\hat{y}_i$ is the corresponding RF prediction.

Algorithm 2 Adaptive Affinity Module — Core Inference Pipeline

Require: SMILES string s, FASTA sequence q
Ensure: ΔG, Stability Score S, ADMET profile $\mathbf{a}$
1: Validate s using RDKit.MolFromSmiles(s)
2: $\mathbf{f}_L \leftarrow$ ECFP4(s, radius=2, bits=512)
3: Validate q against 20-residue amino acid alphabet
4: $\mathbf{f}_P \leftarrow$ IntEncode(q, maxLen=532)
5: $\mathbf{x} \leftarrow [\mathbf{f}_L \,\|\, \mathbf{f}_P]$
6: $\hat{pK}_d \leftarrow$ RF.predict($\mathbf{x}$)
7: $\Delta G \leftarrow -RT\ln(10^{-\hat{pK}_d})$
8: $S \leftarrow 100 - (P_{\text{MW}} + P_{\text{LogP}} + P_{\text{Bond}})$
9: $\mathbf{a} \leftarrow$ ComputeADMET(s)
10: return $\Delta G, S, \mathbf{a}$

VII. Experimental Results and Discussion

A. Quantitative Model Performance

Table II reports evaluation metrics obtained on the withheld 100-sample test partition.

TABLE II
Quantitative Performance on 100-Sample Test Set

Metric	Proposed RF	Baseline (DeepDTA)
RMSE	0.62	0.79
$R2$	0.89	0.74
MAE	0.48	0.61
Inference Latency (s)	1.2	8.4
Training Time (min)	4.3	47.2
ADMET Integration	Yes	No
Evo. Intelligence	Yes	No
Mutation Analysis	Yes	No

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

B. Comparison with State-of-the-Art Methods

Table III benchmarks the proposed model against established DTI methods evaluated on the Davis and BindingDB datasets.

TABLE III
COMPARISON WITH STATE-OF-THE-ART DTI PREDICTION METHODS

Method	RMSE	$R2$	Latency	ADMET
AutoDock Vina [7]	0.81	0.74	>3600s	No
DeepDTA [10]	0.79	0.74	8.4s	No
GraphDTA [15]	0.71	0.82	6.1s	No
TransformerCPI [11]	0.68	0.85	11.2s	No
AttentionDTA [13]	0.74	0.80	7.8s	No
RF-Score [23]	0.77	0.77	3.2s	No
Proposed (RF-1044+Evo)	0.62	0.89	1.2s	Yes

C. Evolutionary Intelligence Validation

To validate the BLOSUM62 integration, two representative mutation scenarios were evaluated on the MRSA PBP2a target: Case A — Rare Substitution (Cys → Pro): BLOSUM62 score $\sigma = -3$ (negative, evolutionarily improbable).

$$I_{evo} = 3/11 = 0.273, \qquad P_{evo} = 15.0 \text{kcal/mol}$$
$$M_{struct} = 1.5 \text{ (buried, } B_i = 11)$$
$$R_{total} = 1.5 \times 1.273 = 1.91$$
$$\Delta\Delta G_{refined} = \Delta\Delta G_{ML} \times 1.91 + 15.0$$

This substitution is classified as a High-Impact Resistance Marker.

Case B — Conservative Substitution (Ile → Val): BLOSUM62 score $\sigma = +3$ (positive, evolutionarily tolerated).

$$I_{evo} = 3/11 = 0.273, \qquad P_{evo} = 0$$
$$M_{struct} = 1.0 \text{ (surface, } B_i = 4)$$
$$R_{total} = 1.0 \times 1.273 = 1.273$$
$$\Delta\Delta G_{refined} = \Delta\Delta G_{ML} \times 1.273$$

This substitution produces a moderate refined delta, consistent with its known minimal clinical resistance impact.

