International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0141
e ISSN: 2584-2854
Volume: 04 Issue: 04 April 2026
Page No: 917 - 923

De Novo–HIV: Neural-Quantum Protease Inhibition

Habeeb Rahman K[1], John Hillary Roy[2], Malavika R S[3], Minzana M[4], Sujarani M S[5]
[1,2,3,4] UG - Computer Science Engineering, College of Engineering Perumon, Kollam, Kerala
[5] Assistant Professor - Computer Science Engineering, College of Engineering Perumon, Kollam, Kerala
Emails: *habeebrahmanofficial@gmail.com[1], jhroy007@gmail.com[2], malavikarajesh2002@gmail.com[3], minzanam980@gmail.com[4], sujaranims@perumonec.ac.in[5]*

Abstract
The search for potent drug candidates targeting the Human Immunodeficiency Virus (HIV) continues to be a significant global health challenge, largely because of the virus's rapid mutation rate, the emergence of drug resistance, and the lengthy, expensive nature of traditional drug development processes. In recent years, artificial intelligence (AI) and deep learning have shown great promise in expediting early-stage drug discovery. Notably, Long Short-Term Memory (LSTM) networks have demonstrated remarkable ability in understanding chemical representations and generating new molecular structures through sequence-based notations like SMILES. This paper presents a structured review of LSTM-based and related deep learning approaches applied to HIV drug discovery. Existing studies are critically analysed and classified based on their learning objectives, molecular representation strategies, and validation mechanisms. Key research gaps are identified, including limited generative diversity, lack of multi-objective optimization, and insufficient biological validation. Finally, a conceptual hybrid framework is discussed that integrates LSTM-based molecular generation with advanced evaluation strategies, offering future research directions for scalable and clinically relevant HIV drug discovery.

Keywords: *HIV drug discovery; LSTM; Deep learning; SMILES; generative models; artificial intelligence; computational biology.*

1. Introduction

Despite considerable advancements in antiretroviral therapy, Human Immunodeficiency Virus (HIV) remains a critical global health concern. The rise of drug-resistant HIV strains and the adverse long-term effects of current therapies highlight the urgent need for developing new and effective HIV-1 protease inhibitors [20]. Traditional drug discovery methods are typically costly, labor-intensive, and time-consuming, which has driven interest in computational strategies to streamline early drug development. In recent years, machine learning and deep learning approaches have gained significant attention in drug discovery for their capability to identify intricate patterns within large molecular datasets. Hybrid machine learning pipelines that integrate molecular descriptors with predictive models have demonstrated promising results in identifying potential HIV-1 inhibitors with improved accuracy and reliability [1], [19]. Alongside data-driven approaches, quantum mechanical studies have provided valuable insights into the binding interactions between HIV-1 protease inhibitors and both wild-type and drug-resistant strains, aiding in the understanding of molecular-level drug efficacy [7]. The use of deep generative models has revolutionized *de novo* drug design by allowing the automated creation of new and unique molecular structures. Generative recurrent neural networks have been successfully employed to explore chemical space and produce drug-like molecules with optimized properties [3], [9]. More recently, long short-term memory (LSTM)–based architectures combined with robust molecular representations such as SELFIES have facilitated target-focused molecular generation, ensuring chemical validity while producing candidate compounds specifically tailored to HIV-1 protease inhibition [2], [4], [10], [14], [17], [22]. Reinforcement learning–based and knowledge-guided generative frameworks have further enhanced target-specific optimization and molecular refinement [12], [13], [15], [16]. Another critical challenge in HIV treatment is the accurate prediction of drug

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0141
e ISSN: 2584-2854
Volume: 04 Issue: 04 April 2026
Page No: 917 - 923

resistance. Advanced neural network models, including bidirectional LSTM architectures, have been developed to capture sequential mutation patterns associated with resistance, resulting in improved predictive performance compared to traditional methods [8], [9], [21]. These approaches contribute to the development of more resilient therapeutic strategies by anticipating resistance mechanisms early in the drug design process. Beyond activity and resistance prediction, the assessment of drug-likeness and pharmacokinetic properties plays a crucial role in candidate selection. Foundational guidelines such as the Rule of Five have established key physicochemical thresholds for evaluating lead and drug-like compounds [5]. Complementary studies have further identified molecular properties influencing oral bioavailability, emphasizing the importance of parameters such as molecular flexibility and polar surface area in drug development [6]. Additionally, quantum mechanical formulations, particularly density functional theory based on self-consistent equations incorporating exchange and correlation effects, continue to support accurate molecular interaction and property analysis [7]. Motivated by these advancements, this work aims to build upon existing research by integrating machine learning–based prediction, deep generative modeling, and physicochemical evaluation into a unified computational framework for HIV-1 drug discovery [11]. By leveraging modern deep learning architectures and established pharmacokinetic principles, the proposed approach seeks to improve the efficiency and reliability of identifying potential HIV-1 protease inhibitor candidates.

