AI FOR DRUG DISCOVERY LSTM-DRIVEN DRUG DESIGN USING SELFIES FOR TARGET-FOCUSED DE NOVO GENERATION OF HIV-1 PROTEASE INHIBITOR CANDIDATES IN THE TREATMENT OF AIDS

MSc Thesis

M.Taleb ALBRIJAWI

Graduate School of Engineering and Natural Sciences

Biomedical Engineering and Bioinformatics

July, 2023

AI FOR DRUG DISCOVERY

LSTM-DRIVEN DRUG DESIGN USING SELFIES FOR TARGET-FOCUSED DE NOVO GENERATION OF HIV-1 PROTEASE INHIBITOR CANDIDATES IN THE TREATMENT OF AIDS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
ENGINEERING AND NATURAL SCIENCES
OF ISTANBUL MEDIPOL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
BIOMEDICAL ENGINEERING AND BIOINFORMATICS

By

M.Taleb Albrijawi

July, 2023

LSTM-DRIVEN DRUG DESIGN USING SELFIES FOR TARGET-FOCUSED DE NOVO GENERATION OF HIV-1 PROTEASE INHIBITOR CANDIDATES IN THE TREATMENT OF AIDS

By M.Taleb ALBRIJAWI

14 July 2023

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Master of Science.

_____

Prof. Dr. Reda Alhajj (Advisor)

_____

Prof. Dr. Cihan Bilge Kayasandık

_____

Prof. Dr. Tansel Özyer

Approved by the Graduate School of Engineering and Natural Sciences:

_____

Prof. Dr. Yasemin Yüksel Durmaz

Director of the Graduate School of Engineering and Natural Sciences

I hereby declare that all information in this document has been obtained and presented in accordance with the academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Signature :

Name, Surname: M.TALEB ALBRIJAWI

ACKNOWLEDGEMENT

Completing this thesis has been made possible by the unwavering support and contributions of amazing individuals, without whom this achievement would not have been possible.

Advisor:

I would like to express my heartfelt gratitude to my advisor, Prof. Dr. Reda Alhajj, for his incredible support, belief in my abilities, and trust throughout my thesis journey. His guidance and mentorship have profoundly shaped both my academic and personal lives. He has been an exceptional advisor, displaying flexibility in his approach and respecting my individual learning style. He provided invaluable insights and guidance while allowing me the freedom to navigate my research journey. His patient demeanor and amazing support have greatly contributed to my progress.

Words fail to capture the depth of my heartfelt appreciation to Prof. Dr. Reda Alhajj, and I will carry the lessons I learned from him throughout my lifetime.

Family and friends:

I would also like to thank my family and lab-mate friends Nawazish, Malek and Ziad for their unconditional motivation, support, and understanding throughout this process. Their encouragement and belief in me have been a constant source of strength.

Lastly, I would like to express my sincere appreciation to the faculty members and research funding committees for their generous funding and support of my thesis. Their financial assistance has been instrumental in the successful completion of this work. I am grateful for the opportunity they have provided me to pursue my academic endeavors and contribute to the field. Their investment in my education and research is truly invaluable, and I am honored to have been a recipient of their support.

M.Taleb Albrijawi

July, 2023

CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

% **:** Percentage
∑ **:** Summation Symbol
σ **:** Sigma
e **:** Euler's number, it approximately equals to 2.71828.
log **:** Logarithm function

ABBREVIATIONS

AI : Artificial Intelligence
ML : Machine Learning
DL : Deep Learning
DNA : Deoxyribonucleic Acid
RNA : Ribonucleic acid
LSTM : Long Short-Term Memory Network
HIV : Human Immunodeficiency Virus
AIDS : Acquired Immune Deficiency Syndrome
AUC : Area Under Curve
TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative
ROC : Receiver Operator Characteristic
GPU : Graphical Processing Unit
2D : 2 Dimensional
3D : 3 Dimensional
PDB : Protein Data Bank
PR : Protease
VAE : Variational Autoencoder
Min : Minimum
Max : Maximum
Std : Standard Deviation
FDA : Food and Drug Administration
GAN : Generative Adversarial Network
[nop] : No Operation
MCS : Maximum Common Sub-structures

YAPAY ZEKA TABANLI LSTM DESTEKLİ İLAÇ TASARIMI: AIDS TEDAVİSİNDE SELFIES KULLANARAK HIV-1 PROTEAZ ODAKLI İNHİBİTÖR ADAYLARININ TASARLANMASI

ÖZET

M.Taleb ALBRIJAWI

Biyomedikal Mühendisliği ve Biyoenformatik, Yüksek Lisans

Tez Danışmanı: Prof. Dr. REDA ALHAJJ

Temmuz, 2023

Yeni ilaç keşif süreci hala uzun süren çalışmalar ve yüksek maliyetler gerektiren bir süreç olarak devam etmektedir. Bir ilacın piyasaya sürülebilmesi için çeşitli aşamaların başarıyla tamamlanması gerekmektedir. Bu uzun süreçte laboratuvar çalışmaları ve klinik deneylerle ilacın kalitesi, güvenliği ve etkinliği gibi kriterler değerlendirilmektedir. Ancak, koronavirüs pandemisi gibi acil ve son derece karmaşık hastalıklar durumunda, ilaç keşif süreci, ilaç tasarımcıları için büyük zorluklar ortaya çıkarmıştır. Özellikle virüs salgınları ile mücadelede, virüslerin adaptif yetenekleriyle sürekli olarak yenilikçi stratejilere ve yeni teknolojilerin keşif sürecine entegre edilmesine ihtiyaç duyulduğu görülmektedir.

Son zamanlarda, mevcut ilaçların farklı hastalıkların tedavisinde yeniden kullanılması ve hızlı testlerin geliştirilmesi gibi "ilaç yeniden konumlandırma" adı verilen yaklaşımlar büyük ilgi çekmiştir. Bu yöntemler, yapay zeka teknolojilerinin uygulanmasıyla birlikte ciddi tehlike oluşturan hastalıkların tedavisinde önemli ilerlemeler sağlanmıştır. Bu yeni yöntemler sayesinde, HIV gibi enfeksiyonlara karşı büyük başarılar elde edilmiştir. Bunlardan biri LSTM (Long Short-Term Memory) tabanlı modeldir. Bu model, metin veya müzik gibi yaratıcı içeriklerin üretiminde, özellikle doğal dil işleme (Natural Language Processing - NLP) alanında büyük bir yetenekle dikkat çekmektedir. Bu model, araştırmacıların ilaç keşif sürecinde yenilikçi moleküllerin üretiminde kullanılmasıyla büyük bir başarı elde etmiştir. Bu gelişme, gelecekteki ilerlemelerde umut verici sonuçlar sunarak yeni fırsatlar ortaya çıkarmıştır.

Bu araştırmada, HIV-1 proteazına karşı etkili olabilecek yeni moleküllerin tasarımında "LSTM-ProGen" adlı bir üretici modelin kullanımı önerilmiştir. Bu model, HIV-1 proteazıyla etkileşime girebilecek moleküllerin üretimi için LSTM tabanlı bir üretim modelidir.

ChEMBL veri tabanından elde edilen moleküller, LSTM-ProGen modelinin aşamalı bir eğitim sürecinden geçirilmiştir. Bu süreçte, modelin yeni molekül üretme yeteneklerini ve moleküler kalitesini artırmak için çeşitli yöntemler uygulanmıştır. Yeni üretilen moleküller, ilaç benzeri özellikler gibi çeşitli performans ölçütleriyle ayrıntılı olarak analiz edilmiştir. Moleküler kenetleme yöntemiyle yapılan analizler sonucunda, HIV proteazına karşı etkili olduğu tespit edilen 12 aday molekül belirlenmiştir.

LSTM-ProGen, yüksek optimizasyon teknikleriyle geliştirilmiştir ve hedef proteine özgü etkili ilaç adayları tasarlamak için büyük potansiyele sahiptir. Bu inovasyon sistemi, gelecekte ilaç tasarımı ve geliştirilmesi alanında önemli katkılar sağlayabilir.

Anahtar sözcükler: AIDS, HIV-1 Proteaz İnhibitörü, LSTM, Yapay Zeka, Derin Öğrenme, İlaç Tasarımı.

LSTM-DRIVEN DRUG DESIGN USING SELFIES FOR TARGET-FOCUSED DE NOVO GENERATION OF HIV-1 PROTEASE INHIBITOR CANDIDATES IN THE TREATMENT OF AIDS

ABSTRACT

M.Taleb Albrijawi

MSc in Electrical, Electronics Engineering and Cyber Systems

Advisor: Prof. Dr. REDA ALHAJJ

July, 2023

The undeniable urgency of the quest to discover groundbreaking antiviral drugs far removed from traditional approaches has been dramatically exemplified by the COVID-19 pandemic. Additionally, the ongoing battle against drug resistance in viruses further underscores this need to surpass conventional methods and bridge the gap for the benefit of humanity. While traditional approaches rely on manual exploration by experts, which require a significant amount of effort, time, and money, innovative strategies to combat the virus's adaptive nature and safeguard effective treatment options are required, especially in the fight against the HIV virus, which still poses a serious challenge. Drug repurposing strategies involve identifying new therapeutic uses for existing drugs that have already been approved or are in advanced stages of development. These approaches leverage the extensive knowledge and vast libraries of compounds that have undergone testing and safety assessments. Using these strategies, many of the challenges associated with traditional drug discovery can be passed and lead to the discovery of innovative treatment options for diseases with unmet medical needs, providing hope for patients who may have exhausted conventional treatment options. In parallel, through the magic of artificial intelligence, these innovative strategies hold tremendous promise for accelerating drug discovery and development. Particularly, the emergence of deep generative models has gained attention due to their high ability to generate new data that obeys the underlying distribution of the training data. These models have unveiled the potential for automating the discovery process with their extensive capability to synthesize new molecules with diverse chemical properties and features. Various approaches have been put forth to tackle this task, but controlling and guiding the generation process has proved challenging, accompanied by increasing complexity and computational demands. To truly revolutionize real-world drug development pipelines, it is essential to empower these models with the capability to design molecules tailored to specific targets, thereby driving the field towards remarkable progress.

LSTM-based generative models have inspired researchers for their proficiency in processing sequential data in natural language applications. Building upon this foundation, they have embarked on a captivating exploration, employing these models to generate innovative molecules by interpreting them as snippets of text. Encouragingly, these models have exhibited promising outcomes, fueling optimism for future advancements. An application of an LSTM-based deep generative model named "LSTM-ProGen" is proposed to be tailored explicitly for the de novo design of drug candidate

molecules that interact with a specific target protein (HIV-1 protease). LSTM-ProGen distinguishes itself by employing a long-short-term memory (LSTM) architecture, augmented with protein-protease interaction knowledge, to generate novel molecules with remarkable precision and target specificity against the HIV-1 protease inhibitor. Following an extensive training process involves fine-tuning LSTM-ProGen on a diverse range of compounds sourced from ChEMBL, a reputable database. The model was optimized to meet specific requirements, with multiple iterations to enhance its predictive capabilities and ensure it generates molecules that exhibit favorable target interactions. The training process encompasses an array of performance evaluation metrics, such as drug-likeness properties. LSTM-ProGen outperforms previous state-of-the-art generative models, demonstrating a significant leap forward in the generation of target-specific drug candidate molecules. Our evaluation includes extensive silico analysis using molecular docking and PCA-based visualization to explore the chemical space that the new molecules cover compared to those in the training set. These evaluations reveal that a subset of 12 de novo molecules generated by LSTM-ProGen exhibit a striking ability to interact with target proteins, rivaling or even surpassing the efficacy of native ligands.

LSTM-ProGen represents a transformative breakthrough in the area of drug discovery. By leveraging the power of LSTM networks, our model enables the automated design of target-specific drug candidate molecules. The implications of LSTM-ProGen are far-reaching, as it has the potential to be extended to expedite the drug development process and pave the way for the discovery of novel therapeutic options for druggable targets in various diseases. With further refinement and validation, LSTM-ProGen holds promise as a versatile tool for designing efficacious and customized drug candidates tailored to specific targets and diseases.

Keywords: HIV, AIDS, HIV-1 Protease, LSTM, Deep Learning, SELFIES, Drug Discovery, Drug Design.

CHAPTER 1

1. INTRODUCTION

Over time, humanity has been striving to eliminate all illnesses and their root causes. Plants have been the source of medicines used to treat different diseases for centuries. The knowledge of using various parts of plants, such as bark, seeds, and fruit, has been passed down throughout the ages and gradually accumulated. One example is the use of salicylic acid extracted from willow bark to treat pain and inflammation. As societies progressed, the process of finding new treatments became more systematic and regulated with the immense advancements in the pharmaceutical industry. Today, discovering new novel drugs involves a range of advanced techniques, including computer modeling, high-throughput screening, and the use of genomic data. The astonishing advances in artificial intelligence (AI) have paved the way for ground-breaking transformations across a variety of industries, including drug discovery. By leveraging the power of AI, many of the hurdles that have long stymied the field have been overcome, enabling more efficient and effective drugs for the betterment of humanity. Particularly, deep generative models have been a game-changer in the process of drug discovery; they are shaking up the industry in a huge way that is unprecedented in history. These models can unlock exciting new avenues for drug design, ushering in a new era by unleashing a tidal wave of potential that promises to revolutionize the rapid identification of novel drug candidates and the development of highly effective treatments for a range of diseases. With the ever-increasing complexity of diseases and the need for more personalized treatments to bring hope to patients around the world, there is a growing urgency to employ AI in the process of drug design.

1.1. Problem Definition and Motivation

Human immunodeficiency virus (HIV) is the insidious culprit behind acquired immunodeficiency syndrome (AIDS) [1]. It continues to pose a major global public health challenge, affecting millions of people worldwide [2]. Once this virus enters the host blood stream, it attacks the CD4 immune cells, weakening the immune system and decreasing the body's ability to fight other infectious diseases. According to the latest statistics published by "Joint United Nations Program on HIV/AIDS" [3], there were 38.4 million people living with HIV in 2021, and 1.5 million people became newly infected with the virus and tragically, 650,000 individuals lost their lives to AIDS-related illnesses. This virus presents a menace to the health and well-being, especially for those individuals living in poverty-stricken countries with limited resources, and lack of access to prevention and treatments.

HIV/AIDS spreads through diverse ways, including unprotected sexual intercourse, sharing of needles or syringes, and mother-to-child transmission during pregnancy, delivery, or breastfeeding [4]. These modes of transmission continue to provide considerable hurdles to HIV prevention efforts. Despite the progress in expanding access to antiretroviral therapy, a staggering number of people living with HIV, estimated at 16.3 million, still do not have access to life-saving treatment [3]. This not only puts their health at risk, but also contributes to the ongoing spread of the virus. The HIV-1 life cycle is complex, involving multiple steps that are essential for the virus to replicate and spread throughout the host's body. It can be divided into six distinct stages [5], beginning with the attachment to the host cell membrane and fusion with a CD4+ T-lymphocyte (**Figure 1.1**). Inside the cell, the viral RNA is reverse transcribed into viral DNA by the enzyme reverse transcriptase. Then, the viral DNA is transported to the host cell nucleus and integrates into the host chromosome, with the help of the integrase enzyme. Next, the virus uses the host's cellular machinery to produce copies of its genomic material and creates messenger RNA strands, which are then translated into long chains of HIV-1 precursor proteins. These precursor proteins are then cleaved by the viral enzyme protease into smaller, active proteins that are then assembled into mature virions [6]. Once the viral RNA strands and active proteins are assembled into a new viral particle, the virus buds off from the host cell to infect another cell. The mature virion contains two copies of the viral genomic RNA and functional viral proteins, including reverse transcriptase, integrase, and protease.

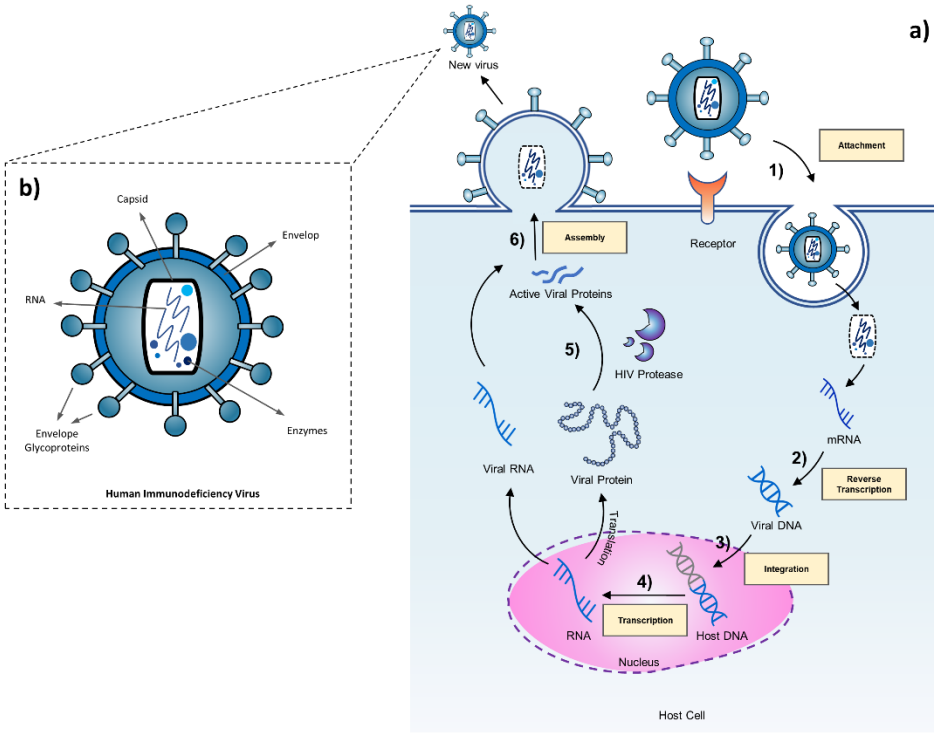


Figure 1.1: a) Life cycle of HIV-1 occurs in 6 major steps: 1) attaching and fusion of HIV virus, 2) reverse transcription, 3) integration of viral DNA into host DNA, 4) expression of viral genes, 5) protein cleavage process by HIV-1 PR, 6) viral assembly and produce new mature virion. **b)** The virus contains two copies of the genetic material RNA and enzymes such as reverse transcriptase, integrase, and protease, which are crucial for its replication cycle, surrounded by a capsid protein shell. The capsid encased in a lipid membrane containing glycoprotein spikes.

