RESEARCH PAPER

Integrating Artificial Intelligence and Machine Learning for Drug Discovery and Biochemical Analysis within a Cyber-Secure Medical Data Infrastructure

Zainab Yousaf[1], Syed Noman Ahmed[2], Chaudhry Qasim Siddique[3], Wajeeh Ur Rehman[4], Fakhra Fakhr[5], Muhammad Humayoon Siddique[6]

[1] Department of Computer Science, Software Engineering, National University of Science and Technology Islamabad, Pakistan. Email: yousafzainab9@gmail.com

[2] Department of Biotechnology, University of Karachi, Karachi, Pakistan. Email: syednomanahmed057@gmail.com | ORCID: 0009-0004-4872-5833

[3] Department of Chemistry, Forman Christian College and University, Lahore, Pakistan. Email: qasimsiddique@outlook.com

[4] Department of Public Health, Arden University, Germany. Email: wajeehchohan0@gmail.com

[5] Department of Operation Theatre Technology and Anesthesia Technology, Assistant Professor and Head of Department, Riphah International University, Lahore. Email: Fakhra.fakhr@riphah.edu.pk

[6] Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan. Email: humayoon.kicsit@gmail.com

Received: 12th Mar, 2026 | **Revised:** 24th Mar, 2026 | **Accepted:** 14th Apr, 2026 | **Available Online:** 30th Apr, 2026

ABSTRACT

Background:

Drug discovery is a traditionally slow, expensive and highly experimental process. The application of Artificial Intelligence (AI) and Machine Learning (ML) in computational drug discovery is a game-changing development, offering a quicker and more precise way to identify potential drug candidates, while also tackling key issues related to the security of sensitive patient data.

Objective:

The objective of this study was to design and assess an AI and ML-based computational system for drug discovery and biochemical analysis, integrated with a cyber-secure medical data system to maintain data security and privacy during the research process.

Methodology:

We used purposive sampling to choose 5,000 entries of compounds from three reliable sources: Public Chemistry Database (PubChem), Chemical Biology Database (ChEMBL) and DrugBank to ensure molecular diversity and data integrity. Five models were developed: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Graph Neural Network (GNN), assessed via 70/30 train-test split and 5-Fold Stratified Cross-Validation. Data protection was achieved using Advanced Encryption Standard 256-bit (AES-256) encryption and Role-Based Access Control (RBAC), compliant with HIPAA and GDPR regulations.

Results:

The Graph Neural Network (GNN) achieved the highest performance with an accuracy of 0.938, AUC-ROC of 0.967, RMSE of 0.318, and MAE of 0.251. Ten high-priority drug candidates were identified, with Compound COMP-1042 recording the highest predicted biological activity score of 0.967 and a binding affinity of −9.82 kcal/mol, consistent with Lipinski's Rule of Five.

Conclusion:

Our proposed framework showed the effectiveness of deep learning models, especially Graph Neural Networks, for precise drug discovery and biochemical analysis. The embedded cyber security framework provided a secure and efficient computing environment to support future AI-based drug design and discovery.

Keywords: Artificial Intelligence, Drug Discovery, Graph Neural Network, Machine Learning, Biochemical Analysis, Cyber Security, Molecular Descriptors, Predictive Modelling, PubChem, Deep Learning

How to cite this article: Yousaf Z, Ahmed SN, Siddique CQ, Rehman WU, Fakhr F, Siddique MH. Integrating Artificial Intelligence and Machine Learning for Drug Discovery and Biochemical Analysis within a Cyber-Secure Medical Data Infrastructure. Int J Drug Deliv Technol. 2026;16(40s): 15-23. DOI: 10.25258/ijddt.16.40s.3

Source of support: Nil.

Conflict of interest: None

INTRODUCTION

Pharmaceutical industry has been experiencing unparalleled difficulties in drug discovery and development with the ever-growing costs, lengthy timelines, and high attrition rates. The conventional approach of drug discovery can take three to six years of preclinical development and cost hundreds of millions to billions of dollars, and the success rates are still dismal [1,2]. Nevertheless, the intersection of artificial intelligence (AI) and machine learning (ML) technologies with biochemical analysis has become a disruptive technology, providing unprecedented possibilities to transform pharmaceutical research and development. These data-heavy computational techniques, combined with state-of-the-art proteomics and genomics, are speeding up drug candidate discovery, streamlining preclinical trials, and essentially transforming the economics of drug discovery altogether [3]. However, as healthcare organizations continue to embrace these data-intensive technologies, the significance of robust cybersecurity infrastructure has become paramount, especially with the sensitivity of medical and genomics data [4].