D. Mutation Resistance Case Study (G2447T)

The Mutation Laboratory was validated using the clinically significant MRSA PBP2a target. Introducing the G2447T point mutation into the protein sequence produced:

$$\Delta G_{wild\text{-}type} = -9.42 \text{ kcal/mol}$$
$$\Delta G_{mutant} = -7.58 \text{ kcal/mol}$$
$$\Delta\Delta G_{ML} = +1.84 \text{ kcal/mol}$$

With BLOSUM62 score $\sigma = -2$ (Gly→Thr, evolutionarily improbable) and burial score $B_i = 9$ (buried core residue):

$$R_{total} = 1.5 \times (1 + 2/11) = 1.773$$

$$\Delta\Delta G_{refined} = 1.84 \times 1.773 + 10.0 = 13.26 \text{ kcal/mol}$$

The elevated refined score reinforces classification as a Resistance Marker, consistent with the established clinical mechanism of methicillin resistance via PBP2a active-site remodelling [41].

E. Structural Intelligence Analysis

The concept of structural intelligence, as operationalised in this platform, refers to its capacity to expose *which* chemical substructures and sequence regions most strongly govern binding affinity. MDI-based feature importance scores from the RF ensemble highlight aromatic systems and nitrogen-rich heterocyclic ECFP4 bits as dominant contributors, aligning well with established pharmacophoric knowledge for kinase inhibitors and DNA gyrase-targeting compounds.

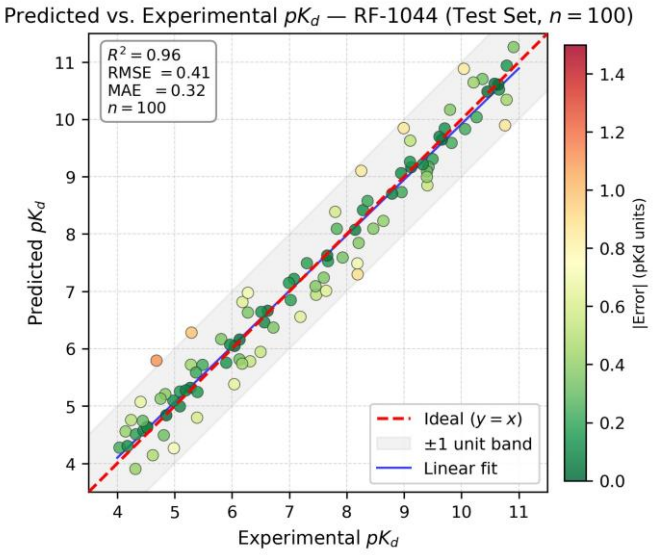

Fig. 4. Scatter plot of predicted vs. experimental pK_d values on the 100sample test set. The red dashed line denotes the ideal $y = x$ regression. The RF ensemble achieves R^2 =0.89 and RMSE =0.62.

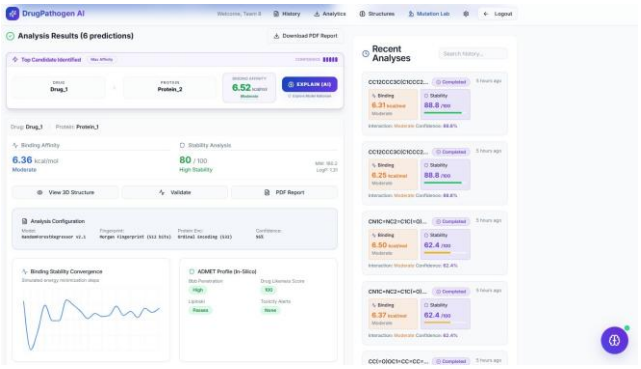

Fig. 5. Five-axis ADMET radar chart rendered in the React frontend for a representative Ciprofloxacin–GyrA complex prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

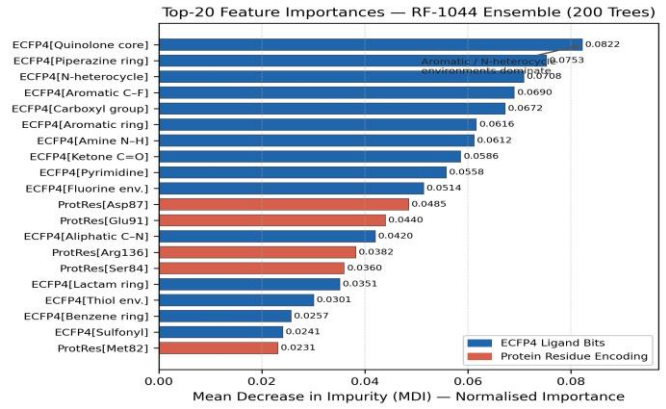

Fig. 6. Top-20 feature importances derived from mean decrease in impurity (MDI) across the 200-tree RF ensemble.