As shown in **Figure 1.1**, the significance of HIV-1 protease in the virus's life cycle is critical [1]; without its participation in cleaving viral polyproteins into functional active pieces, the process of reproducing new viruses will be halted. Hence, HIV-1 PR was a prime target for the development of antiviral drugs to treat HIV-1 infection [7]. The structure of HIV-1 protease is a dimer composed of two identical subunits, each of which contains 99 amino acid residues and a catalytic Asp at position 25 **Figure 1.2**.

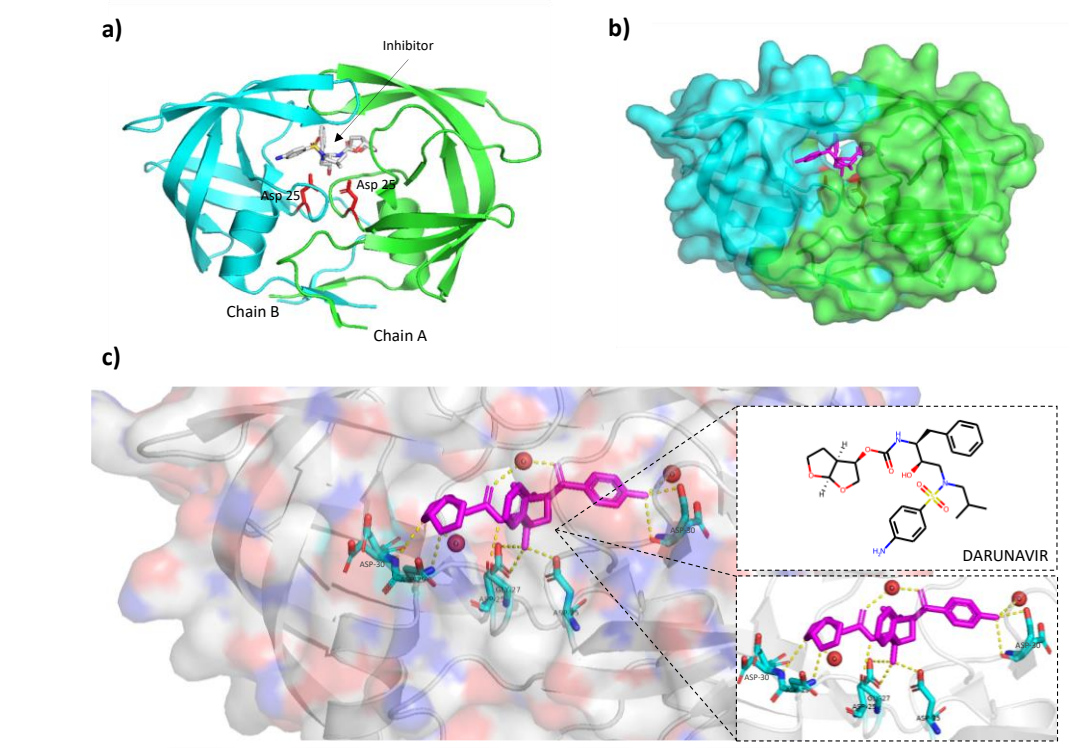


Figure 1.2: General overview of HIV-1 protease wild type in complex with Darunavir inhibitor, **a)** demonstrates the secondary structure of HIV-1 protease, and active site residues highlighted in red. **b)** The protein-ligand complex represented in sphere with different color for each chain. **c)** Shows the protein-ligand interactions within the active site, the water molecules that interact with the ligand in the active site are represented by the red spheres. PyMOL software was used to create these illustrations (PDB id "4LL3" [8]).

Drugs that target the protease active site have been successful in treating HIV infection [9] **Figure 1.3**. However, these drugs are often associated with low efficacy due to the frequent emergence of protease structural mutations, which mostly can lead to drug resistance. Doctors occasionally combine medications to boots up the efficacy of these drugs and control the progressing of the disease depending on the patient's conditions [10]. These facts emphasize the ongoing pressing need for finding new effective antiviral drugs, together with raising awareness and educating societies to prevent the spreading of the virus and support the efforts to save lives.

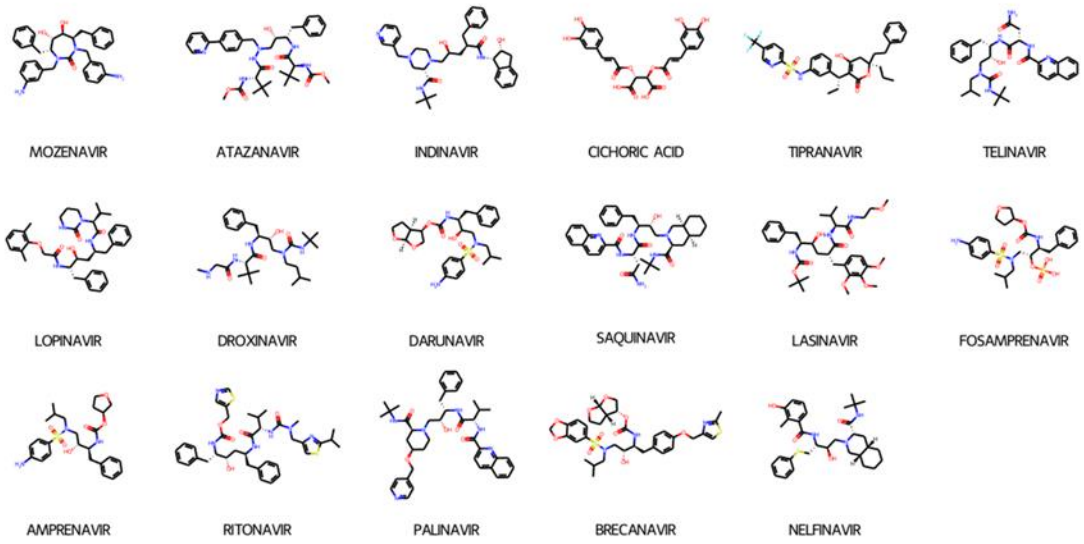


Figure 1.3: This figure shows the structures of common HIV-1 protease inhibitors. These inhibitors have been thoroughly researched and used in the treatment of HIV-related infections (AIDS).

The regular process of bringing a new drug to the market requires serious investments of time, resources, and finance, as it is a complex and rigorous process of development, testing, and regulatory approval. It is estimated to average 10 years [11] with more than one billion dollars [12]. It starts with testing thousands of compounds to identify valuable drug candidates, then it goes through clinicals and most of the candidates fail **Figure 1.4**.

Researchers always try to develop innovative techniques to address the pressing need for new pharmaceuticals by deviating from standard procedures. With recent advances in DL and ML approaches, existing drug and drug-like datasets, and the vast unknown chemical space [13], as well as the rise of molecular languages [14], it is now feasible to uncover novel drug candidates faster and at a lower cost.

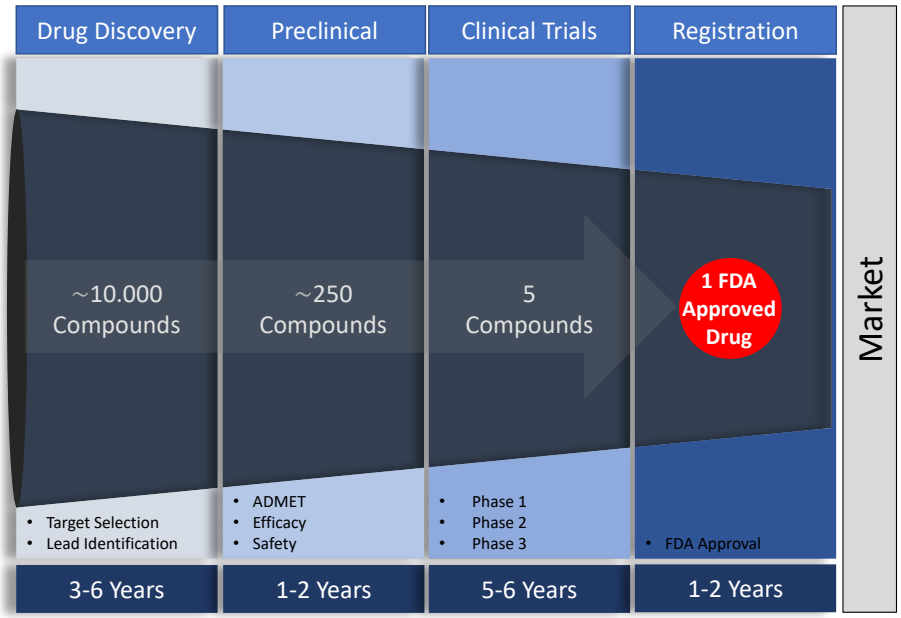


Figure 1.4: The drug development pipeline is divided into five stages: drug discovery, preclinical investigations, clinical trials, and FDA approval. Target selection and lead identification research typically require 3–6 years. The preclinical stage lasts 1–2 years and includes trials on chosen candidates to measure absorption, distribution, metabolism, excretion, and toxicity (ADMET). Clinical trials are undertaken in three stages over 5 to 6 years. Finally, drug candidates that have passed the clinical stages are assessed and registered with the FDA before being released to the market.

LSTM (Long Short-Term Memory) deep learning models (**1.2.4)**, **(1.2.5**) are one of the widely used, gained popularity and attention in the field of molecular synthesis [15], [16]. However, one major challenge in using LSTM models is the difficulty to bias the generation process to be focused on a specific target. This means that the generated molecules may not be suitable for any application. Researchers have proposed various strategies, including using reinforcement learning techniques to optimize the output molecules towards a certain goal, which increased the computing cost. Though, how can we ensure that the molecular models we train are particularly focused on generating inhibitors with the desired properties? Is there a way to explore the potential of the popular LSTM, to maintain the low computational expense while achieving this goal?

In this study we tried to answer these questions by designing a novel strategy for training an LSTM-based model to generate potential HIV-1 protease inhibitors. With the use of SELFIES molecular language as input and filtering the new generated molecules based on properties observed in known HIV-1 protease inhibitors, we were able to direct the model towards generating new inhibitors that look promising for further development.

The results are remarkably encouraging, ensuring the potential of LSTM to be an invaluable tool for identifying new drug candidates that specifically target HIV-1 protease. By refining our model, increasing the data size, and doing further optimization techniques, we believe it can be eminently valuable not only for HIV/AIDS but also for any another draggable targets.

1.2. The Necessary Background

1.2.1. Artificial Intelligence, Machine Learning and Deep Learning

Artificial intelligence (AI) is the emulation of human intellect in robots that are built to think and act like humans. It is reshaping our daily lives quickly [17], [18], and as technologies advance it continues to impact widely on society. Artificial intelligence systems have been employed in a wide range of applications, including natural language processing, picture recognition, health, and self-driving cars. Machine learning (ML) is a subtype of artificial intelligence that allows computers to learn without being explicitly programmed. ML models are utilized in divergent functions, including spam filtering, fraud detection, and product suggestions. Deep learning (DL) [19] is a branch of machine learning that utilizes artificial neural networks to learn from complex data. DL models have complex structures consisting of multilayers, and they are often used to solve hard and complex tasks due to their capability of detecting patterns from large databases. The human brain inspired neural networks, which are used to discover complicated patterns in data. Image identification [20], natural language processing [21], and speech recognition [22] are all examples of deep learning networks implementations.

1.2.2. Molecular representations

Chemical compounds are complex entities that consist of atoms connected by chemical bonds. Representing molecules as input for computer models has been a long-standing challenge. Different methods proposed for this task (**Figure 1.5**). In the end of eighties, the emergence of Simplified molecular-input line-entry system (SMILES) [23] was a significant step forward in chemical informatics, as it allowed for the efficient storage, retrieval, and sharing of chemical information across different platforms. However, SMILES representations have some limitations and weaknesses for chemicals with complex stereochemistry [14], [24]. According to Bjerrum et al. in their investigation, the latent representations derived from canonical SMILES may be less effective since

they become more tied to canonical SMILES grammatical rules rather than the chemical structure of the underlying molecule [25]. This is seen as an obstacle for interpretation and optimization because latent spaces, rather than SMILES syntax rules, contain underlying chemical features and capture ideas of chemical similarity. Thus, the development of new representations was progressing to overcome these problems. Self-Referencing Embedded Strings (SELFIES) another new representation language was recently launched [14], [24], it addressed many of SMILES's weaknesses. SELFIES provides a more flexible and powerful way to represent complex molecules [14], allowing for the inclusion of a wider range of chemical features and functionalities. These languages made it possible to turn molecules into vectors, which could then be used as input for models. Another popular representation method for molecules is molecular graph [26], [27]. There are several sorts of graphs that depict molecules **Figure 1.5d**. The most frequent method is to portray atoms as nodes and chemical connections between atoms as edges. Attempts to portray molecules as graphs date back to the late nineteenth century. With the growing interest in using deep learning techniques for analyzing molecular graphs, it is feasible to anticipate the attributes of novel compounds, such as their solubility and toxicity, by training neural networks using chemical graphs and benefiting from the metrics of graph. However, depending on the research question, the type of constructed graph remains a challenge to choose, as does the rising computational cost of models trained on graphs. Likewise molecular fingerprints are a well-liked style to represent molecules in cheminformatics. A molecular fingerprint is a fixed-length binary vector that records the presence or absence of a molecule's substructural characteristics. The circular fingerprint, also known as the Morgan fingerprint (**Figure 1.5b**), is the most frequent form of fingerprint [28]. This fingerprint is created by iterating through all the atoms in the molecule and considering all potential substructures up to a particular radius. The final fingerprint is formed by hashing each substructure to a fixed-length bit vector and concatenating the resultant vectors. The resultant fingerprint is usually sparse, with most bits being zero.

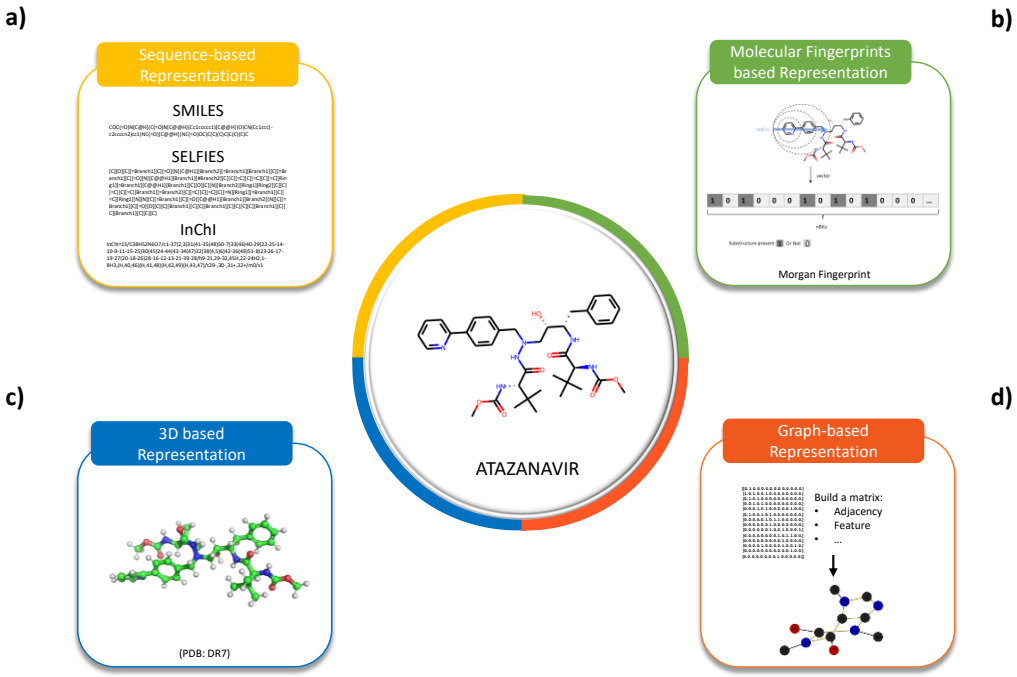


Figure 1.5: Different Representations of "ATAZANAVIR" compound, **a)** shows sequence-based representation such as Smiles, Selfies and InChI. **b)** shows molecular fingerprints-based representation, it shows the resulting vector of the Morgan fingerprint represents the structure encoded into a binary vector, indicating the presence or absence of certain substructures within the molecule. **c)** shows the three-dimensional representation of the molecule, which displays its position in space using the x, y, and z coordinates. **d)** illustrates the graph-based representation of the molecule, where the resulting graph of the adjacency matrix is presented. This graph highlights the bonds and connections between atoms in the molecule.