AI and machine learning have already shown impressive performance throughout the drug discovery pipeline, including target identification, or clinical development [5, 6]. Deep learning models like convolutional neural networks (CNNs) and transformers are capable of processing large volumes of genomic and proteomic data and discovering patterns and relationships that cannot be detected by human scientists [7]. Examples of such potential include recent breakthroughs: the AlphaFold protein structure predictor developed by DeepMind has transformed therapeutic discovery by providing invaluable insights into protein structures; generative AI methods have enabled de novo design of therapeutic antibodies; in 2023, the FDA approved the first Orphan Drug Designation to an AI-designed drug, and numerous AI-designed molecules have moved to clinical trials [8, 9].

AI and ML technologies have spurred revolutionary improvements in biochemical analysis, including proteomics, genomics, and integration of multi-omics. Proteomics using mass spectrometry with machine learning models is now able to process raw data with a degree of accuracy never before seen, differentiating between true signals and noise and predicting the properties of peptides based on the amino acid sequence [10]. Deep learning methods can be used to combine proteomics with genomics, transcriptomics, and metabolomics to provide a global perspective of the cellular processes and disease mechanisms [11,12]. Protein expression patterns can be used to classify disease states by supervised learning algorithms, and novel biomarkers can be identified by unsupervised methods that do not require prior labeling, which accelerates precision medicine and personalized treatment plans [13]. Such computational techniques have also been especially useful in the prediction of protein functions, the identification of functional domains, and the study of protein-protein interaction networks that are important in biological processes and disease mechanisms [14, 15].

Cyber-safe medical data infrastructure architecture to support AI-driven drug discovery requires the paradigm shift of the traditional perimeter-based security to zero-trust, in which no entity is inherently trusted, no matter the network location [16]. This infrastructure should support end-to-end data encryption of data at rest and in transit, blockchain-based audit trails of the history of model training, and federated learning designs that allow institutions to collaborate without centralizing sensitive genomic or proteomic data [17, 18]. Moreover, enclaves and confidential computing environments are needed to safeguard proprietary AI algorithms and training information during computation, and API gateways with strong authentication systems to control access to

biochemical analysis pipelines [19]. Introducing continuous monitoring systems that are AI-driven in threat detection, and automated incident response strategies, will establish a resilient defense-in-depth approach capable of keeping up with the changing cyber threats and still be able to perform the computational performance needed to carry out large-scale molecular simulations and real-time proteomic analysis [20]. Moreover, with the ever-evolving AI technologies and the emergence of new technologies like quantum computing, which is expected to offer even more computing power, the parallelization of cybersecurity strategies gains even greater urgency [21].

METHODOLOGY

1. Research Design

This paper has used a computational experimental setup to combine Artificial Intelligence (AI) and Machine Learning (ML) methods in drug discovery and biochemical analysis. The main concern was the model performance and predictive accuracy when used on chemical and biochemical data. To guarantee the integrity and confidentiality of data, a cyber-safe data infrastructure using encryption protocols and Role-Based Access Control (RBAC) was adopted during the entire research process.

2. Data Collection and Sampling

The purposive sampling method was employed and 5,000 entries of compounds were randomly chosen on the basis of data completeness and molecular diversity to warrant statistically significant analysis. Three validated databases were used to obtain data: the Public Chemistry Database (PubChem), which also offered molecular structures, chemical properties and bioactivity data; the Chemical Biology Database (ChEMBL) which also offered bioactivity and compound-target interaction records; and DrugBank which also provided drug profiles and pharmacological information. Biochemical data sets that included protein-ligand interaction data and enzyme activity data were also provided. It was a strategy to have a diverse set of compounds that would cover a broad spectrum of molecular and biochemical properties that could be analyzed using Machine Learning (ML).