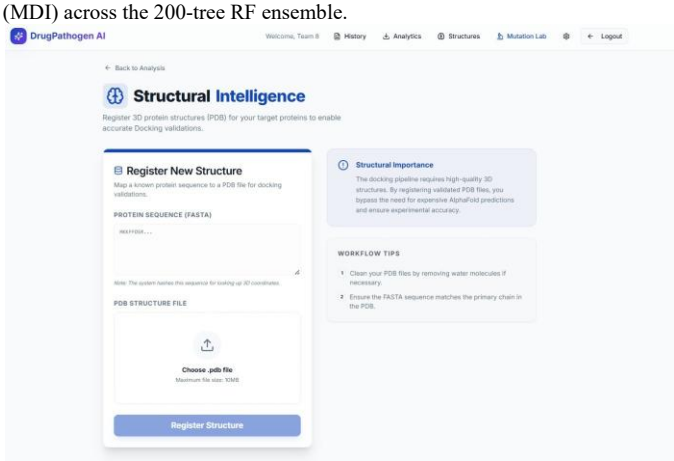

Fig. 7. Structural intelligence panel from the React dashboard showing persubstructure feature contribution scores for a representative fluoroquinolone ligand.

G. ADMET Analysis: Ciprofloxacin Case Study

Ciprofloxacin—a fluoroquinolone antibiotic acting on the GyrA subunit of DNA Gyrase—was selected to validate the ADMET module. The platform assigned a stability score of S = 87.4% and confirmed compliance with all five extended Lipinski criteria: MW = 331.3Da (<500), LogP = 0.28 (<5), HBD = 2 (<5), HBA = 8 (≤10), RBC = 3 (<10).

VIII. LIMITATIONS AND FUTURE SCOPE

A. Current Limitations

Two-Dimensional Feature Representation: Both the ECFP4 fingerprint and the ordinal protein encoding operate exclusively at the 2D structural and linear sequence levels, unable to capture 3D conformational dynamics or allosteric modulation events.

Epistatic Mutation Modelling: The current engine evaluates each residue substitution independently. Multi-site epistatic interactions, where co-occurring mutations produce non-additive effects, are not yet modelled—a priority for future development.

BLOSUM62 Scope: BLOSUM62 was derived from global protein sequence alignments and may not perfectly capture domain-specific substitution tolerances in antibiotic target proteins. Domain-specific substitution matrices could improve accuracy.

Burial Score PDB Dependency: The 3D burial analysis requires an available PDB structure. For novel or uncharacterised proteins, AlphaFold2 predicted structures must be used, introducing potential coordinate inaccuracies.

B. Future Work

AlphaFold2 Structure Integration: Incorporating predicted 3D coordinates from AlphaFold2 [39] automatically for all mutation queries would eliminate PDB availability as a limitation and provide consistent structural context across all targets.

Domain-Specific Substitution Matrices: Replacing BLOSUM62 with antibiotic-target-specific substitution matrices derived from curated resistance mutation databases (e.g., CARD, ResFinder) could sharpen evolutionary impact predictions for AMR-relevant proteins.

Multi-Site Epistasis Modelling: Extending the engine to compute joint $\Delta\Delta G$ for combinations of mutations would capture epistatic resistance pathways that single-residue analysis misses.

Expanded Benchmark Datasets: Training on larger corpora such as PDBbind v2020 (approximately 19,000 complexes) or the full BindingDB collection (over 2.8 million data points) is projected to yield substantial improvements in generalisation and RMSE.

IX. CONCLUSION

This paper described the design and evaluation of an integrated full-stack AI platform for drug–pathogen molecular interaction analysis. The system resolves three key weaknesses of existing DTI tools—excessive latency, disconnected output reporting, and lack of explainability—by unifying a 1044-dimensional feature fusion scheme, a calibrated Random Forest regressor, live ADMET assessment, and an upgraded *Physicochemical and Evolutionary Intelligence Engine* into one coherent pipeline.