1.2.3. Supervised Learning and Unsupervised Learning

Supervised and unsupervised learning are the two basic types of machine and deep learning algorithms [29]. Supervised learning is the use of labeled data to train a model to predict outcomes for fresh, unseen data. Unsupervised learning, on the other hand, is employed when there is no labeled data, and the aim is to detect patterns and structure in the data, then use them to predict outputs for new data. In supervised learning, the model learns from the available training data to generalize predictions to new data. with contrast, with unsupervised learning, the model discovers links and structures in the data without any prior information or assistance. These two types of learning are employed in a variety of deep learning applications.

1.2.4. Deep learning generative models

Deep generative models have transformed the fields of machine learning and artificial intelligence. Generative Adversarial Networks (GANs) [30], [31], Variational

Autoencoders (VAEs) [32], and Recurrent Neural Networks (RNNs) [33] are three of the most popular and successful deep generative models. The main use of these models is generating new data with favorite properties. These models can learn the uncovered patterns and structures of the data, then use them in sampling new similar data to those in the training sets. These models have been used in various domains, such as image and audio generation [34], natural language processing and many others.

As shown in **Figure 1.6c**, GAN [35] is made up of two primary models, a generator that learns the prior map distribution of the input data to sample new data, and a discriminative model that learns to identify whether samples are real or fake. These two models were stochastic gradient-descent trained in parallel. Through repeated iterations, the generator network learns to generate increasingly realistic samples that can be used for various tasks. GANs have been utilized successfully for picture generation, audio synthesis, and text production.

VAE [32] is another widely used generative model. Encoder and decoder are the two components of the basic architecture (**Figure 1.6a**). The input data is encoded by the encoder network into a latent space, and new data is then decoded from the latent space by the decoders network. Latent space is a lower-dimensional space that captures the essential features of the input data. VAE is particularly flexible for a variety of applications since it can employ multiple types of data, including images, audio, and text.

RNN [33] is another class of deep generative models, these networks are primarily designed for processing sequential data, they can learn the dependencies within sequences. As shown in **Figure 1.6b**, repeating units, also called cells, which process each element in a sequence and have hidden state vectors that contain information for all preceding elements are the standard structure of an RNN. Each cell is taking the currently input and previously hidden State Vector as inputs, updating it with a specified set of learnt parameters. When the next element is added in a sequence, an updated hidden state vector will be returned to another cell for processing. Therefore, the recurrent nature of RNNs' gives them the ability to keep memory of one step back input while processing the current input of the sequence and make them powerful and adaptable classes of generative models in generating new data. RNNs have been successfully used in a wide range of applications. Natural language processing, speech recognition, music creation, and time series prediction are just a few of the many uses for RNNs. The capacity of RNNs to handle variable-length sequences is one of their key advantages, which makes them

suitable for a variety of problems from the real world. However, on the other hand, RNNs are vulnerable to the vanishing gradients problem [36], in which the gradients get progressively smaller over time. Long Short-Term Memory (LSTM) [37] and Gated Recurrent Unit (GRU) [38] are two RNN variations that were developed to overcome this problem. Gating methods are added to LSTM and GRU architectures to control the information flow, which improves their ability to capture long-term dependencies and prevent the vanishing gradients problem as we will present in next section (**Figure 1.7**).

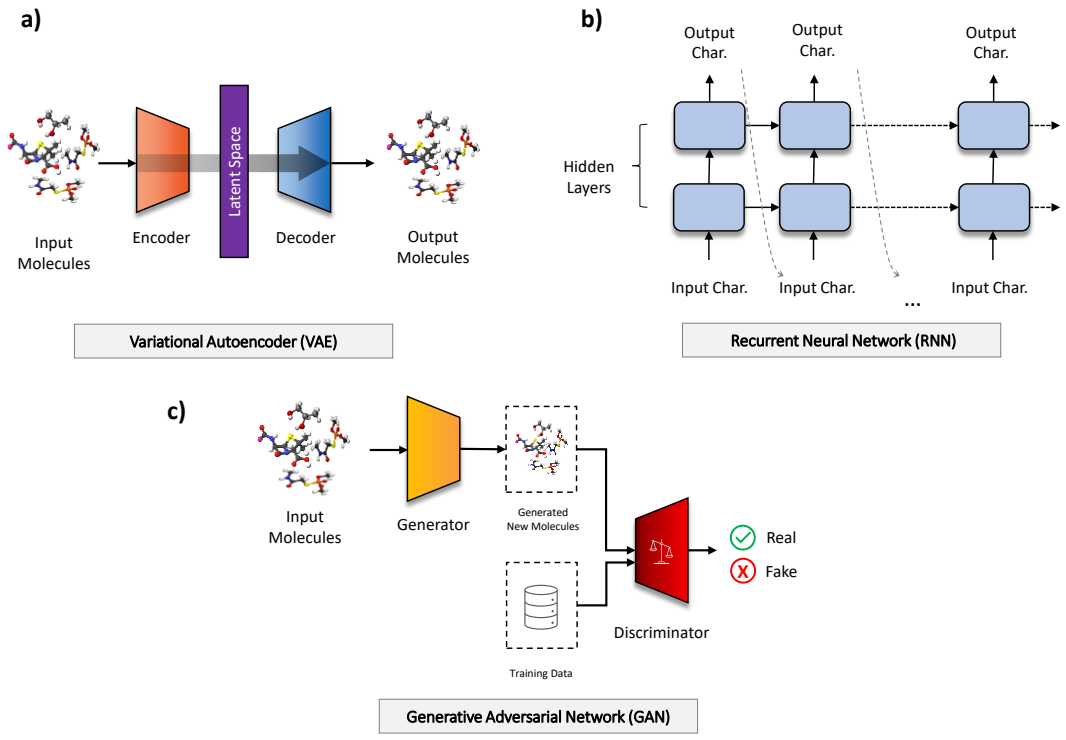
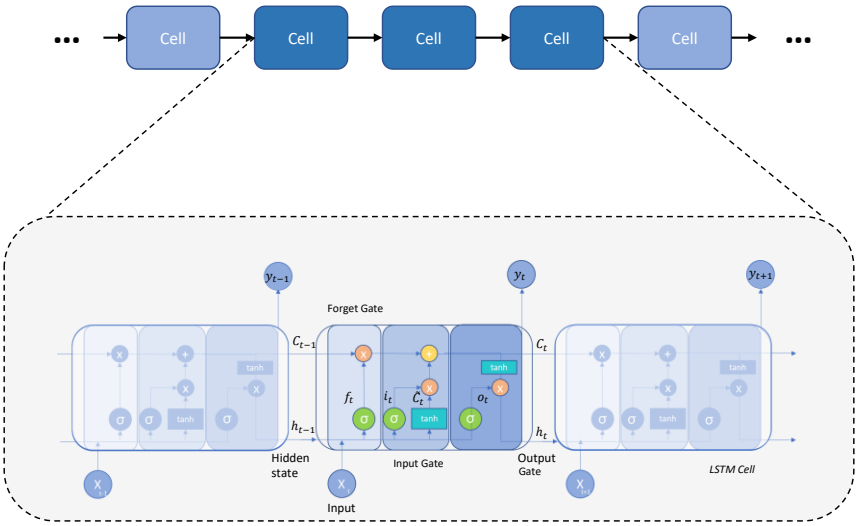


Figure 1.6: Shows various generative deep learning models used for molecule generation. **a)** shows a variational autoencoder (VAE) model, which consists of two sub-models, an encoder that maps input molecules into the latent space and a decoder that scans the latent space for new molecules. **b)** highlights the Recurrent Neural Network (RNN), a model used to generate new sequential data. **c)** introduces the architecture of Generative Adversarial Networks (GANs). A GAN consists of two models, a generator that generates new molecules and a discriminator that is individually trained to distinguish between real and fake molecules.

1.2.5. Long Short-Term Memory (LSTM)

LSTM [37] is a sort of neural network, it has transformed the way we think about artificial intelligence, opening new possibilities for handling of sequential data. It proposed to solve the problem of vanishing gradients in recurrent neural networks (RNNs). LSTM

networks function by forming a memory cell that can recall or forget information selectively over time. This enables the network to remember information from previous inputs and use it to predict future inputs. The cell is regulated by a set of gates that control the flow of information through the cell (**Figure 1.7**). A LSTM cell is made up of these three gates and a memory cell, the first gate is known as the Forget gate, the second as the Input gate, and the final as the Output gate and the cell state is a memory that is passed from one LSTM cell to the next. As shown in **Figure 1.7**, The input gate controls the amount of current that enters the cell. The forget gate manages how much of the previous cell state is forgotten; it selects whether to maintain or discard information from a previous time. The output gate regulates how much of the cell state is sent out. One of the most notable advantages of LSTM networks is their capacity to handle long-term dependencies, which are critical when working with sequences.



Long Short-Term Memory Network

Figure 1.7: The architecture of Long Short-Term Memory (LSTM) network for sequential data processing. First, the forget gate $\boldsymbol{f_t}$ multiplies the state of earlier cell $\boldsymbol{C_{t-1}}$. The cell's candidate state $\boldsymbol{\widetilde{C}_t}$ is then multiplied by $\boldsymbol{i_t}$ and added to the input gate value $\boldsymbol{C_t}$. Afterwards, the hidden state $\boldsymbol{h_t}$ and the output state $\boldsymbol{o_t}$ computed. To establish the value gate's output, the input x_t and final hidden state $\boldsymbol{h_{t-1}}$ were first computed using sigmoid. The cell's hidden unit state is acquired by multiplying the cell $\boldsymbol{\widetilde{C}_t}$ by the output gate $\boldsymbol{o_t}$ following a tanh operation.

LSTM models, originally proposed for natural language processing problems [39], [40]. They have evolved as a popular deep learning technique [21], finding application in audio and image analysis. One of the most fascinating uses is in drug discovery [16]; it inspired researchers to investigate the use of natural language processing models in the field of

drug development by representing molecules as text snippets. LSTM networks showed incredible success in learning the underlying patterns and correlations between chemical formulas and properties by training them on datasets of existing molecules [41]. This information is then utilized to create new molecules with the required characteristics.

1.2.6. Inductive Biasing

Incremental training [42] is a deep learning technique for incrementally updating a pre-trained model with new data. This approach empowers the model to adjust in the distribution of data over time, boosting its performance on future predictions. This assists the model to generalize new data more effectively and produce more accurate predictions. Furthermore, incremental training can help to alleviate the problem of catastrophic forgetting, which occurs when the model forgets what it has learned about previous data when training on new data. This problem was avoided by gradually feeding new data into the model and continuing to enhance its performance over time.

1.2.7. Molecular Docking

The purpose of molecular docking studies is to predict the pose and affinity of small molecules (i.e., drug candidate) to a protein target [43], [44]. As shown in **Figure 1.8**, the procedure entails creating several ligand and receptor conformations and then computationally docking them together to identify the best binding position. The docking score, which assesses the intensity of the ligand-receptor interaction, is then computed, and used to rank the ligands based on their binding affinity. This approach is extensively used in drug development to detect and optimize prospective drug candidates' binding characteristics.

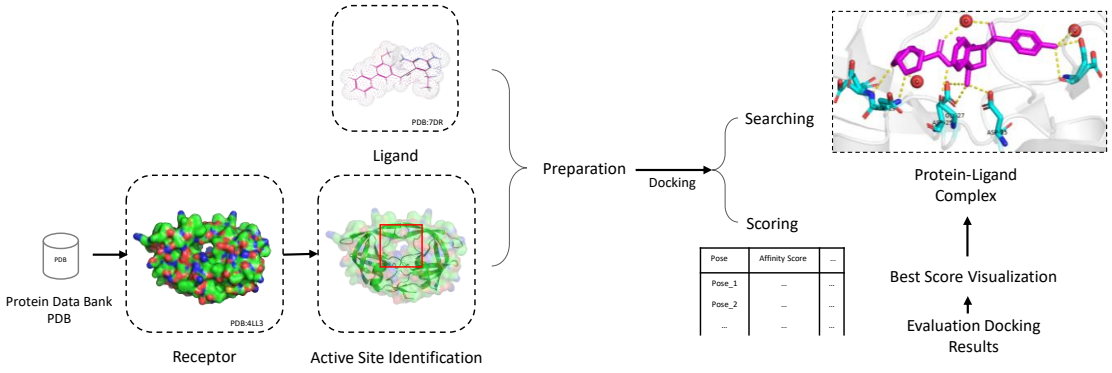


Figure 1.8: Shows the stages involved in docking a ligand into a receptor. starts with obtaining the receptor's structure from the Protein Data Bank and identifying the active site as well as preparing the ligand. The (**2.10)** section thoroughly explains the

preparation process. Once the ligand and the receptor are ready, the docking procedure begins, which employs search and scoring algorithms to estimate ligand binding affinity to a protein target.

1.3. Related Work

The area of drug discovery research rapidly advances to meet the ongoing need for new, and effective medications. Deep learning generative models have shown untapped potential for the discovery of novel drug candidates along with the huge public chemical compound databases such as ChEMBL [45] and DrugBank [46], and the vast unexplored chemical space [13], [47]. The number of potential undiscovered chemical compounds has been estimated to be upwards of 10^{60} [47], [13]. This suggests that there might be millions of new molecules with unique chemical characteristics that could lead to the development of innovative and more effective medications. Recently, various deep learning models have been proposed and applied to generate novel molecules with desirable properties.

In this section, we review some promising proposed models for molecular generation and show how they have assisted in the drug development process. The unifying thread connecting all these models is the need to discover a means to steer the generation process after training to be focused on the target.

Gómez-Bombarelli et al. [48] used a variational autoencoder (VAE) based model in their approach for developing compounds with specified attributes such as high chemical stability and solubility. They trained the VAE model to understand the underlying distribution of chemical structures in the used dataset, allowing it then to build unique compounds with desired characteristics. They proved the efficacy of their method on many datasets, including the QM9 database [49]. Their technique generates novel compounds with certain properties; however, it requires very large datasets to train the model on, which significantly increases the computing cost.

Jin et al. [27] have developed an innovative approach for generating molecular graphs, named junction tree variational autoencoder (JT-VAE). It utilizes the advantages of a graph-based representation of molecules, with nodes representing atoms and edges representing chemical bonds. The JT-VAE generates new, valid molecular structures with desired properties by iteratively expanding and linking groups of connected atoms. This strategy intriguingly shows a promising methodology in the drug development process.

While the approach is effective in learning the latent distribution of molecular networks, it relies on an algorithm to convert these graphs into actual molecular structures, which may not always fulfill expectations, and the decision on the gradually added fragments. Furthermore, the computational expense of transforming molecules into graphs and training two sub-models may present difficulties.

As previously stated, generative adversarial network (GAN) is also widely employed for molecular generation. For the same purpose, Prykhodko et al. developed a novel design termed LatentGAN [31], which is a truly unique deep generative model for de novo molecule generation that combines an RNN-based heteroencoder with a generative adversarial network with two sub-models: generator and discriminator. They first trained the model on a portion of the ChEMBL [50] database of drug-like molecules and then retrained it on three target-biased datasets to generate target-biased compounds. Following this procedure, their results demonstrated that LatentGAN is a powerful tool for creating novel biased compounds. However, there is little empirical data to support the validity and usefulness of these novel compounds, necessitating more research. Encoding molecules to latent vectors without a thorough investigation and the high computational cost involved in training multiple models separately rais concerns about the methodology's efficacy. MolGAN [26] is another implicit generative model for producing small molecules, introduced by De Cao and Kipf. It is made up of two primary components, as is typical of GAN models: a generator network and a discriminator network. The generator network creates new molecular graphs, while the discriminator network verifies the resulting graphs' validity. The two networks are trained in a competitive way. MolGAN is trained on the Zinc database, which contains over 250,000 compounds, and creates new molecular graphs comparable to those in the database. MolGAN was capable to create high-quality graphs while still retaining features critical for drug discovery and development, such as solubility and drug-likeness.

RNN-based models' success in natural language processing has inspired researchers to adapt this methodology to molecular science. These models have successfully generated new compounds with promising drug discovery potential by modeling molecules in linear presentation languages such as SMILES, SELFIES, and InChI (**1.2.2**). The combination of machine learning and molecular science throws up fascinating potential for the development of novel treatments, leading to a promising future research field. Segler et al. [51] is a potential step forward in the development of novel methodologies for de novo

molecular design. They defined a three-layer RNN-LSTM model for developing customized libraries targeting the serotonin receptor as well as other antibacterial targets. They trained the model with SMILES molecules from the Zinc15 database. They refine the model with small groups of compounds known to be active against that target. The authors compared the chemical space that the new compounds cover to that of the molecules in the training set as a way to evaluate the new molecules. Additionally, another successful study proposed by Gupta et al. presented a LSTM-based model for generating molecular libraries with structural similarity to known actives for $PPAR_{\gamma}$ and trypsin. The model was trained using 550,000 SMILES strings of active chemicals from ChEMBL before being fine-tuned with SMILES strings of 4,367 $PPAR_{\gamma}$ ligands and 1,490 trypsin inhibitors. Around 90% of the created compounds are distinct from the known ligands and from one another. The suggested approach was also evaluated for fragment-based drug development. This, together with the other models we have discussed, highlights the extraordinary potential of new deep generative approaches in the domain of drug discovery. Following this extensive review of the most promising molecular generative models, it is important to highlight their limitations and identify weaknesses in them. By noting these shortcomings, we can open the doors for novel approaches that overcome the current challenges and contribute to the development of more effective approaches. In the following sections, a list of three main notable shortcomings observed in the existing works is included below, focusing on places where changes are needed. By acknowledging these shortcomings, we can lay the groundwork for innovative solutions and methodologies to advance in this field.