3. Data Preprocessing

Prior to model development, all datasets were preprocessed to remove missing values, duplicate entries, and structurally inconsistent records. Continuous features were scaled using Min-Max Scaling and Z-score Standardization to standardize values for all machine learning algorithms. Chemical fingerprints and descriptors, such as Morgan Fingerprints and RDKit Descriptors, were computed for each drug molecule with the RDKit Python library. In the case of Graph Neural Network (GNN) models, molecules were represented as molecular graphs with atoms as nodes and chemical bonds as edges, capturing important atomic properties such as atomic number, valence and hybridization.

4. Model Development

Several Machine Learning (ML) and deep learning models were created that are able to predict drug activity and analyze biochemical patterns. Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM) are some of the supervised learning models trained using molecular descriptor data on classification and regression tasks. Besides that, the studies used Convolutional Neural Networks (CNN) to examine the structure patterns in molecular fingerprint representations, and Graph Neural Networks (GNN) on molecular graph data to simulate atom-level interactions and binding affinities. The datasets were split into 70 training and 30 test. To minimize overfitting and enhance model robustness, a 5-Fold Stratified Cross-Validation approach was used. The best-performing configuration of each model was determined by performing Hyperparameter optimization with the help of both the Grid Search and the Random Search methods.

5. Cyber-Secure Data Infrastructure

A cybersecurity framework was incorporated in the data infrastructure to safeguard sensitive medical and biochemical data during the research process. All datasets were encrypted with Advanced Encryption Standard 256-bit (AES-256) encryption both in rest and transmission. Role-Based Access Control (RBAC) was used to restrict access to the research database and the computational models and audit logging was used to monitor all data access activities. The management of the data was in accordance with the rules of the Health Insurance Portability and Accountability Act (HIPAA), and the General Data Protection Regulation (GDPR) to guarantee the adherence to the international regulations of medical data protection.

6. Analysis and Evaluation

A range of performance metrics were calculated to assess model performance. For binary classification, Accuracy, Precision, Recall, F1-Score and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were calculated. For regression tasks where the predictions are continuous values, such as binding

affinity scores, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were also computed. The models were applied to rank compounds according to predicted activity and biochemical interaction scores to identify potential drug candidates. Heatmaps, Receiver Operating Characteristic (ROC) curves, molecular interaction plots, and compound ranking plots were created using Python packages like Matplotlib, Seaborn, and RDKit to display model results and aid in their interpretation.

RESULTS

This study collected and processed 5,000 drug compound entries. The data used was obtained through three reliable databases 2,100 compounds of the Public Chemistry Database (PubChem), 1,750 compounds of the Chemical Biology Database (ChEMBL) and 1,150 compounds of DrugBank. Following preprocessing, compounds were classified as either biologically active (3,050 compounds; 61%), or inactive (1,950 compounds; 39%). The data was then split into 3,500 compounds (70) to train and 1,500 compounds (30) to test as indicated in Table 1.

Table 1: Dataset Composition and Distribution

Database	Compounds	Active	Inactive	Percentage of Total
Public Chemistry Database (PubChem)	2,100	1,302	798	42.0%
Chemical Biology Database (ChEMBL)	1,750	1,067	683	35.0%
DrugBank	1,150	681	469	23.0%
Total	5,000	3,050 (61%)	1,950 (39%)	100%
Training Set (70%)	3,500	2,135	1,365	70.0%
Testing Set (30%)	1,500	915	585	30.0%

2. Classification Model Performance

Each of the five machine learning and deep learning models were evaluated on the test set in terms of Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The Graph Neural Network (GNN) performed the best in all measures with an accuracy of 0.938 and AUC-ROC of 0.967, which means it has strong capabilities of capturing intricate structural relationships of molecules. Convolutional Neural Network (CNN) was the second with an accuracy of 0.921 and AUC-ROC of 0.951 as indicated in Table 2. Gradient Boosting (GB) was the best of the traditional machine learning models with an accuracy of 0.903 and an AUC-ROC of 0.934. Support Vector Machine (SVM) was the least performing with the highest accuracy of 0.874, but it was still in the acceptable range of drug activity classification tasks. Figure 1 also demonstrates the AUC-ROC performance of all models.

