Ailixir

Mansoura University

Faculty of Computer & Information Sciences Department of Computer Science & Information Technology

Bachelor of Science

in

Computer Sience

2026

Mansoura University, Egypt

Introduction

Discovering a new drug is a notoriously complex, exhaustive, and high-risk process. It is a monumental gamble where average development times span 10 to 15 years and the cost for a single approved medicine often exceeds $2.6 billion [1]. Compounding this challenge, over 90% of drug candidates that enter human clinical trials are ultimately found to be unsafe or ineffective, resulting in catastrophic financial and human-capital losses [2]. Historically, identifying the active ingredients of traditional remedies was the primary method of discovery. More recently, this has been industrialized through high-throughput screening (HTS).

HTS is a process where large libraries of chemicals are tested for desirable effects within acceptable parameters of potency, metabolic stability, and selectivity. Selectivity is one of the more important factors as it is a measurement of the effect the compound has on the selected target compared to its effect on other cells. A less selective drug may be toxic for consumption. After a compound that fulfills these requirements has been identified, clinical trials may be developed. However, it is unlikely for a perfect drug candidate to appear from early HTS runs. Even after a desirable "hit" compound has been found, medicinal chemists must use structure-activity relationships (SAR) to manually improve its properties.

As one can see, modern drug discovery, while more systematic, remains an expensive process that requires massive investments. Despite advances in technology and our growing understanding of biology, it remains a costly and lengthy endeavor with a low success rate. This inefficiency stems from its core reliance on physical trial-and-error screening, which is inherently slow, expensive, and limited in scale.

To overcome these fundamental bottlenecks, a paradigm shift is necessary. The industry and academia are increasingly turning to computational methods and Artificial Intelligence (AI). These in silico (computational) approaches replace the physical test tube with powerful algorithms. They allow for massive-scale virtual screening of chemical libraries numbering in the billions, predictive toxicology to spot non-viable candidates early, and molecular modeling to understand complex protein-ligand interactions at an atomic level. This grants researchers the crucial ability to "fail fast and fail cheap" by simulating outcomes digitally before committing to expensive and time-consuming lab work.

This project leverages these advanced computational techniques to propose a truly integrated, AI-driven

ecosystem. Our system is not just another siloed tool; it is a holistic pipeline designed to automate and accelerate the entire discovery workflow. We aim to build a platform that can predict new compound efficacy, filter for ADMET toxicity, simulate molecular docking, identify new uses for old drugs (drug repurposing), and even optimize promising lead compounds. By architecting this entire powerful system as an intuitive mobile application, we aim to democratize these tools, making them significantly more accessible to researchers and students worldwide.

Problem Statement

The process of drug discovery is one of the most complex, expensive, and time-consuming challenges in modern science. Traditionally, developing a single new drug requires 10 to 15 years of continuous research and testing, with an average cost exceeding 2.6 billion USD. Despite this massive investment, over 90% of drug candidates fail during clinical trials, often after years of effort and financial loss. These limitations highlight an urgent need for a more efficient, intelligent, and accessible approach to discovering new medicines.

1. Slow and Inefficient Discovery Process
   Researchers must manually test thousands of chemical compounds in laboratories to identify potential treatments for a specific disease. This process takes months or even years, and most compounds turn out to be ineffective or toxic. As a result, valuable time and resources are wasted on unpromising candidates.
2. Extremely High Research Costs
   Laboratory experiments, chemical analyses, and safety tests are extremely expensive and require specialized equipment and large research teams. Because of this, small research institutions and universities cannot afford to conduct large-scale drug discovery projects, making this field dominated by major pharmaceutical companies.
3. Limited Accessibility and Analytical Tools
   Not all researchers or students have access to the tools, data, or computational resources needed to analyze compounds or understand protein–ligand interactions. This creates a barrier for students and academics who wish to explore this field or develop AI-based research skills.
4. Lack of Educational and Interactive Platforms
   Most existing AI-powered drug discovery platforms are either closed-source or too complex for beginners. There is no simple, user-friendly system that allows researchers and students to experiment, visualize, and understand how AI can accelerate drug discovery. This limits learning opportunities and innovation in the academic community.
5. Difficulty in Drug Repurposing
   Many existing drugs have the potential to treat new diseases (as seen during the COVID-19 pandemic), but due to the absence of intelligent analytical tools, drug repurposing remains a slow and uncertain process, often based on luck or trial and error.