- Molecular encoding

While the graph representation is extensively employed in most of the suggested techniques to encode molecules due to its capacity to allow the use of graph theory and its application, the complexity and computational cost has seriously grown. Furthermore, deciding what features should be recorded in the molecular matrices that would subsequently be represented as graphs necessitates considerable skill because it changes based on the subject and the purpose of the research, which makes it challenging.

- Model architectural complexity

Several complicated models (e.g., VAE, GAN) have been presented for the task of producing molecules; however, these models need training two sub-models, increasing the computational expense, and work required to assess the performance.

- Guiding the generation process:

Most of the suggested models almost face the same two tough challenges: guiding the generation process to create molecules with desired attributes, together with learning the chemical rules embedded in the molecules during the training process to generate valid molecules. Reinforcement learning was frequently employed to improve these models' output molecules. However, it raises the overall expense and the essential endeavors of the subject.

Despite the presence of these challenges, it is important to acknowledge that these models show a future in which innovative and optimized medications may be created more efficiently and effectively, supporting the industry, and eventually enhancing human health. The combination of machine learning and molecular science is an enthralling topic with limitless potential, and our study contributes to the expanding work of investigating this interesting area of inquiry.

1.4. Overview of the Proposed Solution

In this work, we proposed an LSTM-based model (LSTM-ProGen) to generate new antiviral candidates targeting HIV protease using SELFIES molecular language. We incrementally trained the model on known drugs and protease inhibitors from the ChEMBL database. We employed an augmentation method to increase the number of samples in the training sets. The resulting molecules were filtered depending on desired properties observed from the known HIV-1 protease inhibitors, and different similarity measures were calculated. The top ranked molecules with the highest similarity scores were selected. Finally, we performed a molecular docking study on the resulting set of molecules to test their validity and activity against the HIV-1 protease and select the vest ones. The results were promising, and we believe that with further refining and improvements, this model has the potential to be highly valid for any ligand-able protein targets in drug repositioning research.

1.5. Objectives and Contributions of the Thesis

In the presented study, we propose an innovative deep learning model designed to generate novel candidate compounds specifically targeting HIV-1 protease as a potential therapeutic target for the treatment of AIDS. To achieve the goal, a meticulous multi-step process is employed, leveraging data from the ChEMBL database. The performance of the model is thoroughly evaluated using various metrics, ensuring a comprehensive analysis. The key contributions and noticeable achievements are listed below.

- Development of a multi-layered novel deep learning framework for generating new molecule candidates targeting HIV-1 protease.
- Implementation of an LSTM-based model trained within a pyramid scheme training strategy to bias the generation process towards the target protein.
- Utilization of the SELFIES molecular language for encoding and decoding molecules for feeding up the proposed model.
- Evaluation of the molecular modeling performance of the generated molecules using various methods inspired by RDkit, a widely used cheminformatics library.
- Conducting a molecular docking study to assess the binding affinity and potential interactions between the final filtered generated molecules and the HIV-1 protease target.
- Exploration of different strategies and techniques to enhance the generation of molecules with improved properties and potential for inhibiting the HIV-1 protease.
- Contribution to the field of computational drug discovery by proposing a deep learning-based approach for generating novel molecules targeting specific proteins involved in diseases with a novel training strategy.
- The potential for discovering new lead compounds for developing anti-HIV drugs through the generated molecule candidates.
- Enhancing the understanding of the capabilities and limitations of deep learning frameworks for generating molecules and their evaluation in the context of HIV-1 protease drug discovery.

1.6. Organization of the Thesis

This thesis is organized into five chapters, each of them serves a specific purpose and contributes to the overall understanding of the research.

CHAPTER 1 provides an overview of the problem, along with the inspiration for the proposed solution. It includes background information on HIV/AIDS disease and discusses various related works and contributions to the field. **CHAPTER 2** focuses on explaining the used techniques, concepts, theories, and principles that underpin the proposed solution. **CHAPTER 3** describes the methodology of the research work. **CHAPTER 4** presents the findings of the research. **CHAPTER 5** summarizes the study and illuminates its significance in the realm of pharmaceutical innovation. Furthermore, it delves into the obstacles encountered and outlines future work.

CHAPTER 2

2. THEORETICAL PART

2.1. Components of the Developed Methodology

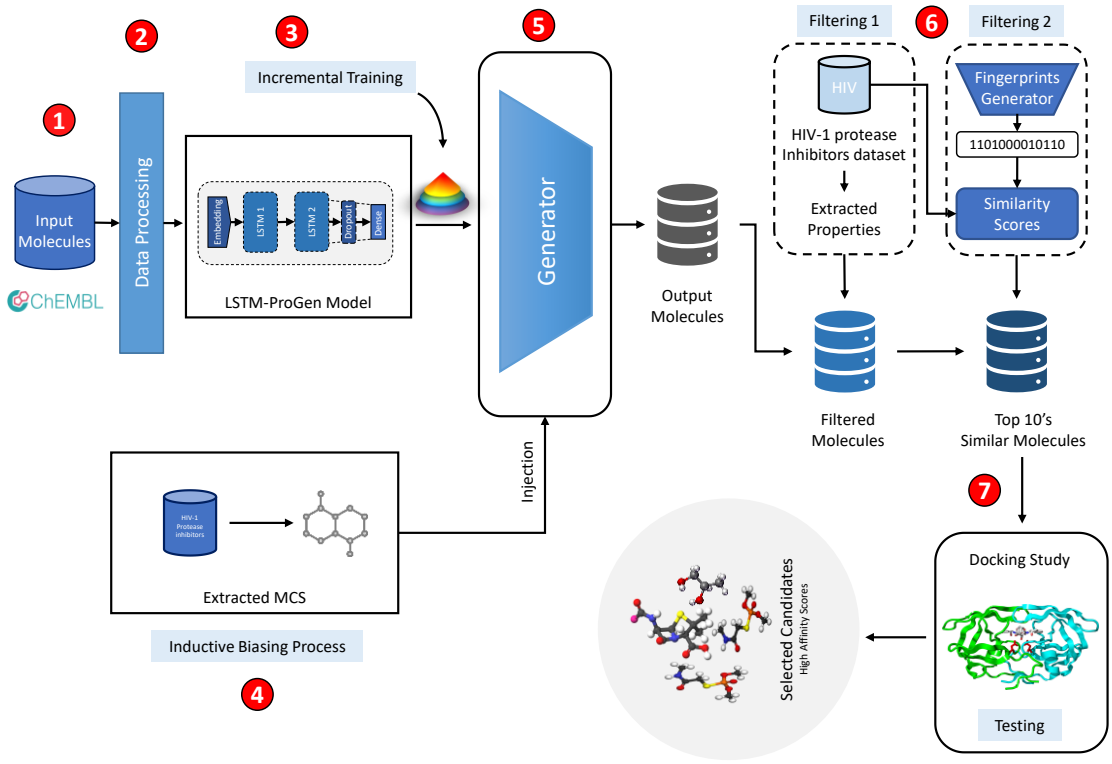


Figure 2.1: An overview of the proposed multi-step approach.

Steps of the proposed approach are depicted in **Figure 2.1**. The process was divided into seven phases. (1) Different datasets were extracted from the ChEMBL dataset, which was utilized as the primary source of input molecules. (2) The input molecules were cleaned and encoded into vectors as part of data processing. (3) A 2-layer RNN-LSTM model was defined and incrementally trained. (4) The maximum common sub-structures from the HIV-1 protease dataset were then extracted and injected into the generator during the process of creating new molecules. (5) The generator was then used to produce new

molecules. (6) The resulting molecules were filtered based on their descriptors, and Morgan fingerprints were generated for calculating similarity scores between the new ones and the original ones. (7) Finally, running molecular docking study to select the inhibitors with high affinity scores.

2.2. Data Collection

In the study, several datasets from the ChEMBL database [45], [50] were utilized for the investigation to accomplish the objective. ChEMBL database is an extensive collection and open-access database. It contains bioactive compounds with drug-like qualities and offers details on their biological functions, pharmacological characteristics, and molecular structures. The initial version of ChEMBL was created and maintained by researchers at the European Bioinformatics Institute (EBI). The ChEMBL project was established to fill the demand at the time for a centralized repository of bioactivity data. The data was collected from a variety of sources throughout the years, including books, patents, and open databases. The database now has more than 2 million bioactivity data points, including more than 1.8 million chemicals and 15,000 targets. The database encourages researchers to share data and work together, which can speed up drug development studies and deepen our understanding of the underlying causes of disease. It served as a helpful resource for drug discovery research, especially for small and medium-sized businesses, university researchers, and non-profit groups that lacked access to private databases.

The first used dataset in our study is “14K drug” the sub-dataset from ChEMBL database. It contains around 14.000 authorized small molecules. Invaluable details regarding the molecular structure, molecular weight, number of hydrogen bond acceptors and donors, partition coefficient, and other physicochemical parameters are all contained in this extensive dataset. We applied filters to remove only the unfavorable molecules that did not share characteristics with well-known HIV-1 protease inhibitors. The resulting dataset was 9480 molecules with their characteristics. The main goal of utilizing this dataset was to improve the model's capacity to produce compounds with drug-like features similar to those in this dataset.

Hepatitis C virus (HCV), Equine Infectious Anemia Virus (EIAV), West Nile virus (WNV), and SARS-CoV-2 protease inhibitors with a comparable mechanism to the HIV-1 protease were the only ones left after filtering the "14K Drugs" dataset to create the

second dataset. After cleaning this dataset, it only contained 391 molecules with properties like molecular weight, octanol-water partition coefficient (LogP), Polar surface area, and others. With the help of this strategy, the model was able to concentrate on the target as a whole and generate compounds that were comparable to those in the training set, starting with drug-like molecules and moving toward protease inhibitors.

The approved HIV-1 protease inhibitors dataset, which includes 17 compounds showed in (**Figure 1.3**), was the final dataset utilized in this study. It was used lastly to train the model to produce similar compounds with drug-like and HIV-1 protease inhibitor characteristics.

2.3. Exploring HIV-1 Protease Inhibitors Dataset

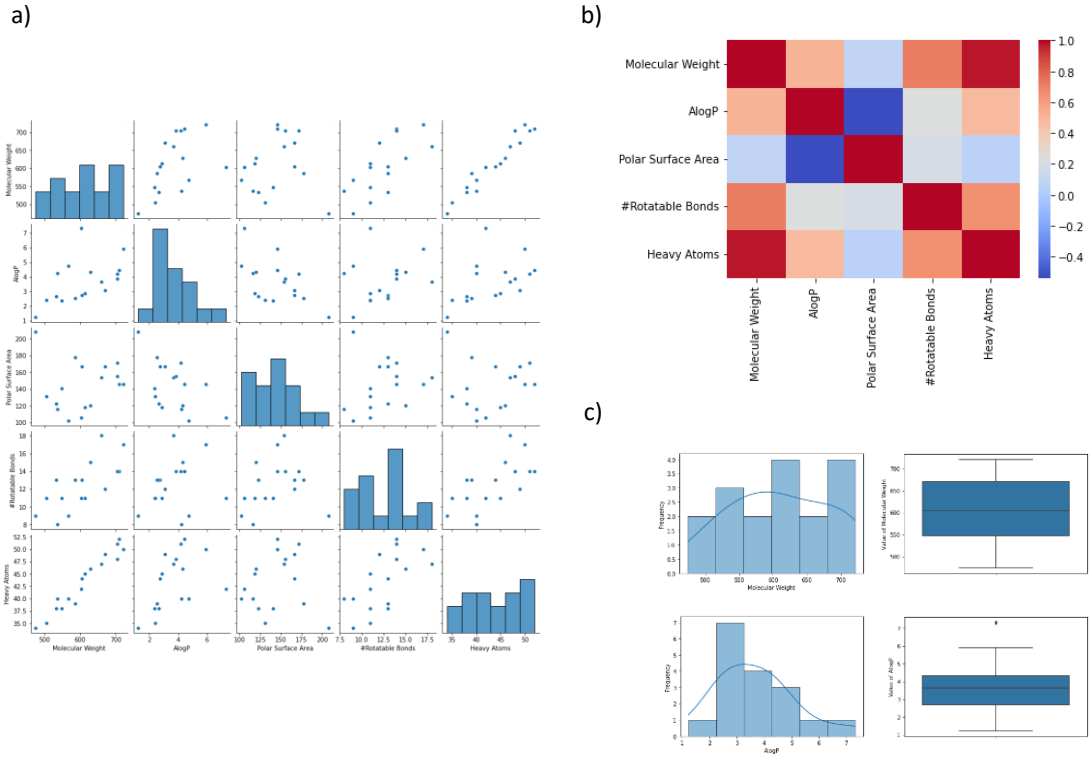


Figure 2.2: Examining the dataset of HIV-1 protease inhibitors. **a)** Display the interactions between certain variables in the dataset. **b)** The heatmap displays the correlations between the dataset attributes. **c)** Histograms and boxplots display the molecular weight and AlogP ranges in the dataset.

We analyzed the dataset of known HIV-1 inhibitors to gain a deep understanding of the relationships between certain features and ensure the efficacy of the newly synthesized compounds against HIV-1 protease. As shown in **Figure 2.2**, We created a heatmap to show the connections between different metrics. The range of values for the selected

molecular features, such as the number of rotatable bonds and polar surface area, was detected and listed in **Table 2.1**. With this knowledge, we can filter the newly created compounds based on these precise ranges of values for the given attributes. This will considerably increase their capacity to target and block HIV-1 protease. Using this method, we can dramatically boost the chance of discovering good candidates for future drug development efforts.

Table 2.1: Examining the distribution of the molecular properties of the HIV-1 protease known inhibitors dataset: Molecular Weight, AlogP, Polar Surface Area, Rotatable Bonds, and Heavy Atoms.

	Molecular Weight	**AlogP**	**Polar Surface Area**	**Rotatable Bonds**	**Heavy Atoms**
Mean	610.048	3.685	143.927	12.529	43.411
Std	76.350	1.470	28.669	2.718	5.657
Min	474.37	1.23	101.9	8	34
25%	547.67	2.68	120	11	39
50%	604.75	3.66	145.78	13	44
75%	670.86	4.33	166.75	14	48
Max	720.96	7.33	208.12	18	52

2.4. Data Augmentation

The limited number of data samples in the last two datasets was one of the main obstacles we faced in this study. Training our model with a low number of samples, particularly in the case of the HIV-1 protease inhibitors dataset, presented a tough challenge in producing similar samples of training examples to guarantee optimal model performance. To master the lack of data, we employed an innovative technique called "Smile Enumeration" to augment our data samples and rise the size of the datasets. It was previously published by E.J. Bjerrum [52] and has proven to be very promising for data augmentation and enhancing the performance of molecular models. The updated sizes of the last two used datasets are as follows: 2756, 1717 for the second, and third dataset respectively.

2.5. Maximum Common Sub-structures

As shown in **Figure 2.3**, the maximum common sub-structures (MCS) between the known HIV-1 protease inhibitors were extracted using a straightforward function from the RDkit library [53]. This function (rdkit.Chem.rdFMCS.FindMCS) takes a list of molecules and returns the MCS as a SMILES Arbitrary Target Specification (SMARTS) pattern. The extracted MCS will then be injected into the generator function to improve the newly generated compound's ability to result in similar structures to the known HIV-1 protease inhibitors as we will show in the coming sections.

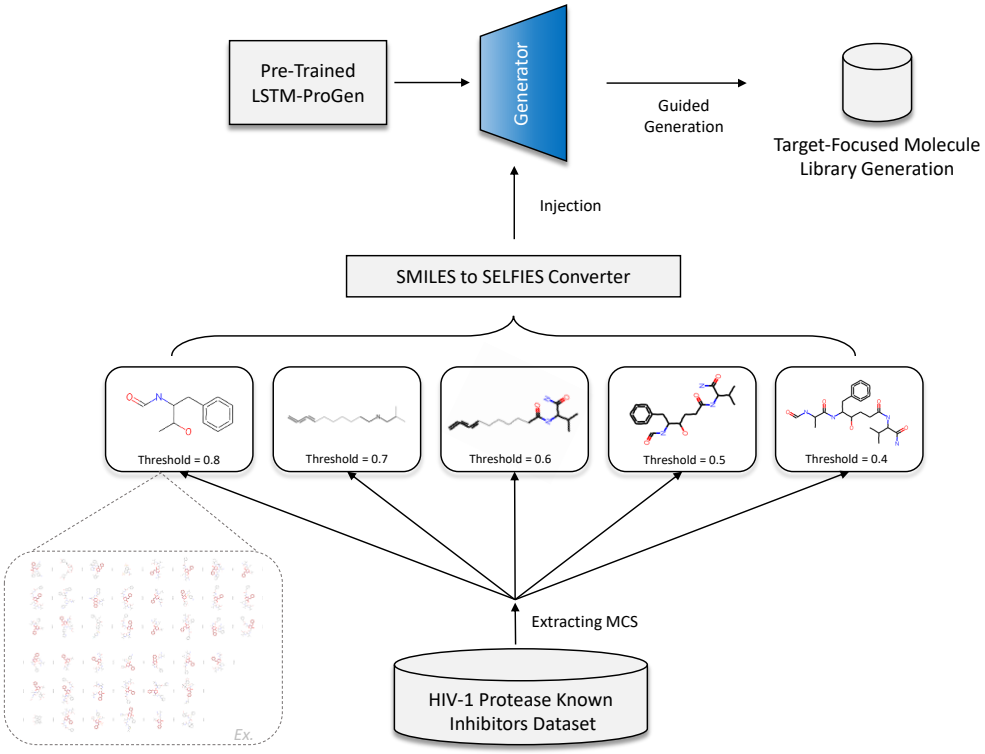


Figure 2.3: The process of extracting the most common substructures (MCS) from a dataset of known HIV-1 protease inhibitors at various thresholds obtained from the ChEMBL database, converting these substructures to SELFIES representation, and injecting them into the generator function to direct the generation process to be focused on the target.