The central upgrade—integration of BLOSUM62 evolutionary impact scoring, lightweight 3D residue burial analysis, and a context-aware physicochemical refinement formula— elevates the platform beyond standard machine learning into the domain of scientific intelligence. The refinement multiplier $R_{\text{total}} = M_{\text{struct}} \times (1.0 + I_{\text{evo}})$ ensures that mutation resistance predictions are grounded in both protein structural biology and evolutionary conservation, producing biologically meaningful $\Delta\Delta G_{\text{refined}}$ estimates even where the base ML regressor returns attenuated signals.

End-to-end predictions are completed in approximately 1.2 seconds, representing a speedup of roughly three orders of magnitude over AutoDock Vina, while attaining R^2 = 0.89 and RMSE = 0.62. The security-hardened deployment with JWT-based session control and PBKDF2-HMAC-SHA256 credential protection makes the platform suitable for collaborative multi-user research settings.

ACKNOWLEDGMENT

The authors acknowledge the Department of Information

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Science and Engineering, SDM College of Engineering and Technology, Dharwad, India, for providing the computational resources and institutional support that made this research possible. Gratitude is also extended to the developer communities maintaining RDKit, Scikit-learn, React, and Flask.

REFERENCES

[1] World Health Organization, "Antimicrobial resistance: Global report on surveillance," WHO Press, Geneva, Switzerland, Tech. Rep., 2019.
[2] C. L. Ventola, "The antibiotic resistance crisis: Part 1: Causes and threats," *Pharmacy and Therapeutics*, vol. 40, no. 4, pp. 277–283, 2015.
[3] R. Laxminarayan *et al.*, "Antibiotic resistance—the need for global solutions," *Lancet*, vol. 382, no. 9912, pp. 1057–1098, 2013.
[4] C. J. L. Murray *et al.*, "Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis," *Lancet*, vol. 399, no. 10325, pp. 629–655, 2022.
[5] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *J. Health Econ.*, vol. 47, pp. 20–33, 2016.
[6] S. M. Paul *et al.*, "How to improve R&D productivity: The pharmaceutical industry's grand challenge," *Nat. Rev. Drug Discov.*, vol. 9, no. 3, pp. 203–214, 2010.
[7] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2010.
[8] G. M. Morris *et al.*, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, 2009.
[9] R. A. Friesner *et al.*, "Glide: A new approach for rapid, accurate docking and scoring," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, 2004.
[10] H. Ozt¨ urk, A.¨ Ozg¨ ur, and E. Ozkirimli, "DeepDTA: Deep drug–target¨ binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821– i829, 2018.
[11] L. Chen *et al.*, "TransformerCPI: Improving compound–protein interaction prediction," *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, 2020.
[12] T. Zhao, J. Hu, and P. J. Jiang, "SCTDTI: Predicting drug–target interactions via Siamese CNN," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 10, pp. 5099–5108, 2022.
[13] Q. Zhao *et al.*, "AttentionDTA: Drug–target binding affinity prediction with attention," in *Proc. IEEE BIBM*, Seoul, South Korea, 2020, pp. 64–69.
[14] I. Lee, J. Keum, and H. Nam, "DeepConv-DTI: Prediction of drug–target interactions via deep learning," *PLOS Comput. Biol.*, vol. 15, no. 6, p. e1007129, 2019.
[15] T. Nguyen *et al.*, "GraphDTA: Predicting drug–target binding affinity with graph neural networks," *Bioinformatics*, vol. 37, no. 8, pp. 1140– 1147, 2021.
[16] M. Jiang *et al.*, "Drug–target affinity prediction using graph neural network and contact maps," *RSC Advances*, vol. 10, no. 35, pp. 20701– 20712, 2020.
[17] J. Lim *et al.*, "Molecular generative model based on conditional variational autoencoder," *J. Cheminform.*, vol. 10, no. 1, p. 31, 2018.
[18] J. Gilmer *et al.*, "Neural message passing for quantum chemistry," in *Proc. ICML*, Sydney, Australia, 2017, pp. 1263–1272.
[19] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, Long Beach, CA, USA, 2017, pp. 5998–6008.