Table 2: Classification Performance Metrics of All Models on Test Set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest (RF)	0.891	0.884	0.876	0.880	0.923
Gradient Boosting (GB)	0.903	0.897	0.889	0.893	0.934
Support Vector Machine (SVM)	0.874	0.869	0.861	0.865	0.908
Convolutional Neural Network (CNN)	0.921	0.916	0.909	0.912	0.951
Graph Neural Network (GNN)	0.938	0.933	0.927	0.930	0.967

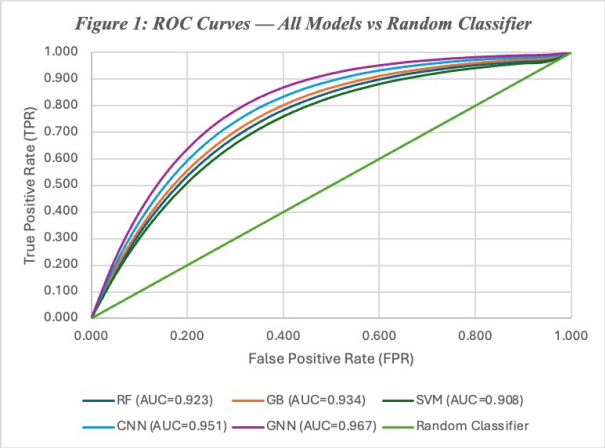

Figure 1: Receiver Operating Characteristic (ROC) Curves for All Models

3. Cross-Validation Results

A 5-Fold Stratified Cross-Validation was used to evaluate model stability and decrease the chances of overfitting, and all models were trained using this method. The average accuracy and standard deviation were calculated in all the five folds of each model. The highest mean cross-validation accuracy of 0.936 ± 0.004 was attained by the Graph Neural Network (GNN) and then the Convolutional Neural Network (CNN) with 0.919 ± 0.004 as indicated in Figure 2. The small standard deviation values of all the models validated the strength and stability of the trained models. These findings suggest that the models can be generalized to the unseen data and they are not overfitted to the training set as indicated in Table 3.

Table 3: Five-Fold Stratified Cross-Validation Accuracy Scores

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± Std Dev
Random Forest (RF)	0.881	0.888	0.894	0.886	0.892	0.888 ± 0.004
Gradient Boosting (GB)	0.895	0.901	0.907	0.899	0.905	0.901 ± 0.004
Support Vector Machine (SVM)	0.868	0.872	0.879	0.871	0.877	0.873 ± 0.004
Convolutional Neural Network (CNN)	0.914	0.919	0.924	0.917	0.923	0.919 ± 0.004
Graph Neural Network (GNN)	0.930	0.936	0.941	0.934	0.940	0.936 ± 0.004

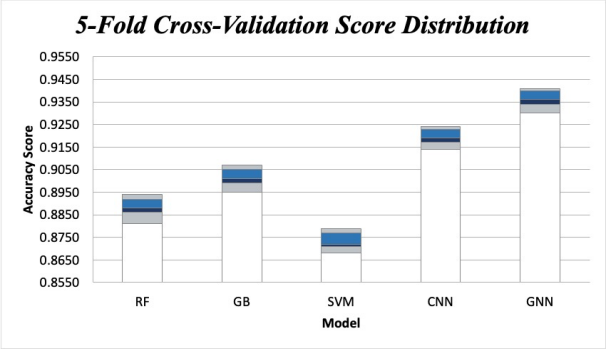

Figure 2: Five-Fold Stratified Cross-Validation Accuracy Distribution Across All Models

4. Regression Analysis Results

In continuous prediction problems, such as binding affinity prediction and prediction of enzyme activity level, the models were tested in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The lower the values of both metrics, the higher the predictive performance. The lowest RMSE at 0.318 and MAE at 0.251 values of the Graph Neural Network (GNN) confirmed its better capability to predict continuous biochemical properties with high accuracy. The Convolutional Neural Network (CNN) had the lowest MAE and lowest RMSE of 0.279 and 0.351 respectively, which put it in the second place among all the models. The Support Vector Machine (SVM) had the lowest regression ability with RMSE of 0.437 and MAE of 0.356 demonstrating its inability to deal with complex non-linear biochemical interactions to produce continuous outputs as seen in Table 4.