Motivation

The primary motivation for this project is born from a critical and undisputed failure in the pharmaceutical industry: the traditional drug discovery pipeline is fundamentally broken. The process as it stands is economically unsustainable and shockingly inefficient. Statistics paint a grim picture of this reality:

- Developing a single new drug is a monumental gamble, consuming **10 to 15 years** of research and an average of **$2.6 billion** [1]. This decade-long process acts as an extreme funnel of attrition, where tens of thousands of initial compounds are screened just to find a handful worthy of testing.

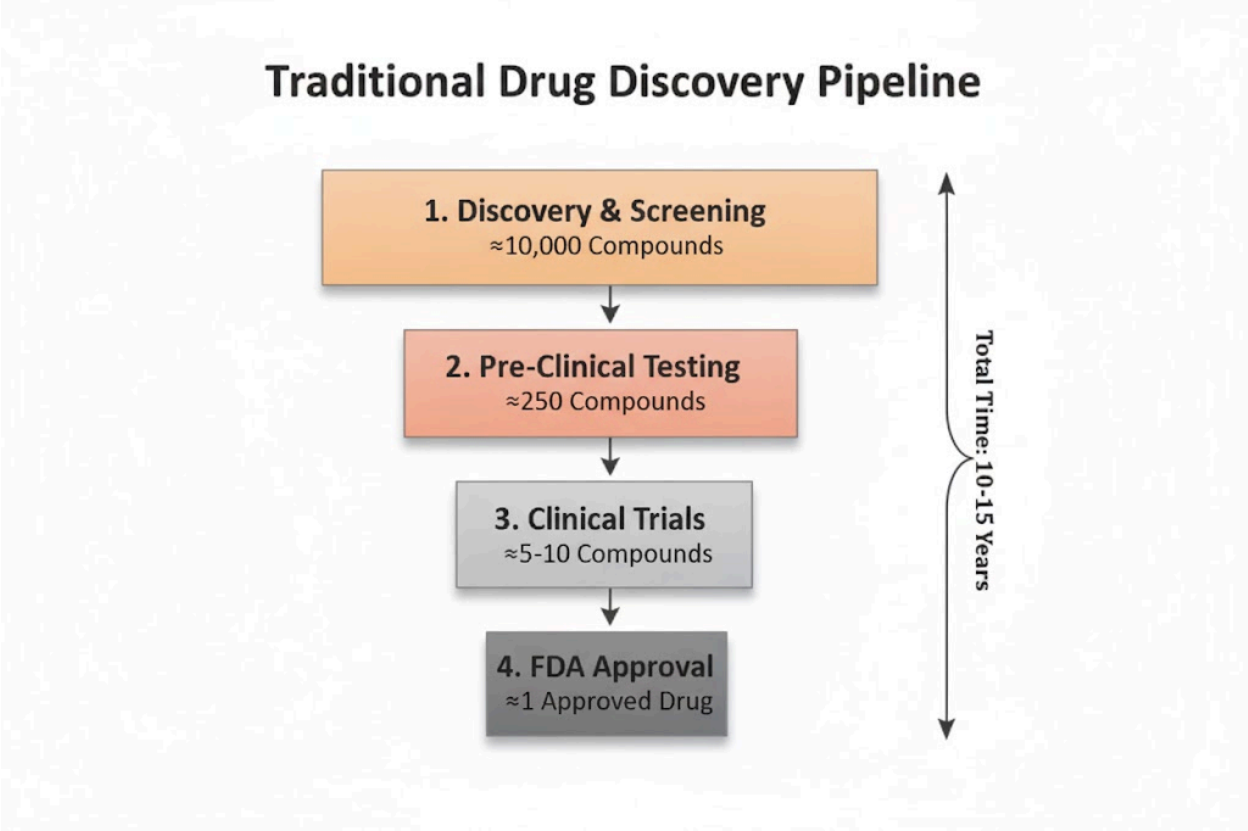


Figure 1: The Traditional Drug Discovery Funnel. This 10-15 year pipeline illustrates the process of attrition where ~10,000 initial compounds are screened to yield only one FDA-approved drug.