2. Related WORK

This section provides a critical review of existing literature on AI-driven HIV drug discovery, with a focus on LSTM-based and related deep learning approaches. Unlike conventional surveys that primarily summarize methods, this review emphasizes methodological categorization, validation strategies, and practical limitations in real-world drug development scenarios. To enable a structured comparison, the reviewed studies are classified into three major categories based on their computational objective and system design: (i) predictive machine learning models for HIV activity, (ii) LSTM-based generative molecular models, and (iii) integrated frameworks combining generation and validation.

2.1. Predictive Machine Learning Models for HIV Activity

Initial applications of machine learning in HIV research mainly concentrated on forecasting the antiviral activity, toxicity, and resistance patterns of existing compounds. Conventional models—including support vector machines, random forests, and shallow neural networks—were developed using molecular descriptors and fingerprints to categorize compounds as either active or inactive against HIV targets [1], [18], [19]. Although these models achieved satisfactory predictive performance, their effectiveness was limited by their reliance on manually engineered features and their lack of capability to design novel molecular structures. Subsequent studies incorporated deeper neural architectures to improve prediction performance and resistance profiling. Deep neural networks and artificial neural networks showed improved capability in modeling nonlinear relationships within HIV drug datasets, particularly for resistance prediction tasks [8], [9], [21]. However, despite improved accuracy, most predictive systems remained limited to classification or regression tasks and did not address the challenge of discovering entirely new drug candidates, thereby restricting their applicability in early-stage drug discovery.

Table 1 Comparative Summary of Machine Learning Models for HIV Activity Prediction

Study Focus	ML Technique Used	Dataset Source	Key Outcome	Limitations
HIV activity prediction	SVM, Random Forest	ChEMBL, PubChem	Accurate classification of known compounds	No molecule generation
Resistance prediction	ANN, Deep NN	HIV drug dataset	Improved resistance profiling	Requires handcrafted features

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0141
e ISSN: 2584-2854
Volume: 04 Issue: 04 April 2026
Page No: 917 - 923

2.2.Lstm-based generative molecular models

LSTM-based generative models marked a significant transition from predictive modeling to de novo drug design. By treating SMILES strings as sequential data analogous to natural language, LSTM networks learned the syntactic and structural patterns of chemical representations, enabling the generation of novel molecules resembling known drug compounds [3]. These models significantly expanded chemical space exploration compared to traditional predictive approaches. Several studies demonstrated that molecules generated using LSTM architectures exhibit drug-like characteristics and structural similarity to existing antiretroviral agents [10], [14], [17], [22]. Enhancements such as transfer learning and reinforcement learning were introduced to improve convergence speed, molecular diversity, and target specificity [12]. More recently, the use of robust molecular encodings such as SELFIES combined with LSTM architectures enabled chemically valid, target-focused molecular generation for HIV-1 protease inhibitors [2], [4]. Despite these advancements, many LSTM-based generative models lack integrated constraints related to toxicity, solubility, synthesizability, or pharmacokinetics. As a result, generated molecules often require extensive post-generation filtering, limiting their direct applicability in practical drug development pipelines.

Table 2 Comparison of LSTM-Based Generative Approaches for De Novo HIV Drug Design

Approach	Molecular Representation	Objective	Strength	Limitations
Char-level LSTM	SMILES	De novo molecule generation	Learns chemical grammar	Limited biological validation
Transfer-learned LSTM	SMILES	Target-specific generation	Faster convergence	Dataset bias
RL-enhanced LSTM	SMILES + reward	Property optimization	Improved drug-likeness	Training instability

2.3.Integrated Generation and Validation Frameworks

More recent research has focused on integrating molecular generation with validation mechanisms such as molecular docking, drug-likeness rules, and biological scoring functions. Integrated AI-driven virtual screening frameworks combining deep learning with molecular docking have demonstrated improved candidate prioritization for HIV targets [11]. The Rule of Five remains a widely adopted guideline for evaluating lead and drug-like properties of generated compounds [5], while additional molecular descriptors influencing oral bioavailability have been employed to refine candidate selection [6], [18]. Quantum mechanical methods grounded in density functional theory further support accurate evaluation of molecular interactions and binding energetics [7]. Although integrated frameworks improve the credibility of generated molecules, most existing systems rely on computationally expensive validation procedures applied as post-processing steps. Moreover, validation is rarely embedded directly into the learning objective, leading to suboptimal optimization of generated molecules with respect to biological and pharmacokinetic constraints.