2.6. Data Processing

Before starting the training process, the datasets must be ready to guarantee that the created models can learn effectively. As seen in **Figure 2.4**, this pretreatment process consists of several steps, including purging the datasets of various contaminants,

including salts, NaNs, and duplicate rows. The next step is to convert the SMILES format to SELFIES representations once the datasets have been cleansed.

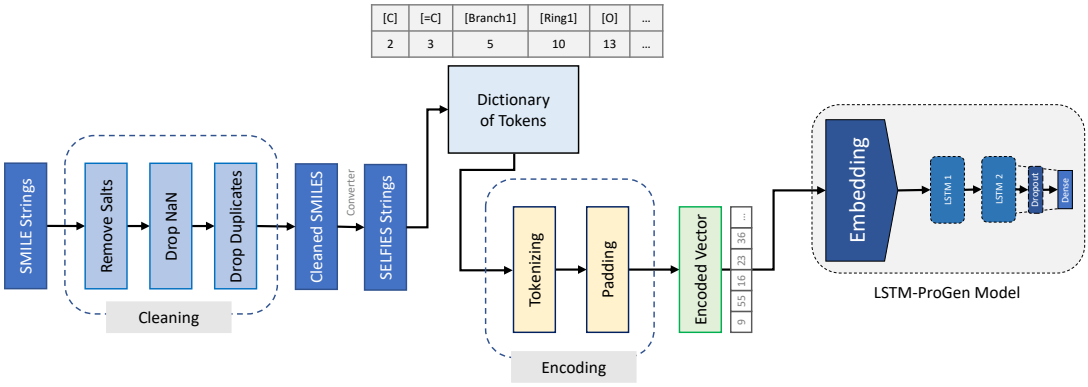


Figure 2.4: An overview of the data processing procedures included cleaning the datasets from salts, NaNs, and duplicated rows; converting SMILES to SELFIES representations; creating a token dictionary that assigned a unique number to each character found in the datasets; encoding each molecule into a vector; and finally padding these vectors into a fixed length of 100 characters. The resulting vectors were fed into the defined model (LSTM-ProGen) at the first embedding layer.

Later, the dataset was split randomly into two sets: a training set with 76% of the data and a test set with 24% of the data. We used a sliding window method that was proposed by Gupta et al. [41] (**Figure 2.5a**). Every sequence is prefixed with the character "[nop]" at the beginning of the sequence to differentiate between the training and testing sequences. Next, a token dictionary is created to give each unique character found in the datasets a special number. There were 86 of these characters. This enables us to vectorize the molecules and encode them for feeding into the model. Finally, each sequence was padded into a constant length of 100 characters, which is the maximum length of string found in the dataset, to ensure stability in the learning process of rules. The LSTM-ProGen model's first embedding layer is then fed with the resulting encoded molecules.

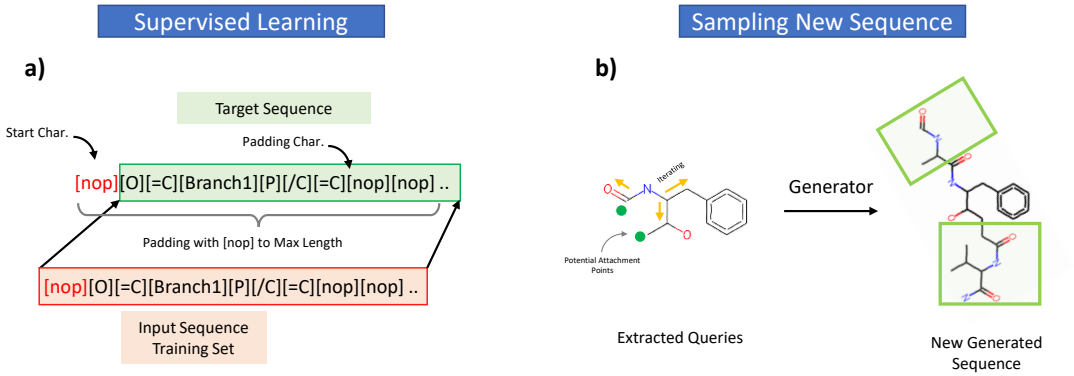


Figure 2.5: **a)** Assigning labels to the sequences. Each sequence had the prefix "[nop]" assigned to the beginning, and the length of each molecule was padded to the longest

string found in the datasets. denoted by “[nop],”. As the model was being trained to learn it, the red sequence was initially used as the input and the green sequence as the goal. **b)** Explains the sampling process of a new sequence using the trained model and the extracted queries (**2.5**), The generator function iterates through each query and adds new characters based on the information it learned from the training phase.

2.7. LSTM-ProGen Model

2.7.1. Model architecture

With the goal of creating new drug candidates for HIV/AIDS, we proposed two distinct models with slightly different processing of the input data. As shown in **Figure 2.6**, the proposed architecture consists of five layers: an embedding layer, followed by two LSTM layers, next a dropout layer, and lastly a dense layer. The embedding layer transforms each character in the input sequence into a vector representation in a dense vector space that can depict the semantic meaning of the characters and their relationships. This enables easier the model to understand the given sequence and helps it to observe more useful characteristics. The model can learn to generalize more effectively by using an embedding layer, which makes it possible for it to recognize and understand similar characters or sequences of characters even when they weren't in the training set of data [54]. The two LSTM layers have 128 and 32 hidden units, respectively, and allow the model to capture the temporal dependencies of the input sequence. The dropout layer helps prevent overfitting by randomly removing connections between the LSTM layers. The dense layer then converts the output of the LSTM layers into an output vector with 86 dimensions. Finally, the dense layer then converts the output of the last LSTM layers into an output vector with 86 dimensions. The total number of trainable parameters in this model is 242,582. The first LSTM layer has 197,120 parameters, the second LSTM layer has 20,608 parameters, and the dense layer has 2,838 parameters. The embedding layer has 22,016 parameters. This design was inspired from the natural language processing application and successfully balances computational efficiency and complexity, making it effective for a range of applications.

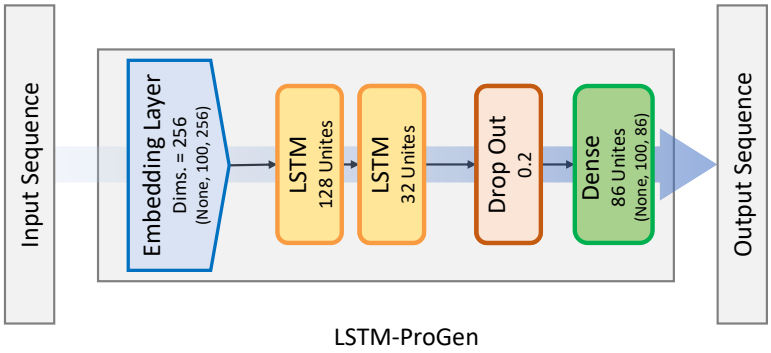


Figure 2.6: Illustration of the LSTM-ProGen architecture used for designing HIV antiviral drugs. Embedding layer, followed by 2 LSTM layers with 128 and 32 units respectively, Dropout layer and Dense layer with 86 units and SoftMax activation function.

2.7.2. Activation function

In our proposed model, we used SoftMax [55] as the activation function, which maps the output of the final dense layer to a probability distribution over the possible characters in the molecule vocabulary. The SoftMax formula is as follows:

$$\boldsymbol{f(s)_i = \frac{e^{s_i}}{\sum_j^k e^{s_j}}} \qquad \textbf{(2.1)}$$

(E 2.1) where s is the input vector made up of ($\boldsymbol{s^0}$, ..., $\boldsymbol{s^k}$), all the $\boldsymbol{s^i}$ values are the elements of the input vector to the SoftMax function, $\boldsymbol{e^{s_i}}$ is the standard exponential function applied to each element of the input vector, $\sum_j^k \boldsymbol{e^{s_j}}$ is the normalization term, which assures that all the function's output values add to 1 and are in the range (0, 1), resulting in a legitimate probability distribution and k the number of classes. SoftMax is particularly useful for multiclass classification problems because it ensures that the sum of the probabilities for all possible classes is equal to 1. This means that the model can generate new molecules that are valid and follow the rules of the chemical language.

2.7.3. Loss function:

The loss function measures the difference between the predicted output and the actual output, and the aim is to minimize this difference during training. We trained the model using categorical cross entropy to ensure optimal performance. It is a commonly used loss function for multi-classes classification problems, such as sequence generation, where the number of unique characters presents the classes. Once the SoftMax activation function gives the predicted distribution output, the categorical cross entropy loss function

calculates the difference between the predicted probability distribution and the actual one, where the actual distribution is represented as a one-hot encoded vector. Throughout the training, the model predicts the character token for each input token in the sequence. At each step the loss L is calculated as the categorical cross-entropy between the predicted and actual next token through this formula:

$$\boldsymbol{L}(\tilde{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{\boldsymbol{i}} \boldsymbol{y_i} \log(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}) \qquad (\mathbf{2.2})$$

(E 2.2) where y is the actual element and $\tilde{\boldsymbol{y}}$ is the predicted element for class $\boldsymbol{i}$. For better understanding, here is an example, suppose the actual distribution is [0, 0, 1, 0] while the predicted distribution is [0.14, 0.54, 0.13, 0.18]. Therefore, the cross-entropy loss function would be:

$$L = -\begin{bmatrix}(0 * \log(0.14)) + (0 * \log(0.54)) + (1 * \log(0.13)) + \\ (0 * \log(0.18))\end{bmatrix} \qquad (\mathbf{2.3})$$

(E 2.3) result will be 2.04, it is relatively high, indicating that the predicted distribution does not match the actual distribution well. A smaller cross-entropy loss suggests a better fit between the expected and actual distributions. By minimizing this difference, the model can learn to generate molecules that are similar to the ones in the training set.

2.7.4. Optimizer

The optimizer adjusts the model parameters during training depending on the gradients of the loss function. In our model, we employed the Adam optimizer to fine-tune the model's parameters [56] and improve the performance. It minimizes the cross-entropy loss function between the prediction and the ground-truth value. It is a variant of stochastic gradient descent that adapts the learning rate for each parameter. It calculates the gradient of the loss function with respect to each parameter and uses this gradient to update the values of the parameters in the direction that minimizes the loss. The adaptive learning rate ensures that the model converges faster and more reliably to the optimal solution.

2.8. Training Strategy

In this study, we trained the defined model with the previously mentioned datasets following a novel strategy (**Figure 2.7**) for the purpose of guiding the generation process to focus on the target (**1.2.6**). It consists of three rounds. Firstly, the model was wormed

up with the "14K drug" dataset, followed by the protease inhibitors dataset in the second round, and finally with the HIV-1 protease inhibitors dataset in the third round. Each time, we decreased the learning rate and assessed the model's performance to ensure its accuracy was improving. In this way, we were able to improve the quality of the generated molecules by biasing the generation to focus on the last added dataset and generate new active molecules against the HIV-1 protease, and quite similar to those in the last training sets.

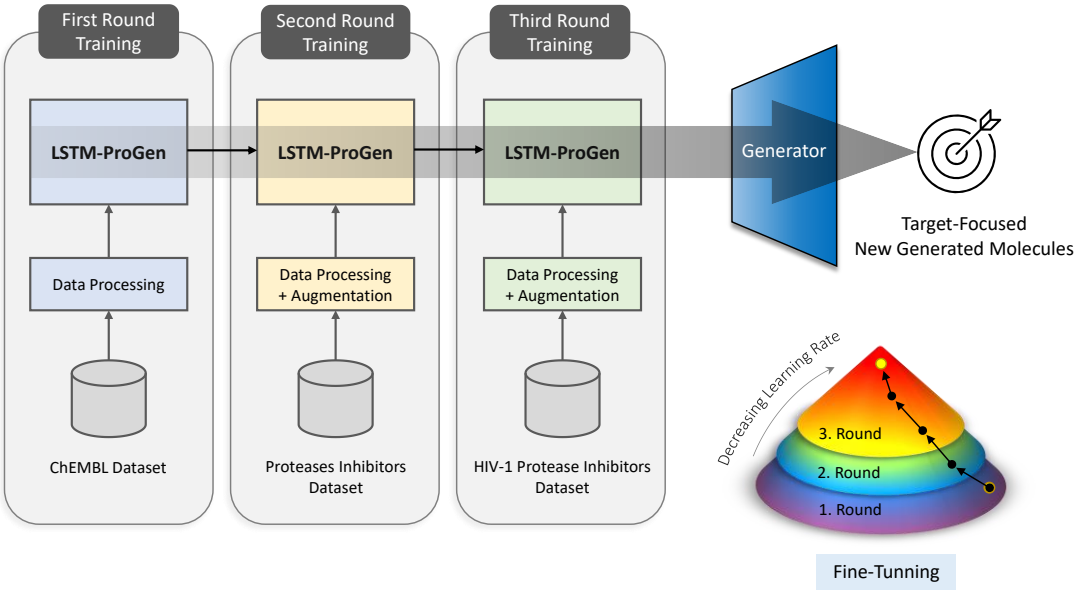


Figure 2.7: Overview of the training strategy we followed to train our model; It includes three rounds: the first round with the ChEMBL 14K drug dataset, the second round with the protease inhibitors dataset, and finally the third round with the HIV-1 protease inhibitors. At the right bottom corner, the figure simply showed the employed fine-tuning process to guide the generation process; in each step, the learning rate decreased to make the molecules focus on the target and generate similar molecules to those in the training sets. It starts in purple, where the model learns to generate drug-like molecules in general, then narrows up to focus only on protease inhibitors, and finally, in the last round, in red, focuses on HIV-1 protease inhibitors.

2.9. Molecule Filtering

The following processes are used to identify the best de novo new candidate molecules produced by the LSTM-ProGen model for targeting the HIV-1 protease: 1) Based on the LogP, molecular weight, and ring types collected from the analysis of the HIV-1 protease

known inhibitors dataset (**2.3**). Compounds that don't share similar features are eliminated. 2) based on the calculated Tanimoto similarity scores (estimated using Morgan fingerprints) between the new generated molecules and those in the original set of HIV-1 protease inhibitors, the top 10 molecules were selected with the greatest Tanimoto similarity scores.

2.10. Molecular Docking

Molecular docking is a computational technique that predicts the binding orientation and affinity of a small molecule ligand to a protein target. It is a crucial step in drug discovery and development, as it can help identify potential lead compounds that can bind to the target protein with high affinity and specificity. The molecules that remain after the filtering operations were analyzed via molecular docking. First the molecule library was created using Open-Babel software [57]. For the docking study, the crystal structures of HIV-1 protease (PDB code “4HLA” [58] and PDB code “3S45” [59]) were prepared by Discovery Studio 2021 software [60]. Missing hydrogen atoms were added, and water molecules were removed. The physical condition of pH was set as 7.4 ± 1.0 for atom typing. The binding site coordinates of 4HLA were defined as center_x = 8.16, center_y = -16.04 and center_z = 0.010, and 3S45 were defined as center_x = 20.51, center_y = 11.30, center_z = 18.08, by detection with same software and cross checked with the data published in the literature. PyRx software [61] was used to find the best binding poses for each ligand with in both used structures. It is a free and open-source software package, combines several open-source software, including AutoDock [62], AutoDock Vina [63] and Open-Babel [57] that provides a user-friendly graphical interface for performing molecular docking studies of large compound libraries against protein targets. After The AMBER force field [64] and AutoDock Vina [63] search algorithm utilized in PyRx for the docking simulation, the results were analyzed and visualized with PyMOL [65] and Discovery Studio software [60].

In summary, this chapter outlines the methodology employed to achieve the study's goal. A comprehensive overview of the process, including the utilization of the ChEMBL database for input molecules and the extraction of relevant sub-datasets, and the idea behind the exploration of the HIV protease known inhibitor dataset, revealed important features for further analysis. Additionally, a data augmentation method was applied to enhance dataset diversity and improve the model's performance. The data processing

technique was thoroughly explained, ensuring transparency and reproducibility. The chapter also discussed the model structure and training strategy. A filtration process was performed on the resulting generated molecules to minimize their size, select the most relevant molecules with high hopes of having inhibitory activity against the target, and decrease the computational cost that the molecular docking study requires. All these elements together enhance the credibility and impact of the study's findings.

CHAPTER 3

3. EXPERIMENTAL PART

This chapter presents an enthralling overview of the resources used to implement the LSTM-ProGen models, with each resource weaved into the fabric of our execution. The LSTM-ProGen models unfold as a testament to the boundless potential of this new field, opening new doors for discovering the unknown.

3.1. Test Environment

In this work, the datasets were processed using the Tesla K80 GPU, which Google Colab provided with 12 GB of GDDR5 RAM. LSTM-ProGen proposed molecular models were trained using this GPU to ensure optimal performance without the need for costly hardware resources. It meets the computing needs of the training and data processing tasks. When the training was completed, the models were stored on disk for future use. Once the training operation was completed, the last round of training was saved on the disk. The generator function executed on the HP ProBook 455 G4 computer, which has an AMD A10-9600P RADEON R5 processor with 10 computer cores, 8.00 GB of installed RAM, and a 64-bit operating system running Windows 10 Pro (20H2).