- Even after this massive investment of time and capital, the catastrophic failure rate reveals the core problem. Over **90% of all drug candidates** that enter human clinical trials fail.

This failure, however, is not random. It is predictable. A deeper analysis reveals why these drugs fail: the two largest drivers are **lack of efficacy** (the drug doesn't work, ~50%) and **unacceptable safety/toxicity** (the drug is harmful, ~30%) [2].

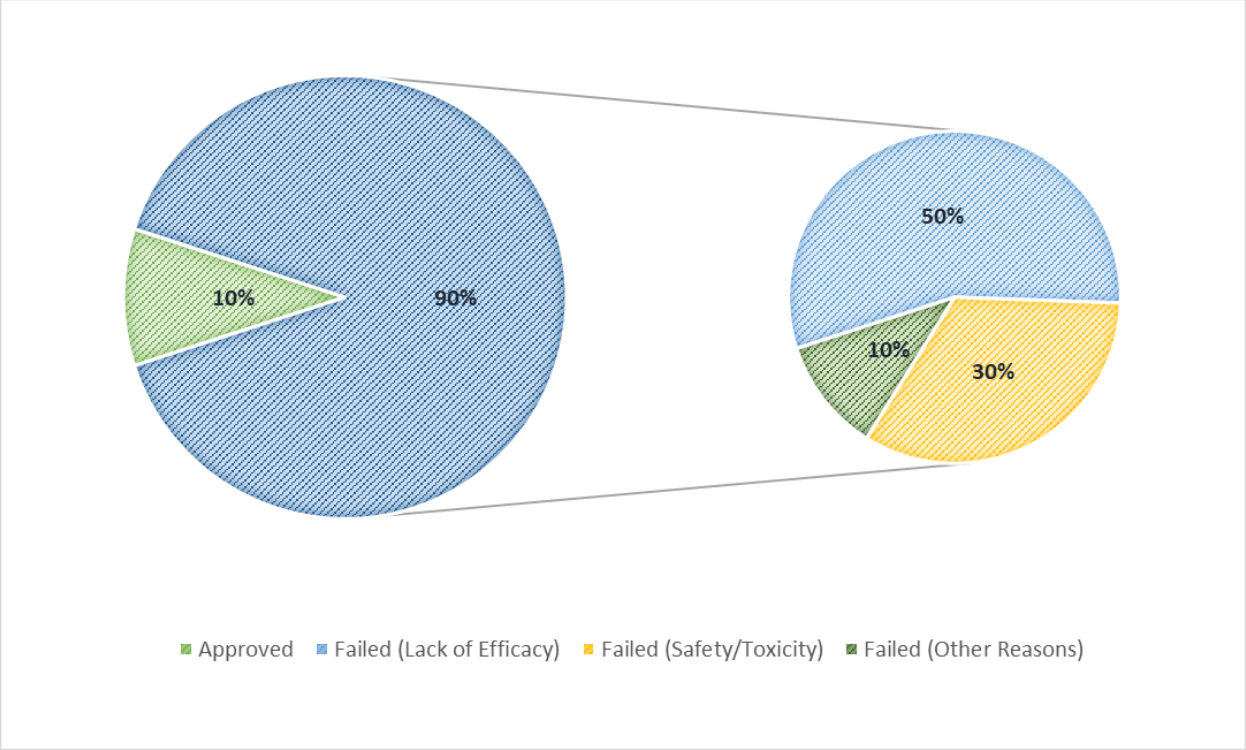


Figure 2: Root Causes of Clinical Trial Attrition. While over 90% of candidates fail (main pie), the breakdown (secondary pie) reveals the primary causes: lack of efficacy and safety/toxicity [2].