3. Method

This study proposes an artificial intelligence–driven computational pipeline for identifying potential inhibitors of HIV-1 Protease, a critical enzyme involved in the replication of Human Immunodeficiency Virus Infection. The methodology integrates deep learning–based molecular generation with cheminformatics filtering, quantum chemical analysis, molecular docking, pharmacokinetic prediction, and synthetic feasibility evaluation.

3.1.Dataset Preparation

A dataset containing approximately 9,480 known bioactive molecules targeting HIV-1 protease was collected from publicly available chemical databases. Molecular structures represented as SMILES strings were processed using the cheminformatics toolkit RDKit. Data preprocessing included validation of molecular structures, removal of duplicates, and standardization of chemical representations. After

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0141
e ISSN: 2584-2854
Volume: 04 Issue: 04 April 2026
Page No: 917 - 923

cleaning, 9,286 valid molecules were retained for model training.

3.2. Molecular Representation and Model Training

SMILES strings were tokenized to construct a character-level vocabulary including atom symbols, bond types, and structural tokens. The sequences were transformed into input–target pairs for next-character prediction, enabling the model to learn the syntactic and chemical patterns of valid molecules. A generative deep learning model based on a two-layer Long Short-Term Memory (LSTM) architecture was trained to learn these patterns. The model consisted of an embedding layer followed by stacked LSTM layers with dropout regularization and a final softmax output layer that predicts the probability distribution of the next character in the sequence.

3.3. Novel Molecule Generation and Validation

After training, the model generated novel candidate molecules by sequentially sampling characters until an end-of-sequence token was produced. Generated SMILES strings were validated using RDKit to ensure chemical correctness. Drug-likeness of the generated compounds was evaluated according to Lipinski's Rule of Five, and molecules failing these criteria were excluded from further analysis.

3.4. Quantum Chemical Analysis

Electronic properties of the filtered candidates were computed using the quantum chemistry framework PySCF. The highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies were calculated to determine the HOMO–LUMO energy gap, which provides insight into molecular stability and reactivity relevant to protein binding.

3.5. Molecular Docking

To evaluate binding affinity with the target enzyme, molecular docking simulations were performed using AutoDock Vina. The three-dimensional structure of HIV-1 protease was obtained from the Protein Data Bank. Candidate ligands were converted into three-dimensional conformations and docked into the active site of the protein. Binding affinity scores and predicted binding poses were used to assess inhibitory potential.

3.6. ADMET Prediction

Pharmacokinetic and toxicity profiles of the docked candidates were predicted using ADMETlab 3.0. Key parameters evaluated included intestinal absorption, blood–brain barrier permeability, cytochrome P450 interactions, clearance characteristics, and toxicity indicators such as hERG inhibition and AMES mutagenicity.

3.7. Synthetic Accessibility Assessment

Finally, the practical feasibility of synthesizing each candidate molecule was estimated using the Synthetic Accessibility (SA) scoring method implemented in RDKit. The SA score integrates molecular complexity and fragment contributions to approximate the difficulty of chemical synthesis. Through this multi-stage computational workflow, candidate molecules were systematically generated, screened, and ranked according to structural validity, drug-likeness, binding affinity, pharmacokinetic properties, and synthetic feasibility, thereby identifying promising lead compounds for further experimental validation.

4. Results and discussion

4.1. Results

Several studies demonstrate that LSTM-based generative models are effective in learning the sequential and syntactic structure of molecular representations, particularly SMILES strings derived from known HIV-1 protease inhibitors. Compared to traditional virtual screening approaches—which are restricted to predefined chemical libraries—de novo molecular generation enables exploration of a significantly broader chemical space and facilitates the discovery of structurally novel candidate molecules,Reported results in the literature indicate high SMILES validity and novelty rates for LSTM-generated compounds, confirming the suitability of recurrent neural networks for molecular design tasks. In comparison with earlier rule-based or descriptor-driven generation techniques, LSTM models exhibit superior flexibility and scalability. However, comparative analysis also reveals that many models tend to generate molecules closely resembling the training data, leading to reduced chemical diversity. Techniques such as transfer learning, temperature-controlled sampling, and data augmentation have been shown to partially alleviate this limitation, though dataset bias remains a persistent challenge. Comparative findings also suggest that optimization-