In this work, a harmonic ensemble of libraries gathered within the execution of our models, by leveraging the capabilities of Pandas, NumPy, Selfies, Scikit-learn, Seaborn, RDkit, Matplotlib, and TensorFlow libraries. These libraries acted as trusted companions, offering practical solutions, and improving our workflow, propelling us closer to our goals. The model architecture was defined using the sequential model type provided by TensorFlow library [66], giving optimum training flexibility. Each training iteration used the well-known Adam optimizer and consisted of 64 batches and 20 epochs. A molecular docking study was performed on the final filtered molecular library. The docking study employed the same gear as previously described. PyRx software was the main software

utilized for this task together with different software such as Discovery Studio and OpenBabel to achieve the goal. PyMOL software [65] was also used to visualize the results.

3.2. Testing Plan

For further examination, we tested two models with a slight variation in the data processing part for the purpose of identifying the most effective strategy for the best performance of our model. The two models have the same multi-layer architecture and are gradually trained using the same three-round technique as described before (**Table 2.1**). Through the implementation of these two models (Alpha and Beta), we highlight the significant impact of specific factors on the model outcomes. We evaluated the significance of maintaining a fixed maximum sequence length that we padded all the sequences to as a data processing part and a common token dictionary used to turn the molecules into vectors.

Table 3.1: Comparison of two LSTM-based models implemented for molecular generation: improving performance and stability.

Model Specifications		Model 1			Model 2		
		1.Round	2.Round	3.Round	1.Round	2.Round	3.Round
Data Processing	Number of input molecules	9480	3819	2600	9480	3147	1717
	Number of unique characters	84	57	56	86	86	86
	Max length	95	162	108	100	100	100
Model Parameters	Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
	Learning rate	0.005	0.005	0.003	0.005	0.004	0.003
	Epochs	20	20	20	20	30	30
	Batch size	64	64	64	64	64	64

The varying lengths used in the first model throughout each training cycle increased the complexity of the molecule production process. Consequently, the quality of the molecules produced decreased, and many of them proved to be invalid.

The Beta model ensured a more regulated and organized generating process by enforcing a set maximum sequence length. This limitation allowed for more control over the output, boosting the possibility of regularly creating legitimate and high-quality molecules. Furthermore, the constant token dictionary employed throughout all training rounds provided a consistent foundation for the model's knowledge and representation of molecules. As a result of these modifications, the second model greatly outperformed the first model. The stated maximum sequence length and constant token dictionary made molecule formation more reliable and genuine. This circumvented the constraints

indicated in the original model, allowing for the synthesis of higher-quality and more valuable molecules.

Moreover, the performance of LSTM-ProGen models in designing de novo molecules was evaluated using well-known metrics with conducting a comparative analysis to address the limitation of each proposed model and indicate the impact of the length and number of tokens factors on the learning process. For better understanding of the chemical space occupied by the de novo molecules generated by Beta model that was considered as baseline model for this study, we used the common powerful technique of principal components analysis (PCA) in conjunction with 2D and 3D visualization. We were able to translate the high-dimensional molecular data into a lower-dimensional representation while keeping the fundamental structural properties by using PCA. Because of the decrease in dimensionality, we were able to explore and understand the complicated chemical space in a more intuitive and interpretable manner. In the 2D visualization, the created de novo molecules were projected onto a two-dimensional plane based on the dataset's major components. Each point on the diagram represented a distinct molecule, with its position representing the overall structural features of the molecule. We learned about the similarity and variety of the created compounds in respect to known chemical compounds by comparing the distribution and clustering patterns of the de novo molecules with those of training molecules. We also expanded the technique to 3D visualization to improve our grasp of the chemical space. We acquired a more thorough depiction of the structural changes and connections between the molecules by integrating an extra dimension. The resulting 3D plot showed the de novo compounds in a visually appealing manner, allowing for a more in-depth investigation of their structural variety and possible chemical attributes. We were able to show areas inside the chemical space filled by the de novo molecules by rotating and studying the 3D map from various perspectives. This comparison of de novo molecules with real molecules allowed us to evaluate the novelty and differentiation of the generated compounds. We were able to determine if the produced molecules located in the same spots known for compounds or explored uncommon chemical space. This evidence provides important insights into the possibility of developing novel chemical structures and providing prospective leads for drugs research by the proposed approach of LSTM-ProGen. Therefore, the combination of PCA-based 2D and 3D visualization methods produced a visually appealing and educational representation of the chemical space explored by de novo molecules. It

allowed us to investigate their distribution, diversity, and likeness to real molecules, giving valuable information about the quality and potential applications of the generated compounds. In addition to examining the chemical space of the de novo molecules, we used a molecular docking experiment to ensure their target-specific activity. This method revealed important information about the possible interactions and binding affinities of the generated molecules by LSTM-ProGen with our specific target proteins. Molecular docking was used to simulate the binding process between the de novo molecules and the HIV-1 protease target proteins. The compounds were rated for binding affinity, which indicates their potential activity and efficacy as target protein ligands. Furthermore, two structures were employed in the investigation, and they are available on the Protein Data Bank (PDB) to guarantee the validity of the findings. We were able to discover promising compounds with high binding affinities and strong interactions with the target proteins through this experiment investigation. These molecules might benefit from further experimental validation or optimization in drug development efforts. By integrating the results of the molecular docking analysis with the chemical space exploration, we were able to obtain a complete picture of the potential biological activity of the generated compounds. We were able to prioritize and choose compounds with desirable structural features as well as favorable binding behaviors by benefiting the combination of computational predictions and visualization-based research. The target-specific activity ranking based on molecular docking added an important dimension to the evaluation of the de novo molecules. It bridged the gap between their structural features and likely biological activities, allowing for a more thorough evaluation of their overall drug-like properties and therapeutic implications.

As we wrap off this session, we look back on our rigorous testing method, which includes the evaluation of two versions of LSTM-ProGen. Minor adjustments were made in the data processing stage, which is critical to ensuring high-level performance. Our main objective was to discover the best-performing model for creating biased chemicals. In our quest for excellence, we did further examination the major impact of various methodologies. We explored the intricacies of various tokenization techniques and evaluated different sequence lengths, extensively assessing their influence on model performance. We were able to discover the ideal arrangement for moving us closer to our goal. However, our investigation did not end with a performance evaluation. We descended into the depths of chemical space, eager to ascertain the authenticity and

potential activity of the generated molecules. The principal component analysis (PCA) was performed in 2D and 3D visualizations way. Finally, we sought to validate their putative action by putting these de novo molecules in the intriguing realm of molecular docking. We were able to determine their ability to interact with and inhibit the target protein of our highest interest using this strategy. We used molecular docking to analyze the binding affinities and interactions of the produced compounds with the target protein, indicating their potential as promising candidates in our hunt for selective inhibition.

CHAPTER 4

4. RESULTS AND DISCUSSION

This section focuses on the outcomes of using the Alpha and Beta LSTM-ProGen models. The two models' performances were compared and well discussed for the task of generating new HIV-1 protease inhibitor candidates. Novel compounds were produced by the LSTM-ProGen architecture that had not been found before in the training dataset. Additionally, the results of the molecular docking experiments are well reported.

Before diving into the details of our discussion, let’s establish a solid foundation by understanding the primary objective of any generative model. These exceptional models employ their talent to create new samples that are comparable to the training instances while also recognizing the distribution of the hidden features in the training examples, sparking a creative symphony of possibilities (**1.2.4**). While undergoing the process of training, one way to evaluate the model’s performance is the training accuracy metric, which measures how well the model recreates the training data during the training process. As the model progresses in learning, it expected to improve its capability to sample the training data, resulting in a rising training accuracy curve, while the validation accuracy of a generative model is another one that often depends on how well the model generalizes untested data. **Figure 4.1** shows the full training progress over the three training rounds of the two variant models (Alpha and Beta).

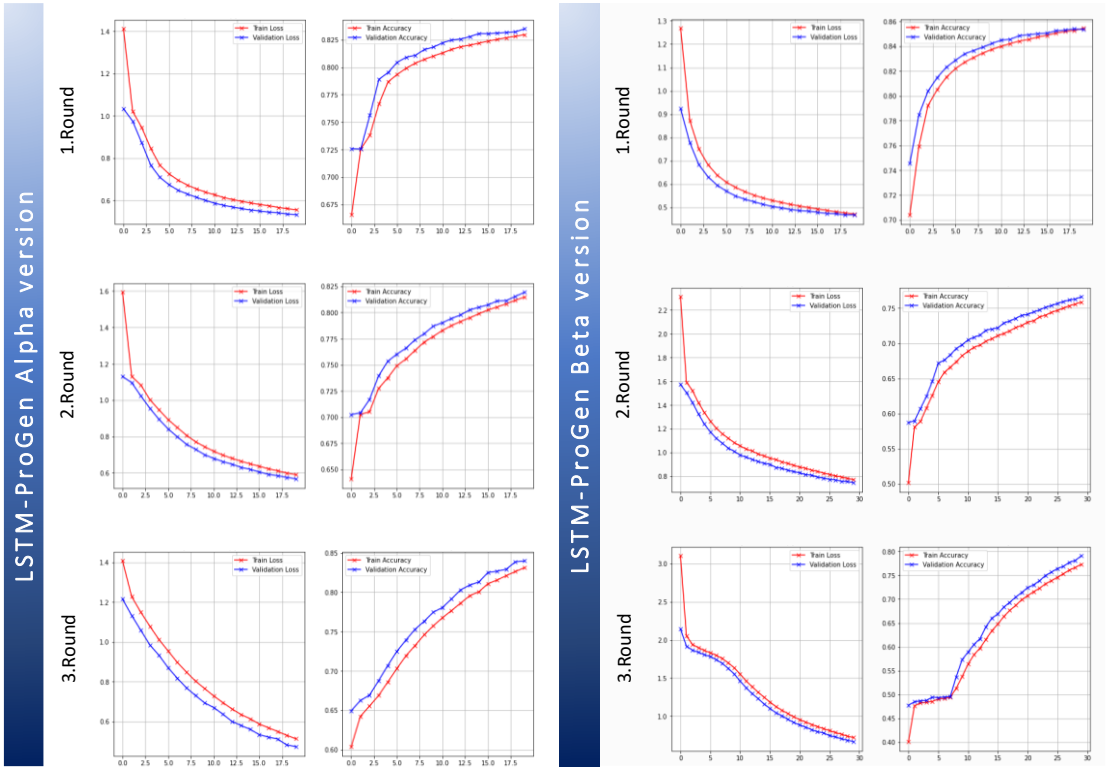


Figure 4.1: Performance improvement across three training rounds for both the Alpha and Beta LSTM-ProGen models. For each scenario, plots of accuracy and loss against the number of epochs demonstrated a consistent improvement in model performance.

The Alpha model demonstrated consistent, smooth training curves, displaying a relatively balanced relationship between loss and accuracy over all training iterations. Similarly, the Beta model achieved similar performance across the three training rounds, with a slight variation in the first round, indicating a sign of overfitting tendencies towards the end of training. It is presented by the intersection of the training and validation accuracy curves during the last epochs. It might have happened because the model started to focus on memorizing the training data or capturing specific features. It mostly suggests a potential likelihood for the model to become less effective at generalizing to unseen data. Though it is worth noting that this situation did not continue in the next two rounds of training, revealing that it was not a serious worry that required immediate action, an early stopping regularization procedure with a patience of 3 was applied in the subsequent rounds. This preventative step helps to keep the worst-case scenario and any prospective difficulties from negatively impacting performance during the training process.

Table 4.1: Comparative information about the difference between the two implemented LSTM-ProGen models.

	Model 1			**Model 2**		
	1.Round	**2.Round**	**3.Round**	**1.Round**	**2.Round**	**3.Round**
Accuracy	83.49%	81.92%	83.97%	85.34%	76.61%	79.07%
True positives (TP)	25630	16207	1557	25759	11259	8129
Ture negatives (TN)	103551	69661	16053	113046	24999	8459
False positives (FP)	459	343	0	405	124	46
False negatives (FN)	157	117	7	189	97	21

In general, both models Alpha and Beta, performed similarly across all training rounds, and the scores of performance metrics reported in **Table 4.1** and **Table 4.2** clearly show that both models scored quite close to one other. However, upon further examination of the generated molecules utilizing both models, despite the slightly inferior overall performance of the Beta model, it provided a notable capability to generate molecules with higher validity scores compared to the Alpha model. Training the Alpha model on different lengths with different numbers of unique characters for encoding molecules probably caused some confusion throughout the training process, resulting in low-validity molecules created using this model (**Figure 4.2**). Conversely, training the Beta model with a consistent length and the same set of unique characters for the encoding process during all the training rounds increased its ability to learn rules and generate valid molecules better than the Alpha model (**Table 4.1**). Based on this, these two elements are essential for guaranteeing stability throughout training and producing reliable molecules.

Table 4.2: Performance metrics.

	Accuracy	**Precision**	**Recall**	**F1-Score**
Alpah Version	0.999	1.0	0.995	0.997
Beta Version	0.995	0.994	0.997	0.995

Considering this significant observation, the Beta model was chosen as the benchmark model to forward the research into developing de novo compounds that target the HIV-1 protease. With the Beta model exhibiting exceptional promise and the ability to provide big outcomes. Furthermore, it is regarded as a beginning point for future advancements,

since it opens the door to additional work to generalize the used framework to be robust and effective in the development of molecular modeling.

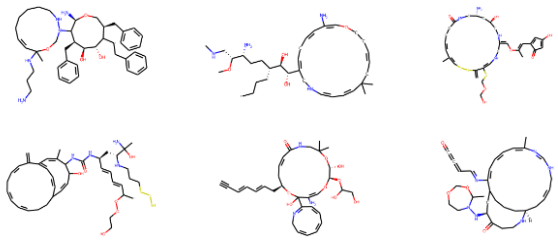


Figure 4.2: Examples of molecular structures that were sampled using the LSTM-ProGen Alpha model include some uncommon ring structures that were not included in the list of rings for all currently known drugs provided in [67].

Utilizing the Beta model, approximately 3000 novel molecules were generated, which were entirely distinct from those present in the training set. Scaffold variety measures the diversity of key molecular scaffolds in generated compounds [68]. It provides higher-level insights into the range of generated chemical structures. The 3000-molecule sample contained 1760 distinct core scaffolds. It is an important clue that the dataset's newly produced molecules have a wide range of diversity. In addition, the novelty score calculated the percentage of produced molecules that differed from the molecules in the training dataset. It evaluates the model's ability to create previously undiscovered chemicals. It was 100%, suggesting that all newly generated molecules are original and unique, with no duplicates or similarity with the training samples. Moreover, to gain insights about the physicochemical properties of these newly generated molecules, the density plots of three descriptors, namely molecular weight, octanol/water partition coefficients (AlogP), and polar surface area (PSA), were analyzed (**Figure 4.3**). The plots revealed significant similarities in the ranges of these descriptors among the generated molecules, HIV inhibitors, and the training dataset of ChEMBL molecules.

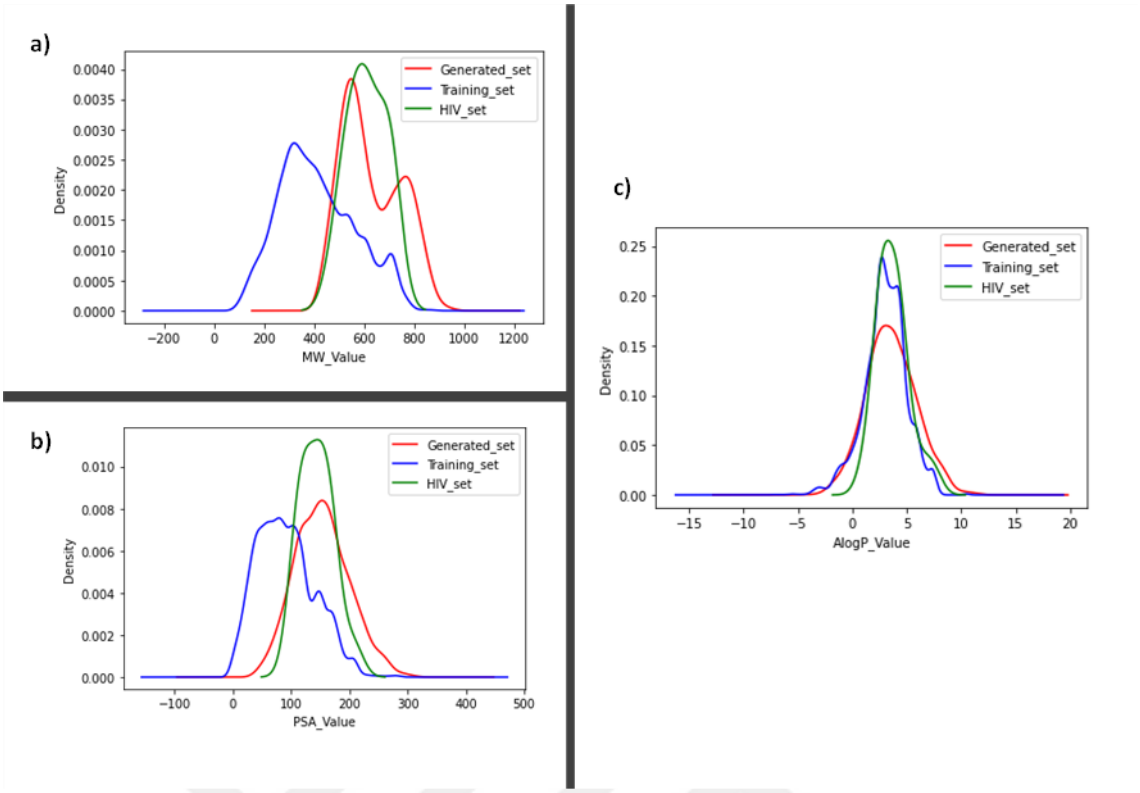


Figure 4.3: Kernal density similarity plots for comparative analysis of the distribution of three chemical descriptors, including molecular weight (MW), octanol-water partition coefficient (AlogP), and polar surface area (PSA), among the three datasets, training (blue), HIV inhibitors (green), and the generated molecules by LSTM-ProGen (red) datasets. **c)** highlights the similarity in the range of AlogP values across all three sets. **a)** displays the molecular weight range varies slightly between the mentioned datasets, but the newly generated set adequately covers the same ranges as the HIV inhibitors dataset. **b**) shows the polar surface area ranges are very similar, with only a slight upward trend observed in the HIV dataset.