This is where our motivation lies.

These statistics are not just numbers; they are a clear map of the problem. They show that the industry wastes billions of dollars and years of effort advancing drugs that are destined to fail for reasons that can be predicted.

While Artificial Intelligence has emerged as a solution, existing tools (as reviewed in our Related Work section) remain fragmented. They often focus on one piece of the puzzle (like generation) while ignoring another (like safety). Furthermore, many are computationally prohibitive, closed-source, or too complex for non-experts, creating a significant barrier to entry.

This project is directly motivated by these gaps. We aim to build a holistic, end-to-end pipeline that directly attacks the two main causes of failure. Our system is designed to integrate:

1. **New Candidate Prediction** (to solve the efficacy problem).
2. **ADMET/Safety Filtering** (to solve the toxicity problem).

Furthermore, we are motivated to *democratize* this technology. By architecting our system as a user-friendly mobile application, we aim to break down the accessibility barrier, placing these powerful predictive tools directly into the hands of students, academics, and researchers.

In essence, we are motivated to transform a slow, fragmented, and broken process into a fast, integrated, and accessible one, guided by the data on why drugs fail.

Objectives

The main objective of this project is to design and develop an AI-powered mobile application that automates and accelerates the process of drug discovery using deep learning and molecular simulation techniques.

To achieve this primary goal, the following specific objectives are defined:

1. **Predict New Compound Effectiveness:** Develop and integrate AI models capable of predicting the potential biological activity of new chemical compounds against specific disease targets.
2. **Enable Drug Repurposing:** Implement a feature that allows the AI to analyze an existing drug and predict other diseases it may effectively treat.
3. **Optimize Promising Lead Compounds:** Implement AI-based optimization models that suggest structural modifications to enhance the efficacy or reduce side effects of moderately active compounds.
4. **Implement Compound Similarity Search:** Develop a feature to find compounds similar to a given input, offering alternative options if the original is rare, costly, or unavailable.
5. **Integrate ADMET (Safety) Filtering:** Implement AI models to filter out unsafe or non-viable compounds before experimental testing by predicting their ADMET properties.
6. **Perform Molecular Docking Simulations:** Integrate simulation tools to validate molecular interactions and estimate drug effectiveness by calculating binding affinity.
7. **Develop a User-Friendly Mobile Application:** Create an intuitive client-side mobile application that allows users to submit requests and receive responses from the backend services.
8. **Provide 3D Visualization Services:** Implement a dedicated service to transform SMILES codes into 2D/3D structures and render protein-ligand interactions from docking results.

Literature Review – Related Work

1. BindDM: Reinforcement Learning-Based Molecular Optimization

- **Problem Addressed:** Most generative drug design models fail to optimize molecules for protein binding affinity directly. BindDM was proposed to overcome this limitation by integrating reinforcement learning with docking feedback as a reward function.
- **Methodology:** The model generates molecular candidates and evaluates them through iterative docking simulations. A reinforcement learning agent adjusts the molecular structures based on docking scores, guiding the generation toward higher affinity compounds.
- **Key Results:** BindDM achieved improved docking-based affinity metrics and enhanced diversity compared to standard VAEs and GANs. However, the method was computationally intensive and relied heavily on accurate docking approximations.
- **How This Project Differs:** Our project reduces computational dependency by using efficient post-generation filtering based on QED, SA, and similarity scoring, rather than incorporating docking in every training iteration. This improves scalability and allows faster large-scale molecular screening.

2. DrugGEN: Structure-Aware Molecular Generation

- **Problem Addressed:** Traditional generative models often ignore the protein's structural information, producing molecules with uncertain binding relevance. DrugGEN aims to generate ligands that are aware of protein pocket structures.
- **Methodology:** It employs a graph-based deep learning framework where both the ligand and protein pocket are encoded as graphs. The model then learns to generate molecules conditioned on protein features to improve structural compatibility.
- **Key Results:** DrugGEN achieved high chemical validity (100%), strong internal diversity (>0.85), and balanced novelty. Yet, it did not include synthetic accessibility or QED-based evaluation in its generation phase.
- **How This Project Differs:** Our model extends DrugGEN's approach by incorporating comprehensive post-generation evaluation — including QED, SA, and Lipinski rule checks — to ensure both biological relevance and chemical feasibility before docking.