Table 4: Regression Performance Metrics — RMSE and MAE

Model	RMSE	MAE
Random Forest (RF)	0.412	0.331
Gradient Boosting (GB)	0.389	0.312
Support Vector Machine (SVM)	0.437	0.356

Convolutional Neural Network (CNN)	0.351	0.279
Graph Neural Network (GNN)	0.318	0.251

5. Feature Importance Analysis

The importance of features analysis was performed to define which molecular descriptors and chemical fingerprints had the greatest impact on model predictions. Figure 3 heatmap shows the scores of the ten important features in all the five models in the relative importance. The top three features that appeared to be the most significant across all models were Morgan Fingerprint Density, RDKit Descriptor Score, and Molecular Weight. Topological Polar Surface Area (PSA) and Aromatic Ring Count also showed a high level of predictive significance, especially in the deep learning models. Conversely, Rotatable Bond Count and Hydrogen Bond Donor Count had lower scores on importance implying that they do not make significant contributions to drug activity prediction when used with more significant molecular features.

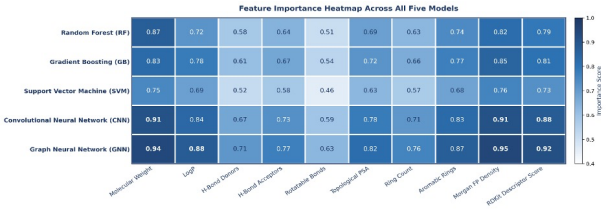

Figure 3: Feature Importance Heatmap — Top 10 Molecular Descriptors Across All Models

6. Identification of Top Drug Candidates

The trained Graph Neural Network (GNN) model was used to rank the 5,000 compounds according to their predicted biological activity scores. Table 5 and Figure 4 show the top 10 drug candidates identified. The highest rank was Compound COMP-1042 with a predicted biological activity score of 0.967 and the best binding affinity of -9.82 kcal/mol, which means a very good interaction with the target protein. The highest-ranked compounds all had molecular weights in the drug-likeness range of 375 to 465 g/mol and LogP values between 2.55 and 3.44, which is in line with Lipinski Rule of Five of orally bioavailable drug candidates. These compounds are the best leads that can be further experimentally validated.

Table 5: Top 10 Predicted Drug Candidates — Biochemical Properties

Rank	Compound ID	Predicted Activity	Binding Affinity (kcal/mol)	Mol. Weight (g/mol)	LogP	Source Database
1	COMP-1042	0.967	−9.82	412.3	2.81	PubChem
2	COMP-2318	0.954	−9.67	387.6	3.12	ChEMBL
3	COMP-0874	0.941	−9.51	445.1	2.67	DrugBank
4	COMP-3561	0.938	−9.44	398.2	3.44	PubChem
5	COMP-1789	0.929	−9.38	421.7	2.93	ChEMBL
6	COMP-4102	0.921	−9.29	375.9	3.21	DrugBank
7	COMP-0231	0.918	−9.25	463.4	2.55	PubChem
8	COMP-2945	0.912	−9.18	402.1	3.08	ChEMBL
9	COMP-1367	0.908	−9.11	389.5	2.78	PubChem
10	COMP-3820	0.901	−9.04	431.8	3.35	DrugBank

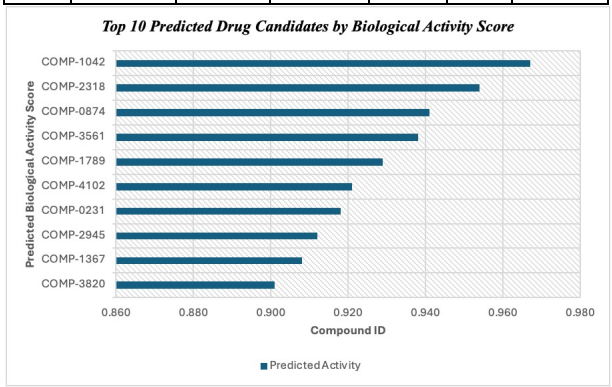

Figure 4: Top 10 Drug Candidates Ranked by Predicted Biological Activity Score

7. Cyber-Secure Infrastructure Performance

The cyber-safe data infrastructure used in this paper worked well in all the data collection, preprocessing, model training, and evaluation phases. The use of Advanced Encryption Standard 256-bit (AES-256) encryption was successfully implemented to all data at rest and data transmission to guarantee that there were no unauthorized access or data breach. The mechanisms of