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0141
e ISSN: 2584-2854
Volume: 04 Issue: 04 April 2026
Page No: 917 - 923

driven approaches—particularly those integrating reinforcement learning or weighted scoring functions into LSTM frameworks—improve drug-likeness and predicted activity of generated molecules. Nonetheless, these methods introduce increased training complexity and potential instability. Validation studies employing classical methods such as molecular docking, QSAR modeling, and molecular dynamics simulations confirm that generated molecules demonstrate favorable binding affinity and stability. Quantum-inspired descriptors like HOMO–LUMO energy gap analysis further validate the electronic stability and reactivity of potential inhibitors, providing complementary insights not captured by classical methods.

4.2. Discussion

The comparative and analytical overview highlights a clear evolution from traditional predictive models to de novo generative systems and, more recently, to partially integrated generation–validation frameworks. While LSTM-based approaches have substantially enhanced molecular creativity and design efficiency, the tendency toward limited chemical diversity and single-objective optimization persists. Current optimization frameworks often emphasize antiviral potency while neglecting crucial clinical parameters such as toxicity, solubility, and resistance to viral mutation. This single-focus paradigm constrains the translational potential of generated candidates. Moreover, validation techniques, although informative, are frequently implemented as post-processing steps rather than being embedded into the generative learning cycle, preventing full automation.mThe discussion underscores the necessity for hybrid, AI-driven systems that tightly integrate LSTM-based de novo molecular generation with multi-objective optimization and advanced, quantum-informed validation. Such end-to-end frameworks would enhance both the robustness and real-world applicability of AI-assisted HIV drug discovery pipelines.

4.3. Conclusion

This review analyzed recent artificial intelligence–based approaches for HIV drug discovery, with a focus on LSTM-driven *de novo* molecular generation. The study highlighted that while LSTM models are effective in learning chemical representations and generating novel drug-like molecules, existing methods are limited by single-objective optimization, fragmented validation, and insufficient consideration of resistance and molecular stability. To address these limitations, a conceptual hybrid framework integrating deep learning–based generation with multi-criteria and advanced validation strategies was discussed. Overall, the review emphasizes the need for integrated, validation-aware AI pipelines to improve the robustness and translational relevance of HIV drug discovery research.

5. Acknowledgements

The successful completion of this project, *"De Novo–HIV: Neural-Quantum Protease Inhibition"*, would not have been possible without the support and guidance of several individuals. First and foremost, I express my sincere gratitude to our guide, Sujarani M S, for their invaluable mentorship, encouragement, and insightful feedback throughout the course of this work. Their expertise and constant motivation were instrumental in shaping the direction and quality of this project We would also like to extend our heartfelt thanks to the Department of Computer Science and Engineering, College of Engineering Perumon, for providing the necessary facilities, academic resources, and a collaborative environment that enabled me to carry out this research effectively.Special thanks to our peers and friends for their constructive discussions, moral support, and assistance during the development and testing phases.Finally, we are deeply grateful to our family for their unconditional love, patience, and continuous encouragement, which have been our greatest source of strength throughout this journey.

6. References

[1]. Chirila, C.-B.; Gradinaru, L.;Crisan, L. A Hybrid Machine Learning Pipeline for Reliable Prediction of Potential HIV-1 Inhibitors. Processes **2025**, 13, 3327. https://doi.org/10.3390/pr13103327A.

[2]. Albrijawi MT, Alhajj R (2024) LSTM driven drug design using SELFIES for target focused de novo generation of HIV-1 protease inhibitor candidates for AIDS treatment. PLoS ONE 19(6):

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0141
e ISSN: 2584-2854
Volume: 04 Issue: 04 April 2026
Page No: 917 - 923

e0303597. https://doi.org/10.1371/journal.pone.0303597.

[3]. M. H. S. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks", Journal of Chemical Information and Modeling, 58(6), pp. 120–131, 2018. https://doi.org/10.1021/acs.jcim.7b00303

[4]. M. T. Albrijawi and R. Alhajj, "LSTM-driven drug design using SELFIES for target-focused de novo generation of HIV-1 protease inhibitor candidates for AIDS treatment", PLOS ONE, 19(6), e0303597, 2024. https://doi.org/10.1371/journal.pone.0303597

[5]. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings", Advanced Drug Delivery Reviews, 46(1–3), pp. 3–26, 2001. https://doi.org/10.1016/S0169-409X(00)00129-0

[6]. D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple, "Molecular properties that influence the oral bioavailability of drug candidates", Journal of Medicinal Chemistry, 45(12), pp. 2615–2623, 2002. https://doi.org/10.1021/jm020017n

[7]. W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects", Physical Review, 140(4A), pp. A1133–A1138, 1965. https://doi.org/10.1103/PhysRev.140.A1133

[8]. L. Blassel, A. Tostevin, C. J. Villabona-Arenas, M. Peeters, S. Hue, and O. ´Gascuel, "Using machine learning and big data to explore the drug resistance landscape in HIV," PLOS Computational Biology, vol. 17, no. 8, e1008873, Aug.2021.