3. MolMIM: Molecular Mutation and Optimization Model

- **Problem Addressed:** Existing models often get trapped in limited chemical spaces, producing repetitive or known scaffolds. MolMIM introduces a mutation-based approach to explore new chemical regions.
- **Methodology:** It starts with known inhibitors and applies generative mutation operators guided by molecular property optimization (QED and binding affinity). The system iteratively mutates molecules until desired thresholds are reached.
- **Key Results:** MolMIM successfully generated optimized molecules with better QED and docking scores but showed lower novelty due to strong bias toward initial scaffolds.
- **How This Project Differs:** Unlike MolMIM, our workflow focuses on generating molecules from scratch, not by mutating existing ones. This allows for higher novelty and greater exploration of the chemical space while maintaining drug-likeness constraints.

4. LiGAN: Ligand Generation Using 3D Density Grids

- **Problem Addressed:** Most molecular generative models overlook the 3D spatial compatibility of the ligand with its binding site. LiGAN was designed to directly model spatial constraints through voxelized representations.
- **Methodology:** LiGAN uses a 3D convolutional neural network to predict ligand atom densities inside protein pockets, enabling generation of physically realistic 3D structures.
- **Key Results:** The model achieved high structural compatibility but struggled with computational complexity and limited scalability. The grid-based representation also restricted chemical diversity.
- **How This Project Differs:** Our approach focuses on SMILES-level generation combined with post-processing chemical validation. This enables generation of thousands of valid molecules with less computational cost and higher chemical diversity.

5. IDOLpro: Multi-Objective Drug Design Framework

- **Problem Addressed:** Balancing multiple drug design objectives — such as binding affinity, novelty, and druglikeness — is challenging for most single-objective models.
- **Methodology:** IDOLpro uses a reinforcement learning framework with a multi-objective reward function that simultaneously optimizes affinity, QED, and diversity scores.
- **Key Results:** The model demonstrated strong performance across several metrics but was limited by the need for precomputed docking data, which slows real-time molecule generation.
- **How This Project Differs:** Our project applies a multi-criteria selection system (QED > 0.5, SA <

4, similarity 0.3–0.7, Lipinski compliance) after molecule generation instead of during training. This post-filtering method increases flexibility and reduces computational overhead.

Summary and How This Project Differs

Previous works like BindDM, DrugGEN, and LiGAN achieved impressive results in AI-driven molecule generation, but they often lacked integrated evaluation of drug-likeness, manufacturability, and binding relevance in a unified pipeline.

Our project bridges this gap through a hybrid design that couples deep generative modeling with systematic chemical and pharmacological filtering. Specifically:

- We emphasize post-generation chemical validation (QED, SA, Lipinski compliance) before docking.
- Our workflow allows scalable, target-specific molecule generation for AKT1 inhibitors without retraining.
- It ensures that generated molecules are not only chemically valid and novel, but also synthetically accessible and druglike, offering a more practical step toward experimental validation.

Comparison of Related Work

Project Name	**Primary Focus (Approach)**	**Drug-likeness Filtering (QED, SA, etc.)**	**Docking Integration**	**Key Limitation(s)**
BindDM	Generative Optimization (RL)	Limited (Focus on affinity)	During Training (As reward)	Computationally intensive
DrugGEN	Generative (Structure-Aware Graph)	Not Included (in generation)	N/A (Generates for docking)	Lacks feasibility checks (SA/QED)