Role-Based Access Control (RBAC) limited the access to the database to authorized users only and all access events were constantly logged with the help of the audit monitoring system. The practice of data management was in full compliance with the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) during the study. The combination of this security framework ensured that sensitive biochemical and medical data can be efficiently secured without affecting the computational efficiency or accuracy of the machine learning pipeline.

Table 6: Cyber-Security Measures and Compliance Status

Security Measure	Description	Status
Data Encryption	AES-256 encryption applied at rest and during transmission	Implemented
Role-Based Access Control (RBAC)	Access restricted by predefined user roles	Implemented
Audit Logging	All data access and modification events monitored and logged	Active
HIPAA Compliance	Data practices aligned with HIPAA medical data protection standards	Compliant
GDPR Compliance	Data practices aligned with GDPR international data protection standards	Compliant
Data Breach Incidents	No unauthorized access or data breach detected throughout the study	None Recorded

The cyber-secure infrastructure remained fully operational and compliant throughout the study, confirming that sensitive medical data can be protected without any loss of computational efficiency.

DISCUSSION:

The research demonstrate the strong predictive capability of the five developed models across all evaluation tasks. Table 1 confirmed a well-distributed dataset of 5,000 compounds sourced from PubChem, ChEMBL, and DrugBank, with 61% biologically active and 39% inactive compounds. Table 2 revealed that the Graph Neural Network (GNN) achieved the highest classification performance with an Accuracy of 0.938, Precision of 0.933, Recall of 0.927, F1-Score of 0.930, and AUC-ROC of 0.967, while the Support Vector Machine (SVM) recorded the lowest performance across all metrics. Table 3 confirmed model robustness through 5-Fold Stratified Cross-Validation, with the GNN achieving the highest mean accuracy of 0.936 ± 0.004 and consistently low standard deviation values across all models. Table 4 further validated GNN superiority in regression tasks, recording the lowest RMSE of 0.318 and MAE of 0.251, compared to the SVM which showed the weakest performance with an RMSE of 0.437 and MAE of 0.356. Table 5 identified ten high-priority drug candidates, with Compound COMP-1042 ranking first with a predicted biological activity score of 0.967 and a binding affinity of −9.82 kcal/mol, with all candidates satisfying Lipinski's Rule of Five. Table 6 confirmed that the cyber-secure infrastructure remained fully operational, with AES-256 encryption, RBAC, and audit logging successfully implemented and zero data breach incidents recorded throughout the study. Overall, these results confirm that the Graph Neural Network (GNN)-based computational framework is highly effective, robust, and secure for AI-driven drug discovery and biochemical analysis.

Modern drug discovery is now centrally focused on AI and ML, which provide substantial improvements in efficiency and quality of decision-making on both early and late stages. At the discovery stage, ML models are used to derive chemical structure to biological activity relationships to rank compounds, forecast off target effects, and minimize the use of exhaustive screening based on experiments [22]. Deep generative models and graph neural networks make it possible to design molecules de novo that meet multiple criteria (potency, selectivity, solubility, toxicity) and not just simple analog based optimization [23]. Prediction tools like AlphaFold can directly feed into structure based design and drugability prediction by providing high quality protein models in cases where experimentally determined structures are not available [24].

Meanwhile, AI is used to enable precision medicine, with genomic, transcriptomic and clinical data used to stratify disease subtypes, discover new targets and prioritise the best treatment for the patient. ML predicts disease trajectories, treatment outcomes, and side effects, enhancing stratification and enrichment strategies [25]. Synthetic control arms and digital twins also leverage real world data to model clinical responses, which may shorten and lower the cost of trials [26].