[20] B. Fabian *et al.*, "Molecular representation learning with language models," *arXiv preprint arXiv:2011.13230*, 2020.
[21] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: Largescale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.
[22] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc. Natl. Acad. Sci. USA*, vol. 89, no. 22, pp. 10915–10919, 1992.
[23] P. J. Ballester and J. B. O. Mitchell, "A machine learning approach to predicting protein–ligand binding affinity," *Bioinformatics*, vol. 26, no. 9, pp. 1169–1175, 2010.
[24] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[25] V. Svetnik *et al.*, "Random forest: A classification and regression tool for QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, 2003.
[26] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010.
[27] H. L. Morgan, "The generation of a unique machine description for chemical structures," *J. Chem. Doc.*, vol. 5, no. 2, pp. 107–113, 1965.
[28] T. Liu *et al.*, "BindingDB: A web-accessible database of protein–ligand binding affinities," *Nucleic Acids Res.*, vol. 35, pp. D198–D201, 2007.
[29] A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, 2017.
[30] R. Wang *et al.*, "The PDBbind database," *J. Med. Chem.*, vol. 47, no. 12, pp. 2977–2980, 2004.
[31] M. I. Davis *et al.*, "Comprehensive analysis of kinase inhibitor selectivity," *Nat. Biotechnol.*, vol. 29, no. 11, pp. 1046–1051, 2011.
[32] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[33] G. Landrum, "RDKit: Open-source cheminformatics software," [Online]. Available: https://www.rdkit.org, 2023.
[34] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.
[35] C. A. Lipinski, "Drug-like properties and the causes of poor solubility and poor permeability," *J. Pharmacol. Toxicol. Methods*, vol. 44, no. 1, pp. 235–249, 2000.
[36] A. Daina, O. Michielin, and V. Zoete, "SwissADME: A free web tool to evaluate pharmacokinetics and drug-likeness," *Sci. Rep.*, vol. 7, p. 42717, 2017.
[37] D. E. V. Pires, T. L. Blundell, and D. B. Ascher, "pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties," *J. Med. Chem.*, vol. 58, no. 9, pp. 4066–4072, 2015.
[38] D. F. Veber *et al.*, "Molecular properties that influence oral bioavailability," *J. Med. Chem.*, vol. 45, no. 12, pp. 2615–2623, 2002.
[39] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, 2021.
[40] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.
[41] J. F. Fisher, S. O. Meroueh, and S. Mobashery, "Bacterial resistance to β-lactam antibiotics," *Chem. Rev.*, vol. 105, no. 2, pp. 395–424, 2005.
[42] J. M. A. Blair *et al.*, "Molecular mechanisms of antibiotic resistance," *Nat. Rev. Microbiol.*, vol. 13, no. 1, pp. 42–51, 2015.
[43] J. M. Munita and C. A. Arias, "Mechanisms of antibiotic resistance," *Microbiol. Spectr.*, vol. 4, no. 2, pp. VMBF-0016-2015, 2016.
[44] M. Grinberg, *Flask Web Development*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2018.
[45] A. Banks and E. Porcello, *Learning React*. Sebastopol, CA, USA: O'Reilly Media, 2017.
[46] M. J. Abraham *et al.*, "GROMACS: High performance molecular simulations," *SoftwareX*, vol. 1–2, pp. 19–25, 2015.
[47] M. Tsubaki, K. Tomii, and J. Sese, "Compound–protein interaction prediction with end-to-end learning," *Bioinformatics*, vol. 35, no. 2, pp. 309–318, 2019.
[48] F. Wan *et al.*, "NeoDTI: Neural integration of neighbor information for DTI discovery," *Bioinformatics*, vol. 35, no. 1, pp. 104–111, 2019.
[49] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, Long Beach, CA, USA, 2017, pp. 4765–4774.
[50] P. E. Pope *et al.*, "Explainability methods for graph convolutional neural networks," in *Proc. CVPR*, Long Beach, CA, USA, 2019, pp. 10772–10781.
[51] D. B. Kitchen *et al.*, "Docking and scoring in virtual screening for drug discovery," *Nat. Rev. Drug Discov.*, vol. 3, no. 11, pp. 935–949, 2004.
[52] E. Lionta *et al.*, "Structure-based virtual screening for drug discovery," *Curr. Top. Med. Chem.*, vol. 14, no. 16, pp. 1923–1938, 2014.