MolMIM	Optimization (Mutation-based)	Yes (Guided by QED)	During Optimization	Low novelty (Biased to initial scaffolds)
LiGAN	Generative (3D Density Grids)	Not Included	N/A (Generates 3D structures)	Computationally complex; low diversity
IDOLpro	Generative Optimization (Multi-Objective RL)	Yes (As part of reward)	During Training (Requires precomputation)	Slow; requires precomputed data
Our Project (Proposed)	**Generative + Integrated Filtering Pipeline**	**Yes (Systematic Post-Generation)**	**Post-Filtering (Validation Step)**	**-**

System Architecture

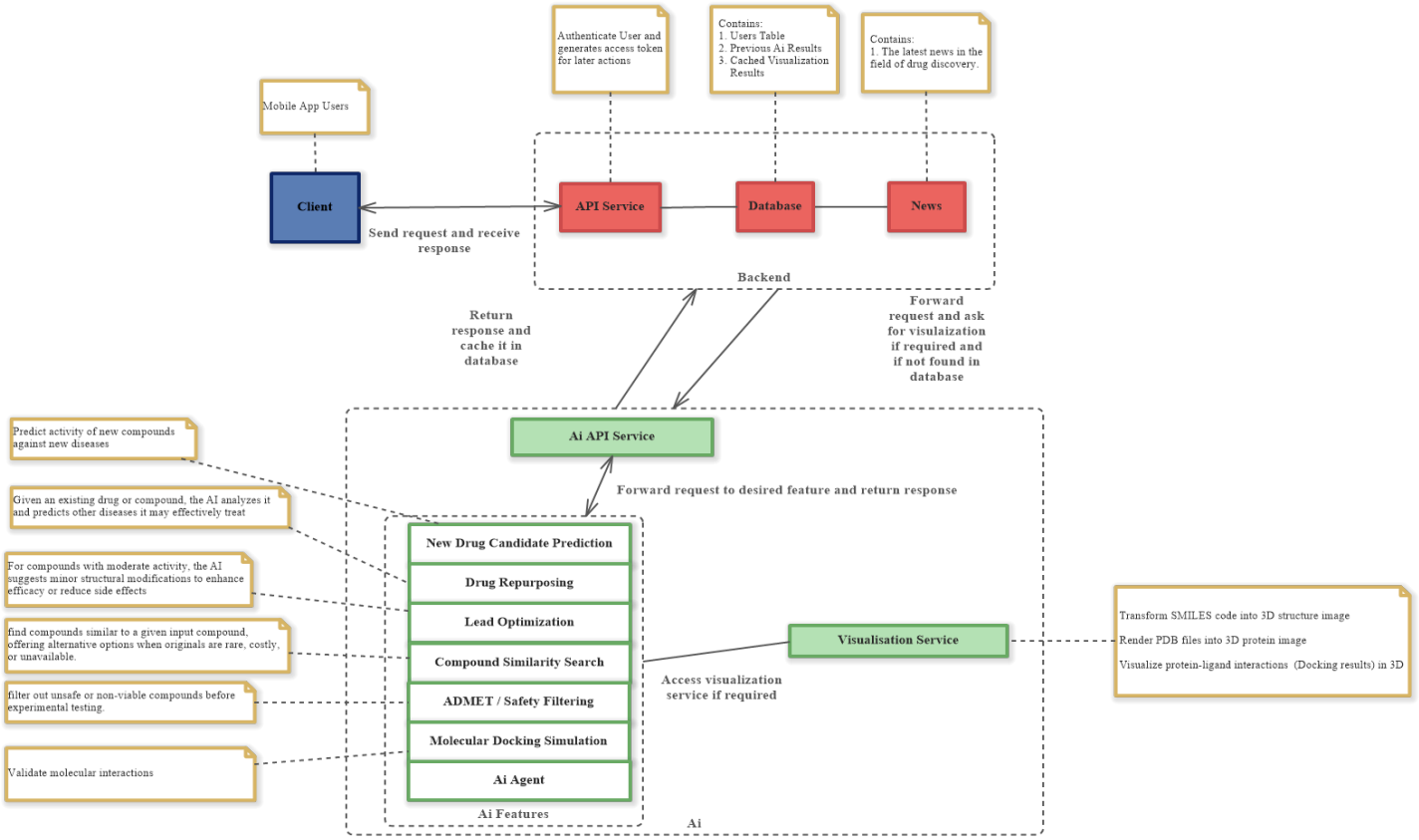
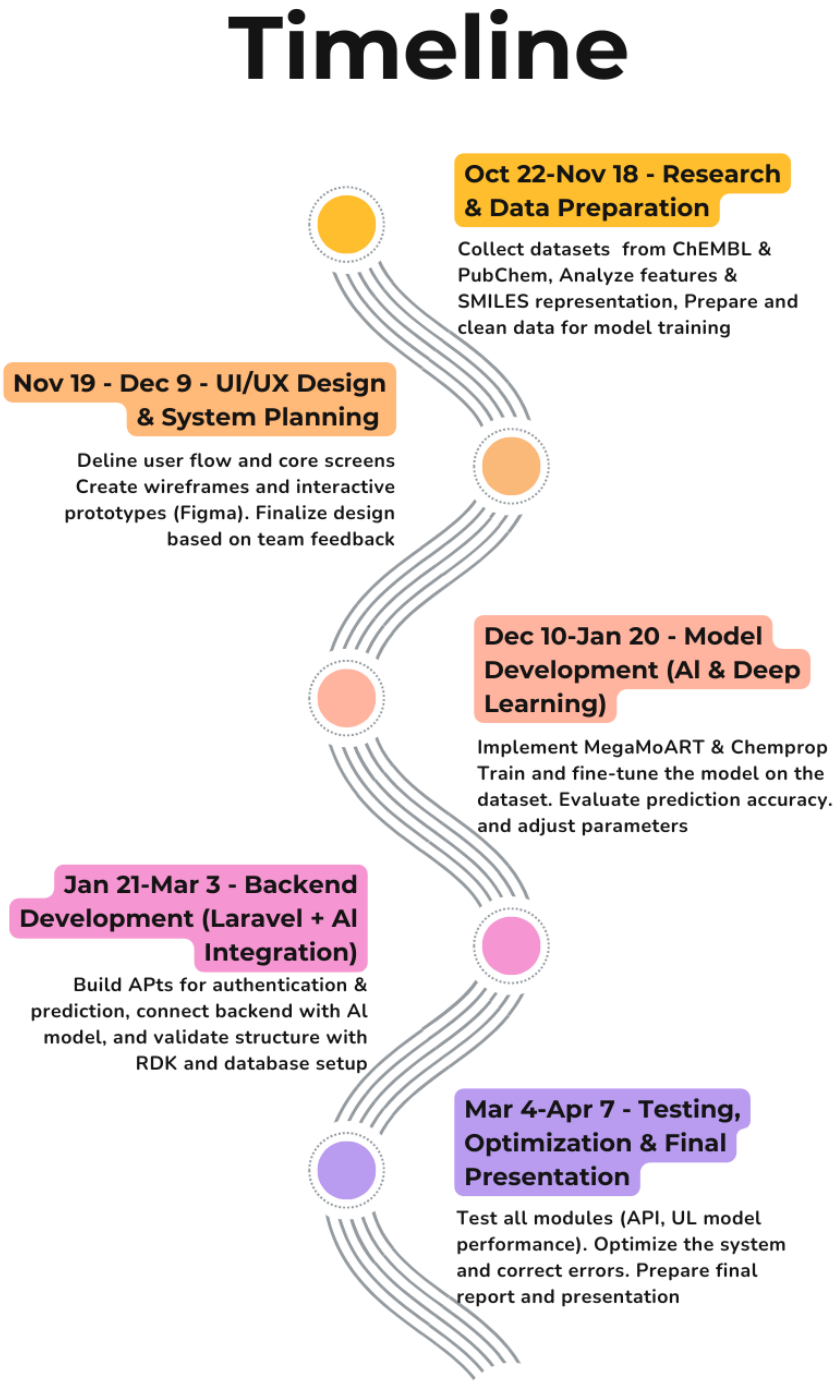