Afterwards, filters derived from known HIV protease inhibitors (**2.3**) were applied to reduce the size of the generated set and focus on molecules with probably high activity against HIV-1 protease. This filtration process resulted in a subset of 360 molecules, (**Table 4.3**) includes some samples from the filtered dataset. To further investigate the chemical space covered by the de novo molecules, a principal component analysis (PCA) was conducted. **Figure 4.4** showcases the relationship between the entire generated molecules dataset, all the HIV inhibitors, and the entire training dataset of ChEMBL with the high spread diversity features. This analysis aimed to explore the proximity of the generated molecules to the HIV inhibitors dataset, which is of particular interest in this study. De novo molecules of LSTM-ProGen model and Chembl molecules are mostly spread out over the same space. The red points indicate the filtered dataset which is

expected to cover the near space that the real HIV inhibitors in yellow cover. This ended up with fair affinities in the molecular docking results as provided in the next section.

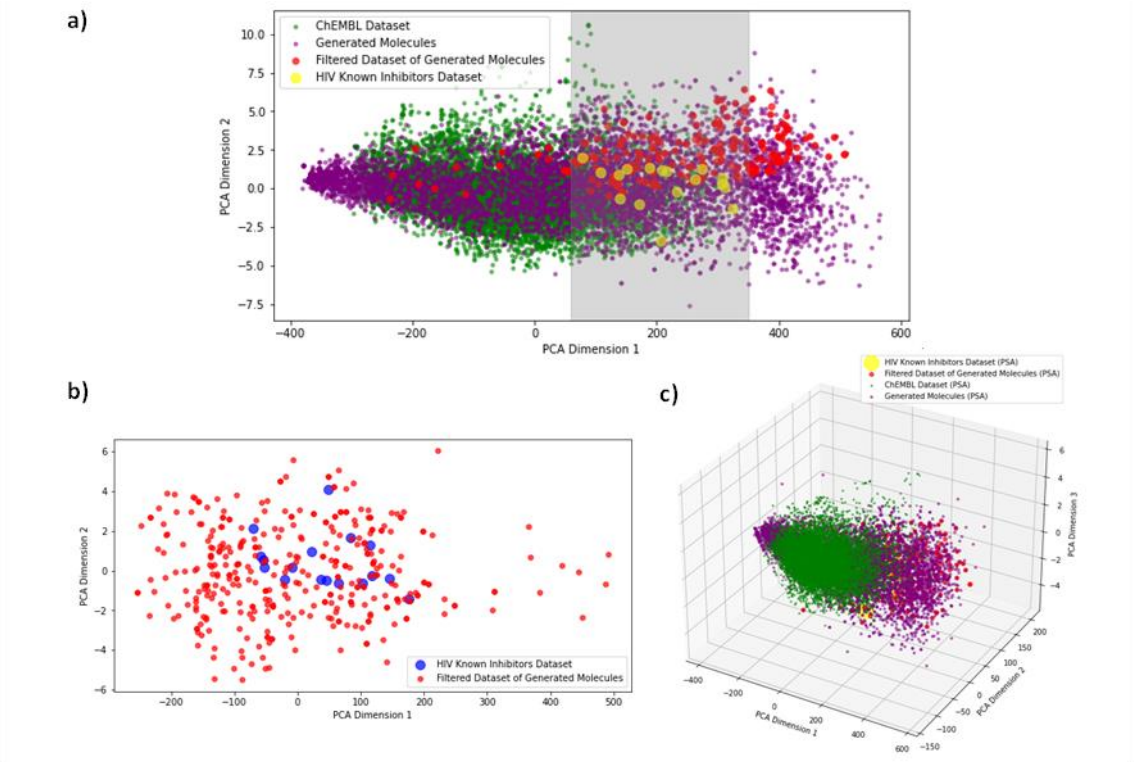


Figure 4.4: Mapping the chemical space through principal component analysis (PCA) among the LSTM-ProGen model-generated molecules, the training dataset, and the HIV inhibitors dataset. **a)** demonstrates physicochemical properties such as the molecular weight and octanol-water partition coefficient (LogP) in three datasets. **b)** HIV inhibitors and the new filtered molecules aligned. **c)** Beyond 2D: A Multidimensional Illustration of Molecular Weight, Octanol-Water Partition Coefficient (LogP), and Polar Surface Area (PSA) in the three datasets.

This reduction of the number of samples was necessary to minimize resources and the computational cost associated with the subsequent molecular docking study. Through a thoughtfully curated expert-led selection process, a subset of 76 compounds was randomly chosen from the pool of filtered de novo molecules. In this way we ensure the high ability of being active and lowering the computational cost required for the docking study.

Table 4.3: This is a list of randomly selected compounds from the collection of newly generated active molecules against the HIV-1 protease by LSTM-ProGen Beta, together with information about their molecular features.

Molecular ID	Molecular Weight	PSA	LogP
No302	129.09	30.82	1.2901
No289	231.319	47.28	2.53059
No224	234.226	39.44	3.3148
No191	268.312	49.69	3.9365
No178	300.402	75.68	2.4845
No137	344.431	90.61	3.63756
No155	360.433	86.08	3.72928
No193	382.528	93.53	3.49689
No88	413.474	101.08	1.59017
No237	413.498	101.99	1.65726
No158	416.522	114.88	0.37925
No218	441.488	60.98	4.28095

4.1. Molecular Docking Results

The molecules that remain after the filtering operations were tested via molecular docking. Strong evidence was determined to support the significant potential of several of our tested de novo compounds to be active against HIV-1 protease with inhibitory activities. These findings are an essential first step in unraveling the complexity of drug-target interactions and paving the way for future therapeutic development advances. The preparation schema of the used crystal structures for HIV-1 protease (PDB 4HLA and 3S45) and the new molecules included in (**2.10**).

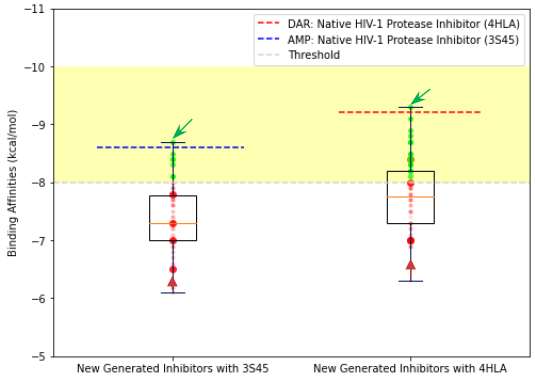


Figure 4.5: Score box plots displaying the binding free energies measured in the docking analysis of dene novo molecules generated by LSTM-ProGen model with two different HIV-1 protease structure (4HLA [58] and 3S45 [59]) provided on protein data bank. The scores of both the native ligand (Darunavir) in the utilized HIV protease complex structure shown by red dashed line, and another known inhibitor (Amprenavir) shown by blue dashed line. The green arrows indicate the new generated compounds with the potential to inhibit the HIV-1 protease.

In our application, a comprehensive cross-evaluation method was implemented to validate the docking process and ensure the reliability of our protocol. The goal was to evaluate not only the activity of our de novo molecules but also the efficacy of well-established and approved HIV protease inhibitors to confirm the validity of our protocol.

The docking results of both the known inhibitors and our novel molecules were compared. To initiate the cross-evaluation, a docking study was performed on the native HIV protease inhibitors, utilizing the structures observed in the Protein Data Bank (PDB). DAR was docked with the 4HLA structure, while AMP was docked with the 3S45 structure. The docking scores of DAR and AMP were carefully determined, considering their interactions, and fit within the binding regions of the target protein. Their scores were -9.2 and -8.6, respectively. These inhibitors served as reference points due to their well-documented affinities and binding abilities. Subsequently, our new compounds underwent the same docking technique, and their docking scores were compared to those of the native inhibitors mentioned above. The activity threshold was set at -8, indicating that any compounds scoring lower than this value had the potential to significantly inhibit the target protein. The docking scores are presented in **Table 4.4**. By ensuring that several of tested de novo molecules achieved docking scores comparable to or higher than those of DAR and AMP in case of Mol 92 with 4HLA and No 145 with 3S45. In this way, the credibility and efficacy of our protocol were established. This cross-evaluation method enabled the validation of the docking process and the verification that our molecules

exhibited similar binding affinities and interactions within the active site of the target protein. **Figure 4.5** shows the de novo molecules that were able to achieve affinity scores smaller than -8, highlighted in green. Furthermore, two of them surpassed the native inhibitors of the two utilized structures, DAR in 4HLA and AMP in 3S45. The results were documented in **Table 4.4**, which highlights the molecules that exhibited activity in both cases in the same color.

Table 4.4: The new generated molecules with the highest binding affinity scores in both structures of HIV-1 protease (4HLA and 3S45). The highlighted molecules in matching colors demonstrate potent inhibition activity against both utilized HIV-1 protease structures.

4HLA PDB Structure		**3S45 PDB Structure**	
Ligand ID	**Binding Affinity**	**Ligand ID**	**Binding Affinity**
Mol 92	-9.3	No 145	-8.7
Mol 154	-9.1	No 42	-8.5
Mol 42	-8.9	No 127	-8.4
Mol 127	-8.8	No 285	-8.4
Mol 145	-8.7	No 245	-8.3
Mol 195	-8.7	No 18	-8.1
Mol 211	-8.7	No 195	-8.1
Mol 181	-8.5	No 154	-8
Mol 245	-8.5	No 295	-8

For the visualization and analysis of the resulting complexes, advanced molecular visualization tools such as PyMOL were employed. The three-dimensional structures of the complexes were rendered, enabling the examination of the spatial arrangement and interactions between the molecules and the target protein (**Figure 4.6**). Additionally, Discovery Studio, a comprehensive software suite, was utilized to explore the interactions into the complexes (**Figure 4.7**). This powerful tool facilitated a thorough investigation of the properties and characteristics of the docking complexes, allowing for a deeper understanding of the molecular interactions, and aiding in the identification of key binding residues. Through this meticulous cross-evaluation method, the reliability of our docking protocol was validated by confirming that our novel molecules exhibited comparable affinities to the well-established and approved inhibitors. This approach not only strengthened the credibility of our results but also provided valuable insights for the identification and development of potential HIV protease inhibitors with promising activity against the target protein.

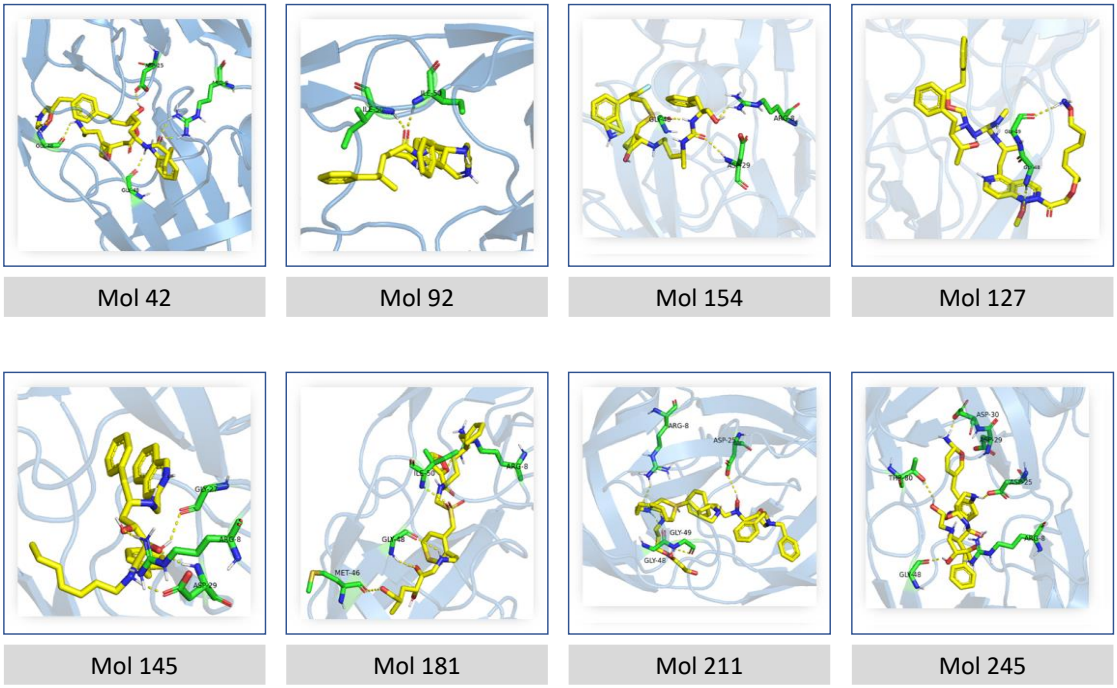


Figure 4.6: HIV-1 protease complex structure with the Darunavir inhibitor (PDB id: "4HLA" [69]); visualization of the best predicated poses of the molecular docking study of the de novo generated inhibitors of HIV-1 protease in its active site. The PyMOL program was used to create these figures [65].

To sum up, a set of 76 molecules were selected from the filtered set to test their inhibitory activity against the target protein through molecular docking studies. A final set of 12 molecules demonstrated significant potential as inhibitors for HIV-1 protease, thus offering potential avenues for the treatment of AIDS/HIV. These selected molecules exhibited favorable binding affinity scores, indicating their ability to interact effectively with the target protein and potentially disrupt its function. However, it is important to note that out of the initial pool of 76 molecules, only 12 showed inhibitory activity in the docking experiments. This observation highlights the need for further optimization and refinement of the selected molecules, as well as improvements to the proposed generation model. Enhancements in these areas can lead to the development of more effective inhibitors with enhanced binding affinities and improved therapeutic potential. The identification of these 12 molecules with promising inhibitory activity through molecular docking represents an important step in the drug discovery process for HIV-1 protease inhibitors. It is important to acknowledge that this study is based on a computational approach; therefore, the feasibility of synthesizing the suggested molecules has not been examined. Additional experimental validation and evaluation are crucial to validate their

inhibitory activity and assess their potential for further development as therapeutic agents in the battle against AIDS/HIV.

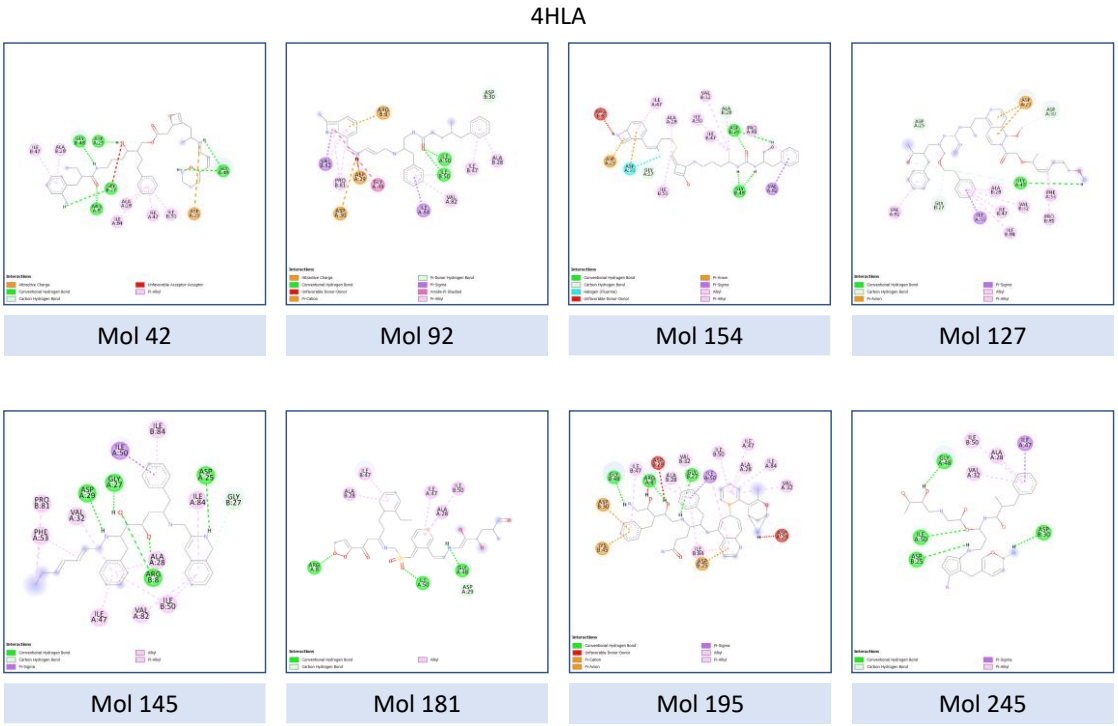


Figure 4.7: Analysis of docking study results using Discovery Studio software [60], revealing crucial receptor-ligand binding interactions.