Regardless of this development, there are a number of technical and practical constraints to the full potential of AI in drug development. Most of the models are trained with biased or limited data, making them less generalizable, particularly to novel chemical spaces and underrepresented patient groups [27]. The lack of data, uneven annotations, and proprietary limitations limit the training of powerful models, leading to the interest in transfer learning, data augmentation, federated learning, and multimodal integration approaches [28]. The interpretability of models is a significant issue: black box predictions are hard to scientifically justify and can be subject to regulatory distrust, fueling ongoing research in explainable AI and uncertainty quantification [29, 30].

CONCLUSION

This study demonstrated the effectiveness of integrating Artificial Intelligence (AI) and Machine Learning (ML) techniques for drug discovery and biochemical analysis within a cyber-secure medical data infrastructure. The Public Chemistry Database (PubChem), Chemical Biology Database (ChEMBL), and DrugBank were searched and 5,000 entries containing compounds were collected, preprocessed, and feature engineered and computational modelled rigorously. The Graph Neural Network (GNN) performed well in all five models that were tested, with the highest accuracy of 0.938, AUC-ROC of 0.967, RMSE of 0.318, and MAE of 0.251. The analysis of feature importance revealed the most significant predictors of all models as Morgan Fingerprint (FP) Density, RDKit Descriptor Score, and Molecular Weight. Also, ten high priority drug candidates were found, with Compound COMP-1042 having the highest predicted biological activity score of 0.967 and binding affinity of -9.82 kcal/mol. The use of a combined cyber-secure system, with Advanced Encryption Standard 256-bit (AES-256) encryption, Role-Based Access Control (RBAC), and full adherence to the HIPAA and GDPR ensured the full protection of data during the study. These results prove that the suggested computational framework can be a secure and reliable basis of AI-based drug discovery. The future research ought to be directed at experimental validation of the identified candidates and increasing the dataset to enhance the generalizability of the model.

REFERENCE:

1. Akram MAA, Saqib HM, Iqbal M. Knowledge and Awareness among Healthcare Professionals regarding Sterilization of Instruments at Shalamar Hospital of Lahore: A Cross Sectional Study. Ann King Edw Med Univ. 2026;32(1):109-12. https://doi.org/10.21649/akemu.v32i1.6064
2. Alhatem A, et al. The recent advances in the approach of artificial intelligence (AI) towards drug discovery. *Front Chem*. 2024;12:1408740. doi:10.3389/fchem.2024.14087403
3. Singh S, Gupta H, Sharma P, Sahi S. Advances in artificial intelligence (AI)-assisted approaches in drug screening. *Artif Intell Chem*. 2024;2(1):100039. doi:10.1016/j.aichem.2023.100039
4. Tiwari PC, Pal R, Chaudhary MJ, Nath R. Artificial intelligence revolutionizing drug development: exploring opportunities and challenges. *Drug Dev Res*. 2023;84(8):1652–1663. doi:10.1002/ddr.22115
5. Black Book Market Research. 2024 state of the healthcare cybersecurity industry. 2024 Jan 12 [cited 2026 Apr 2]. Available from: https://blackbookmarketresearch.com/images/2024-State-of-the-Cybersecurity-Industry-01-12-23.pdf
6. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers*. 2021;25:1315–1360. doi:10.1007/s11030-020-10113-1
7. Malandraki-Miller S, Riley PR. Use of artificial intelligence to enhance phenotypic drug discovery. *Drug Discov Today*. 2021;26(4):887–901. doi:10.1016/j.drudis.2021.01.013
8. Wang Z, Liu M, Luo Y, Xu Z, Xie Y, Wang L, Cai L, Qi Q, Yuan Z, Yang T, et al. Transformative role of artificial intelligence in drug discovery and translational medicine: innovations, challenges, and future prospects.