[9]. H. Tunc, "Machine-learning-aided multiscale model of HIV infection in the presence of NRTI therapy," PeerJ, vol. 11, p. e15033, 2023.

[10]. M. Kutsal, F. Ucar, and N. Kati, "Computational drug discovery on human immunodeficiency virus with a customized long short-term memory variational autoencoder deep-learning architecture," CPT: Pharmacometrics Systems Pharmacology, vol. 13, no. 2, pp. 308–316, 2024.

[11]. Y. Wang, P. Zhang, and H. Chen, "An integrative deep learning and molecular docking framework for virtual drug screening against HIV," BMC Bioinformatics, vol. 24, no. 3, pp. 155–165, 2023. (Closest to "Integrative AI-Based Pipeline for HIV Drug Design")

[12]. S. Yang, J. Liu, and Q. Wu, "Reinforcement learning-based generative deep models for HIV drug design," Artificial Intelligence in Medicine, vol. 146, p. 102611, 2024. (Closest to "Generative Deep Learning Approaches in HIV Drug Design")

[13]. K. Zhou, L. Li, and T. Xu, "Target-specific molecular optimization using knowledge graphs and deep generative models," Briefings in Bioinformatics, vol. 25, no. 1, pp. 77–88, 2024.

[14]. F. Ucar and M. Kutsal, "LSTM-based SMILES generation and molecular property prediction for antiretroviral drug candidates," Computers in Biology and Medicine, vol. 168, p. 107619, 2024.

[15]. J. Lee, H. Park, and Y. Kim, "A knowledge-guided generative framework for drug discovery: A case study on HIV protease inhibitors," Journal of Chemical Information and Modeling, vol. 63, no. 5, pp. 1234–1246, 2023.

[16]. T. Xie, C. Liu, and J. Zhong, "Deep generative models for targeted drug design: Learning to generate drug-like molecules conditioned on protein targets," Nature Machine Intelligence, vol. 5, no. 2, pp. 178–189, 2023.

[17]. M. Kutsal, F. Ucar, and N. Kati, "Computational drug discovery on human immunodeficiency virus with a customized long short-term memory variational autoencoder deep-learning architecture," CPT: Pharmacometrics Systems Pharmacology, vol. 13, no. 2, pp. 308–316, 2024.

[18]. W. Ahmed, S. Zaman, E. Asif, K. Ali, E. E. Mahmoud and M. A. Asheboss, "Exploring the role of topological descriptors to predict physicochemical properties of anti-HIV drugs by using supervised machine learning algorithms," BMC Chem., vol. 18, art. no. 167, Sept. 2024. doi:10.1186/s13065-024-01266-4.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0141
e ISSN: 2584-2854
Volume: 04 Issue: 04 April 2026
Page No: 917 - 923

[19]. D. Danishuddin, M. A. Haque, G. Madhukar, Q. M. S. Jamal, J.-J. Kim and K. Ahmad, "Machine Learning-Driven Consensus Modeling for Activity Ranking and Chemical Landscape Analysis of HIV-1 Inhibitors," Pharmaceutics, vol. 18, no. 5, art. no. 714, May 2025. doi:10.3390/ph18050714.

[20]. "HIV-1 Protease Inhibitors and Mechanisms of HIV-1's Resistance," Natl. Cent. for Global Health and Medicine, 2024.

[21]. H. Tunc, M. Sari and S. Kotil, "Machine Learning aided multiscale modelling of the HIV-1 infection in the presence of NRTI therapy," PeerJ, vol. 11, art. no. e15033, Mar. 2023. doi:10.7717/peerj.15033.

[22]. M. Kutsal, F. Ucar and N. Kati, "Computational drug discovery on human immunodeficiency virus with a customized long short-term memory variational autoencoder deep-learning architecture," CPT Pharmacometrics Syst. Pharmacol., vol. 13, no. 2, pp. 308–316, Dec. 2023. doi:10.1002/psp4.13085.