Figure 3: System Architecture Workflow. This architecture illustrates the workflow of our AI-driven drug discovery system. The Client (mobile app) sends user requests to the API Service, which handles authentication, validation, and data routing. The API forwards requests to the AI Service, which contains specialized modules for tasks such as New Drug Candidate Prediction, Drug Repurposing, Lead Optimization, Compound Similarity Search, ADMET/Safety Filtering, and Molecular Docking Simulation. Each module performs a specific analysis to evaluate or improve potential drug compounds. When visualization is required, results are sent to the Visualization Service to generate graphical or 3D outputs. The final processed results whether analytical, visual or both are returned to the Backend, stored in the Database (for caching and to reduce the work on visualization service), and then sent back to the Client.

Timeline


Timeline
Oct 22-Nov 18 - Research & Data Preparation
Collect datasets from ChEMBL & PubChem, Analyze features & SMILES representation, Prepare and clean data for model training
Nov 19 - Dec 9 - UI/UX Design & System Planning
Deline user flow and core screens Create wireframes and interactive prototypes (Figma). Finalize design based on team feedback
Dec 10-Jan 20 - Model Development (Al & Deep Learning)
Implement MegaMoART & Chemprop Train and fine-tune the model on the dataset. Evaluate prediction accuracy. and adjust parameters
Jan 21-Mar 3 - Backend Development (Laravel + Al Integration)
Build APts for authentication & prediction, connect backend with Al model, and validate structure with RDK and database setup
Mar 4-Apr 7 - Testing, Optimization & Final Presentation
Test all modules (API, UL model performance). Optimize the system and correct errors. Prepare final report and presentation

Conclusion

The traditional drug discovery process is notoriously slow, costly, and inefficient, often taking more than a decade and billions of dollars to bring a single drug to market. Moreover, nearly 90% of candidate drugs fail during clinical trials, highlighting the urgent need for smarter, faster, and more reliable approaches.

Our proposed AI-powered drug discovery pipeline addresses this challenge by integrating artificial intelligence with computational chemistry. Through three key stages — generation, filtering, and simulation — the system can autonomously generate new chemical compounds, evaluate their validity and safety, and predict how effectively they interact with disease-related proteins.

By leveraging advanced AI models such as MegaMolBART, Chemprop, and molecular docking tools, our project aims to accelerate drug discovery, reduce research costs, and improve accuracy in identifying promising compounds. Beyond its scientific value, the system also supports innovation in healthcare and education, providing researchers with a powerful tool to explore new treatments more efficiently.

Ultimately, this project represents a step forward toward a future where AI transforms pharmaceutical research, making the discovery of life-saving drugs faster, more precise, and more accessible to all.

References

[1] DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics, 47*, 20-33.

[2] Waring, M. J., Arrowsmith, J., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M., ... & Armstrong, D. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery, 14*(7), 475-486.

[3] Luo, S., Chen, Z., Li, Y., Wang, Y., Zhang, J., & Hou, T. (2021). BindDM: A deep reinforcement learning-based molecular optimization method for protein binding affinity. *Briefings in Bioinformatics, 22*(5), bbab108.

[4] An, X., Wu, F., Zhou, C., Wang, Y., & Ouyang, P. (2022). DrugGEN: A graph-based deep learning framework for structure-aware molecular generation. *Briefings in Bioinformatics, 23*(1), bbab494.

[5] Yan, W., Li, J., Fang, J., & Zhang, Y. (2021). MolMIM: A molecular mutation and optimization model based on deep learning. *Journal of Chemical Information and Modeling, 61*(11), 5344-5353.

[6] Ragoza, M., Hoch, J. C., Rogers, C., & Wei, G. W. (2019). LiGAN: A generative model for 3D-aware

generation of molecules. *arXiv preprint arXiv:1907.03949*.

[7] Lee, S., Kim, H., Na, B., Lee, D. H., Kang, K., & Lee, J. (2022, August). IDOLpro: A Multi-objective Drug Design Framework using Deep Reinforcement Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)* (pp. 3290-3300).