CHAPTER 5

5. CONCLUSIONS AND FUTURE WORK

5.1. Conclusions

LSTM-ProGen is an automated approach for designing target-specific therapeutic candidate molecules. Combining the power of LSTM-based modeling with the robustness of Selfies molecular language, our model presented a remarkable performance in producing HIV-1 protease inhibitor candidates, demonstrating the efficacy of employing LSTM in this endeavor.

One of LSTM-ProGen's primary strengths is its high generation efficiency and capacity. The model was able to generate de novo compounds with similar physicochemical properties to known HIV-1 protease protein inhibitors by efficiently learning from different datasets. Various physicochemical metrics, such as polar surface area (PSA), molecular weight, and the logarithm of the partition coefficient (logP), revealed that the new molecules were within the same drug-likeness limits as the known HIV-1 protease inhibitors. Further computational studies were conducted to assess the target-specific features of the de novo molecules generated by LSTM-ProGen against the target protein.

The results indicate a high potential for targeting HIV-1 protease, a key enzyme in the HIV replication cycle. Through the molecular docking study, the de novo compounds exhibited binding affinities and interactions comparable to current HIV-1 protease inhibitors, suggesting their potential to be optimized to reduce viral replication.

The findings show LSTM-ProGen’s ability to rapidly explore chemical space and design novel molecules with desirable properties for specific protein targets. The model's ability to generate molecules with properties similar to those of known inhibitors provides a chance for faster drug discovery and development processes. Although LSTM-ProGen faced difficulties in effectively capturing the rules governing ring systems, the LSTM-

ProGen approach remains a valuable tool for streamlining the discovery and development of new drugs for druggable targets, as it helps to reduce the time-consuming and costly aspects associated with traditional hit-to-lead optimization.

While LSTM-ProGen demonstrated promising results, additional experimental validation tests are required to confirm the efficacy and safety of the presented molecules. In vitro and in vivo studies, as well as rigorous toxicity assessments, are critical steps in determining the viability of these de novo molecules as drug candidates.

5.2. Future Work

The advancements in deep learning (DL) and machine learning (ML) approaches have made a significant impact on the field of drug discovery. These approaches have shown promising results with the goal of speeding up the identification and design of novel drug candidates. LSTM-ProGen stands as evidence of the importance of DL and ML in this domain. However, certain areas require more research and improvement to continue to enhance the outcomes of such models. One of the most important critical challenges that must be addressed in future work is the development of a technique to improve learning during the training process to enable the production of molecules with ring structures similar to those found in approved drugs to date. LSTM-ProGen had suffered from the learning of the ring systems, and this is revealed in the resulting molecules. One potential option could be to train LSTM-ProGen on larger and more diverse datasets. By exposing the model to a wide range of molecular structures, including those with ring systems, it may learn to generate molecules more closely resembling approved drugs. However, this approach may necessitate significant computational resources as well as careful curation of the training data. Another possible choice is fragment-based training, which provides a novel and innovative solution to the problem of generating molecules with the correct ring systems. We can capture the rich structural diversity found in existing drugs and medicinal compounds by training a separate model on a dataset consisting of diverse ring fragments. This fragment-focused model learns the intricate relationships between various fragments, such as connectivity and spatial arrangements. As a result, it gains a thorough understanding of the essential characteristics and properties of ring structures, enhancing its ability to generate molecules with appropriate ring structures which are similar to those in the approved drugs to date. Afterwards, we can transfer learning to the main model implemented for a certain subject; following this way, we can ensure to

generate molecules with the desired properties for the subject of the study with proper ring structures. In addition to these structural considerations, and leaving aside the computational cost, reinforcement learning can be a valuable tool for guiding the generation process. Through reinforcement learning, the model can adapt its generation strategy based on the feedback received from the evaluation process. This method allows the model to prioritize specific properties during the generation process, such as efficacy, safety, or target specificity, resulting in the discovery of molecules with improved pharmacological profiles.

To sum up, DL and ML hold great promise for revolutionizing drug discovery and bring tremendous hope in positively impacting patients' lives. However, further advancements are necessary to fill the existing gaps and accelerate this process. By addressing the challenges associated with generating molecules, we can enhance the performance and utility of the AI-based models in the field. This pursuit of improvement is essential to our collective mission to save lives, support medicine, and enhance the well-being of patients worldwide.

BIBLIOGRAPHY

[1] R. A. Weiss, "How does HIV cause AIDS?," *Science (New York, N.Y.),* vol. 260, no. 5112, pp. 1273-1279, 1993.

[2] P. Piot, M. Bartos, P. D. Ghys, N. Walker and B. Schwartländer, "The global impact of HIV/AIDS," *Nature,* vol. 410, no. 6831, pp. 968-973, 2001.

[3] J. U. N. P. o. H. AIDS, "UNAIDS Global AIDS Update 2022," Joint United Nations Programme on HIV/ AIDS, Geneva, 2022.

[4] G. M. Shaw and E. Hunter, "HIV transmission," *Cold Spring Harbor perspectives in medicine,* vol. 2, no. 11, pp. a006965-a006965, 2012.

[5] H. Yang, J. Nkeze and R. Y. Zhao, "Effects of HIV-1 protease on cellular functions and their potential applications in antiretroviral therapy," *Cell & bioscience,* vol. 2, no. 1, p. 32, 2012.

[6] S. ‘. o. P. T. b. B. {German Advisory Committee Blood (Arbeitskreis Blut), "Human immunodeficiency virus (HIV)," *Transfusion medicine and hemotherapy,* vol. 43, no. 3, pp. 203-222, 2016.

[7] A. Brik and C.-H. Wong, "HIV-1 protease: mechanism and drug discovery," *Organic & biomolecular chemistry,* vol. 1, no. 1, pp. 5-14, 2003.

[8] R. P. D. Bank, "Rcsb.org, 4LL3," RCSB Protein Data Bank, [Online]. Available: https://www.rcsb.org/structure/4LL3.

[9] T. Robins and J. Plattner, "HIV protease inhibitors: their anti-HIV activity and potential role in treatment," *Journal of acquired immune deficiency syndromes,* vol. 6, no. 2, pp. 162-170, 1993.

[10] D. R. Bangsberg, D. L. Kroetz and S. G. Deeks, "Adherence-resistance relationships to combination HIV antiretroviral therapy," *Current HIV/AIDS reports,* vol. 4, no. 2, pp. 65-72, 2007.

[11] H. C. S. Chan, H. Shan, T. Dahoun, H. Vogel and S. Yuan, "Advancing drug discovery via artificial intelligence," *Trends in pharmacological sciences,* vol. 40, no. 8, pp. 592-604, 2019.

[12] A. Mullard, "New drugs cost US$2.6 billion to develop," *Nature reviews. Drug discovery,* vol. 13, no. 12, pp. 877-877, 2014.

[13] P. G. Polishchuk, T. I. Madzhidov and A. Varnek, "Estimation of the size of drug-like chemical space based on GDB-17 data," *Journal of computer-aided molecular design,* vol. 27, no. 8, pp. 675-679, 2013.

[14] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka, R. F. Lameiro, D. Lemm, A. Lo, S. M. Moosavi and Nápoles-, "SELFIES and the future of molecular string representations," *Patterns (New York, N.Y.),* vol. 3, no. 10, p. 100588, 2022.

[15] F. Grisoni, M. Moret, R. Lingwood and G. Schneider, "Bidirectional molecule generation with recurrent neural networks," *Journal of chemical information and modeling,* vol. 60, no. 3, pp. 1175-1183, 2020.

[16] P. Ertl, R. Lewis, E. Martin and V. Polyakov, "In silico generation of novel, drug-like chemical matter using the LSTM neural network," *arXiv [cs.LG]*, 2017.

[17] C. Dirican, "The impacts of robotics, artificial intelligence on business and economics," *Procedia, social and behavioral sciences,* vol. 195, pp. 564-573, 2015.

[18] P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, {. Press}, A. Saxenian, J. Shah and M. Tambe, "Artificial Intelligence and Life in 2030: The One Hundred Year Study on Artificial Intelligence," *arXiv [cs.CY],* 2022.

[19] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[20] B. B. Traore, B. Kamsu-Foguem and F. Tangara, "Deep convolution neural network for image recognition," *Ecological informatics,* vol. 48, pp. 257-268, 2018.

[21] F. A. Gers and E. Schmidhuber, "LSTM recurrent networks learn simple context-free and context-sensitive languages," *IEEE transactions on neural networks,* vol. 12, no. 6, pp. 1333-1340, 2001.

[22] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE access: practical innovations, open solutions,* vol. 7, pp. 19143-19165, 2019.

[23] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of chemical information and computer sciences,* vol. 28, no. 1, pp. 31-36, 1988.

[24] M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation," *Machine learning: science and technology,* vol. 1, no. 4, p. 045024, 2020.

[25] E. J. Bjerrum and B. Sattarov, "Improving chemical autoencoder latent space and molecular DE Novo generation diversity with heteroencoders," *Biomolecules,* vol. 8, no. 4, p. 131, 2018.

[26] N. De Cao and T. Kipf, "MolGAN: An implicit generative model for small molecular graphs," *arXiv [stat.ML],* 2018.

[27] W. Jin, R. Barzilay and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," *arXiv [cs.LG],* 2018.

[28] H. L. Morgan, "The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service," *Journal of chemical documentation,* vol. 5, no. 2, pp. 107-113, 1965.

[29] B. C. Love, "Comparing supervised and unsupervised category learning," *Psychonomic bulletin & review,* vol. 9, no. 4, pp. 829-835, 2002.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial networks," *Communications of the ACM,* vol. 63, no. 11, pp. 139-144, 2020.

[31] O. Prykhodko, S. V. Johansson, P.-C. Kotsias, J. Arús-Pous, E. J. Bjerrum, O. Engkvist and H. Chen, "A de novo molecular generation method using latent vector based generative adversarial network," *Journal of cheminformatics,* vol. 11, no. 1, p. 74, 2019.

[32] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning,* vol. 12, no. 4, pp. 307-392, 2019.

[33] L. Medsker and L. C. Jain, Recurrent Neural Networks: Design and Applications, Boca Raton, FL: CRC Press, 1999.

[34] Y. Zhao, X. Xia and R. Togneri, "Applications of deep learning to audio generation," *IEEE circuits and systems magazine,* vol. 19, no. 4, pp. 19-38, Fourthquarter 2019.

[35] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine,* vol. 35, no. 1, pp. 53-65, 2018.

[36] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International journal of uncertainty, fuzziness and knowledge-based systems,* vol. 06, no. 02, pp. 107-116, 1998.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[38] R. Dey and F. M. Salem, "Gate-variants of Gated Recurrent Unit (GRU) neural networks," *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS),* pp. 1597-1600, 2017.

[39] L. Yao and Y. Guan, "An improved LSTM structure for natural language processing," *IEEE,* pp. 565-569, 2018.

[40] M. K. Nammous and K. Saeed, "Natural language processing: Speaker, language, and gender identification with LSTM," in *Advances in Intelligent Systems and Computing*, Singapore, Springer Singapore, 2019, pp. 143-156.

[41] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider and G. Schneider, "Generative recurrent networks for DE Novo drug design," *Molecular informatics,* vol. 37, no. 1-2, 2018.

[42] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid and K. Alahari, "End-to-end incremental learning," *arXiv [cs.CV]*, pp. 233-248, 2018.

[43] I. A. Guedes, C. S. de Magalhães and L. E. Dardenne, "Receptor-ligand molecular docking," *Biophysical reviews,* vol. 6, no. 1, pp. 75-87, 2014.

[44] R. Jakhar, M. Dangi, A. Khichi and A. K. Chhillar, "Relevance of molecular docking studies in drug designing," *Current bioinformatics,* vol. 15, no. 4, pp. 270-278, 2020.

[45] M. Davies, M. Nowotka, G. Papadatos, N. Dedman, A. Gaulton, F. Atkinson, L. Bellis and J. P. Overington, "ChEMBL web services: streamlining access to drug discovery data and utilities," *Nucleic acids research,* vol. 43, no. W1, pp. W612-20, 2015.

[46] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson and Chi, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic acids research,* vol. 46, no. D1, pp. D1074-D1082, 2018.

[47] J.-L. Reymond and M. Awale, "Exploring chemical space for drug discovery using the chemical universe database," *ACS chemical neuroscience,* vol. 3, no. 9, pp. 649-657, 2012.

[48] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS central science,* vol. 4, no. 2, pp. 268-276, 2018.

[49] L. C. B. a. J.-L. Reymond, "970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe," *J. Am. Chem. Soc.,* vol. 131, p. 8732, 2009.

[50] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic acids research,* vol. 40, no. D1, pp. D1100-D1107, 2012.

[51] M. H. S. Segler, T. Kogej, C. Tyrchan and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS central science,* vol. 4, no. 1, pp. 120-131, 2018.

[52] E. J. Bjerrum, "SMILES Enumeration as Data Augmentation for Neural Network," arXiv [cs.LG], Denmark, 2017.

[53] G. Landrum, "Rdkit.org," [Online]. Available: https://www.rdkit.org/.

[54] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv [cs.CL],* 2013.

[55] S. Sharma, S. Sharma and A. Athaiya, "ACTIVATION FUNCTIONS IN NEURAL NETWORKS," *International Journal of Engineering Applied Sciences and Technology,* vol. 4, no. 12, pp. 310-316, 2017.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv [cs.LG],* 2014.

[57] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, "Open Babel: An open chemical toolbox," *Journal of cheminformatics,* vol. 3, no. 1, p. 33, 2011.

[58] {. P. D. Bank}, "Rcsb.org," [Online]. Available: https://www.rcsb.org/structure/4hla.

[59] "3S45," RCSB Protein Data Bank, [Online]. Available: https://www.rcsb.org/structure/3s45.

[60] ". D. S. D. SYSTÈMES, "Dassault Syst mes BIOVIA, Discovery Studio Modeling Environment, Release 2017," *Dassault Syst mes,* 2016.

[61] S. Dallakyan and A. J. Olson, "Small-molecule library screening by docking with PyRx," *Methods in molecular biology (Clifton, N.J.),* vol. 1263, pp. 243-250, 2015.

[62] G. M. Morris, R. Huey and A. J. Olson, "Using AutoDock for ligand-receptor docking," *Current protocols in bioinformatics,* vol. Chapter 8, no. 1, p. Unit 8.14, 2008.

[63] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of computational chemistry,* vol. 31, no. 2, pp. 455-461, 2010.

[64] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, "AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules," *Computer physics communications,* vol. 91, no. 1-3, pp. 1-41, 1995.

[65] L. Schrödinger, "The PyMOL Molecular Graphics System, Version 1.8," 2015. [Online].

[66] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia and Joze, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," *arXiv [cs.DC],* 2016.

[67] R. D. Taylor, M. MacCoss and A. D. G. Lawson, "Rings in drugs: Miniperspective," *Journal of medicinal chemistry,* vol. 57, no. 14, pp. 5845-5859, 2014.

[68] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. Molecular frameworks," *Journal of medicinal chemistry,* vol. 39, no. 15, pp. 2887-2893, 1996.

[69] R. G. H. P. I. D. D. W. P. G. A. M. H. Yedidi, "4HLA," RCSB Protein Data Bank, [Online]. Available: https://www.rcsb.org/structure/4hla.

[71] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias and A. Aspuru-Guzik, "Objective-reinforced Generative Adversarial Networks (ORGAN) for sequence generation models," *arXiv [stat.ML]*, 2017.

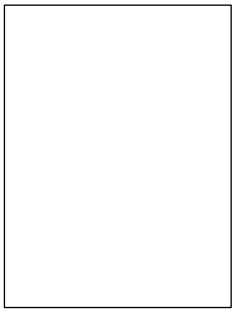
[72] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper and A. Zhavoronkov, "DruGAN: An advanced generative adversarial autoencoder model for de Novo generation of new molecules with desired molecular properties in silico," *Molecular pharmaceutics*, vol. 14, no. 9, pp. 3098-3104, 2017.

[73] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, "PubChem Substance and Compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202-13, 2016.

APPENDIX A

My code is publicly available at [https://github.com/taleb-hub/LSTM-ProGen], which also includes all the used datasets in this work.

CURRICULUM VITAE



Name Surname : M.Taleb ALBRIAJWI

Place and Date of Birth :

E-Mail :

EDUCATION:

B.Sc. : 2014, Aleppo University, Science, Biochemistry

PROFESSIONAL EXPERIENCE AND REWARDS:

- Fully funded scholarship for my MSc studies, Istanbul Medipol University

PUBLICATIONS, PRESENTATIONS, AND PATENTS ON THE THESIS:

- Albrijawi, M. T., Ibrahim, A. H., & Alhajj, R. (2021). "Predictions of drug metabolism pathways through CYP 3A4 enzyme by analysing drug-target interactions network graph". Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York, NY, USA: ACM.

AI For Drug Discovery LSTM-Driven Drug Design Using SELFIES for Target-Focused De Novo Generation of HIV-1 Protease Inhibitor Candidates in The Treatment of AIDS

ORIGINALITY REPORT

15%	12%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	sens.medipol.edu.tr Internet Source	1%
2	huggingface.co Internet Source	1%
3	arxiv.org Internet Source	<1%
4	dokumen.pub Internet Source	<1%
5	deepai.org Internet Source	<1%
6	bdb2.ucsd.edu Internet Source	<1%
7	iq.opengenus.org Internet Source	<1%
8	cellandbioscience.biomedcentral.com Internet Source	<1%