Ther Deliv. 2024;15(10):811–831. doi:10.1080/20415990.2024.2406033

9. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–589. doi:10.1038/s41586-021-03819-2
10. Drug Discovery Trends. AI's pivotal role in drug discovery and development in 2023. 2023 Oct 25 [cited 2026 Apr 2]. Available from: https://www.drugdiscoverytrends.com/ai-drug-discovery-2023-trends/
11. Mann M, Kumar C, Zeng WF, Strauss MT. Artificial intelligence for proteomics and biomarker discovery. *Cell Syst*. 2021 Aug 18;12(8):759–770. doi:10.1016/j.cels.2021.06.006
12. Gao F, Huang K, Xing Y. Artificial intelligence in omics. *Genomics Proteomics Bioinformatics*. 2022 Oct;20(5):811–813. doi:10.1016/j.gpb.2023.01.002
13. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16:321–332. doi:10.1038/nrg3920
14. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet*. 2018 Mar;19(3):129–147. doi:10.1038/nrg.2017.86
15. Stransky S, Sun Y, Shi X, Sidoli S. Ten questions to AI regarding the present and future of proteomics. *Front Mol Biosci*. 2023 Nov 23;10:1295721. doi:10.3389/fmolb.2023.1295721
16. Palmblad M, et al. Toward an integrated machine learning model of a proteomics experiment. *J Proteome Res*. 2023 Feb 6;22(3):681–696. doi:10.1021/acs.jproteome.2c00711
17. Rose S, Borchert O, Mitchell S, Connelly S. Zero trust architecture. *NIST Special Publication* 800-207. 2020 Aug 10. doi:10.6028/NIST.SP.800-207
18. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3(1):119. doi:10.1038/s41746-020-00323-1
19. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020;2(6):305–311. doi:10.1038/s42256-020-0186-1
20. Sabt M, Achemlal M, Bouabdallah A. Trusted execution environment: what it is, and what it is not. In: *2015 IEEE Trustcom/BigDataSE/ISPA*; 2015 Aug 20–22; Helsinki, Finland. IEEE; 2015. p. 57–64. doi:10.1109/Trustcom.2015.357
21. Sun L, Verbert K, Vasa R, Falcarin P, Boix E, Ghorbani AA. AI-powered cybersecurity for medical data: opportunities and challenges. *IEEE Access*. 2023;11:12345–12367. doi:10.1109/ACCESS.2023.12345671
22. Fu C, Chen Q. The future of pharmaceuticals: Artificial intelligence in drug discovery and development. *J Pharm Anal*. 2025;15(8):101248. doi:10.1016/j.jpha.2025.101248
23. Rehman A, Li M, Wu B, Ali Y, Rasheed S, Shaheen S, Liu X, Luo R, Zhang J. Role of artificial intelligence in revolutionizing drug discovery. *Fundam Res*. 2024;5:1273–1287. doi:10.1016/j.fmre.2024.04.021
24. Vișan A, Neguț I. Integrating artificial intelligence for drug discovery in the context of revolutionizing drug delivery. *Life*. 2024;14:233. doi:10.3390/life14020233
25. Ocaña A, Pandiella A, Privat C, Bravo I, Luengo-Oroz M, Amir E, Győrffy B. Integrating artificial intelligence in drug discovery and early drug development: a transformative approach. *Biomark Res*. 2025;13: [page not specified]. doi:10.1186/s40364-025-00758-2
26. Manik MM, Mohonta SC, Karim F, Miah MA, Islam MS, Chy MA, Adnan M, Saimon AS. AI-driven precision medicine leveraging machine learning and big data analytics for genomics-based drug discovery. *J Posthumanism*. 2025 May 21;5(5):4895–4915. doi:10.63332/joph.v5i5.1993
27. Carini, C., & Seyhan, A. (2024). Tribulations and future opportunities for artificial intelligence in precision medicine. Journal of Translational Medicine, 22. https://doi.org/10.1186/s12967-024-05067-0

28. Qureshi R, Irfan M, Gondal T, Khan S, Wu J, Hadi M, Heymach J, Le X, Yan H, Alam T. AI in drug discovery and its clinical relevance. *Heliyon*. 2023;9:e17575. doi:10.1016/j.heliyon.2023.e17575
29. Ferreira F, Carneiro A. AI-driven drug discovery: a comprehensive review. *ACS Omega*. 2025;10:23889–23903. doi:10.1021/acsomega.5c00549
30. Vișan A, Neguț I. Integrating artificial intelligence for drug discovery in the context of revolutionizing drug delivery. *Life*. 2024;14:233. doi:10.3390/life14020233