A Study of Generative AI Techniques for Accelerating Drug Discovery

By

Manuella Kristeva NAKAM YOPDUP (manuella.yopdup@aims.ac.rw)
African Institute for Mathematical Sciences (AIMS), Rwanda

Supervisor: Dr. rer. nat. habil. Abebe Geletu W. Selassie

Co-Supervisor: Dr. Eunice Gandote

AIMS, Rwanda

June 2025

AN ESSAY PRESENTED TO AIMS RWANDA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICAL SCIENCES

AIMS | African Institute for Mathematical Sciences RWANDA

DECLARATION

This work was carried out at AIMS Rwanda in partial fulfilment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at AIMS Rwanda or any other University.

Student: Manuella Kristeva NAKAM YOPDUP

Supervisor: Dr. rer. nat. habil. Abebe Geletu W. Selassie

Co-Supervisor: Dr. Eunice Gandote

ACKNOWLEDGEMENTS

I thank my Lord Jesus for his wisdom, guidance and advice throughout this journey. I am grateful to my supervisor, Dr. rer. nat. habil. Abebe Geletu W. Selassie and to my co-supervisor, Dr. Eunice Gandote, for their time, patience and encouragement. Without their support, this work could not have been completed. To my parents, sisters and friends for their motivation, encouragement and unwavering support. To the entire AIMS community, I appreciate the opportunity and the love.

DEDICATION

I dedicate this work to my father, Mr. Blaise Pascal NAKAM POKAM, who sacrifices himself every day to provide us the best.

Abstract

Generative Artificial Intelligence (Generative AI) is increasingly utilized across various fields, including commerce, finance, and healthcare. Specifically, several Generative AI models have been developed to facilitate drug discovery and expedite the drug development process. In this study, we employed a Generative AI model based on generative adversarial networks (GANs) and transformers, named DrugGEN. This model was trained and tested on a dataset comprising ofAfrican medicinal plants, which offers a promising foundation for discovering anticancer drugs. Our evaluation of the generated molecules revealed that the DrugGEN model successfully produces viable candidates, even without direct training on specific datasets. The observed diversity among the final candidate molecules is a strong indicator of their potential in drug discovery, suggesting that these compounds could lead to innovative therapeutic solutions.

Keywords : Generative AI, Drug Discovery, AfroCancer Database.

Résumé

L'intelligence artificielle générative (IA générative) est de plus en plus utilisée dans divers domaines, comme le commerce, la finance et la santé. Plusieurs modèles d'IA générative ont notamment été développés pour faciliter la découverte de médicaments et accélérer le processus de développement de ces derniers. Dans cette étude, nous avons utilisé DrugGEN, un modèle d'IA générative basé sur les réseaux adversaires génératifs (GAN) et les transformers. Ce modèle a été entraîné et testé sur un ensemble de données comprenant des plantes médicinales africaines(AfroCancer), ce qui offre une base prometteuse pour la découverte de médicaments anticancéreux. Notre évaluation des molécules générées a révélé que le modèle DrugGEN produit avec succès des molécules candidates viables. La diversité observée parmi les molécules candidates est un indicateur fort de leur potentiel dans la découverte de médicaments, suggérant que ces composés pourraient conduire à des solutions thérapeutiques innovantes.

Mots clés : IA générative, Développement de médicaments, Base de données AfroCancer.

Contents

1. Introduction

The integration of generative artificial intelligence (Gen AI) into drug discovery is a transformative shift into the pharmaceutical industry, particularly in light of the Covid-19 pandemic, which has profoundly impacted global health and resulted in significant loss of life [1]. This unfortunate situation has once again highlighted the need to adapt and rapidly create new medicines (Ward et al., 2021). The rapid spread of the virus is one of the causes that challenged traditional drug discovery and approval processes, revealing the crucial importance of speed and flexibility in responding to emerging health crises. However, the traditional pathway of drugs discovery is an expensive and time-consuming process, often marked by low success rate. This is due to the inherent complexity of biological systems, the challenges of identifying viable drug targets, and the high attrition rate during preclinical and clinical testing. Consequently, there is a pressing need to enhance production efficiency and minimize resource consumption in drug development. Artificial Intelligence (AI), particularly generative AI (Gen AI) appears to be a transformative solution. By integrating generative AI into various stages of drug discovery, researchers can reduce costs, increase efficiency, and ultimately bring new therapies to patients faster.

Another most complex and life-threatening disease which requires ongoing drug discovery is cancer which is a major global health problem. Its impact in Africa presents unique and complex challenges due to the fact that radiotherapy equipment, chemotherapy drugs and surgical oncology services are often concentrated in a few urban centers, or even non-existent. There is therefore an urgent need to make drugs accessible and to produce appropriate, effective cancer treatments. However, as most drugs used are imported from outside the continent, there is an urgent need for new medicines (Siqueira-Neto et al., 2023). As these drugs are imported, they are not always adapted to our situation, as they do not always optimally treat the unique genetic and environmental factors that influence disease prevalence. Also, genetic and environmental factors influence cancer prevalence and progression in African populations. Therefore, it becomes necessary for Africa to design its own medicines adapted to our context and conditions (Ferreira et al., 2022). Increasingly, African researchers are turning their attention to modernizing the drug discovery process. We have, for example, the African Natural Products Database (ANPDB) platform[2], which enables researchers to collect databases containing Natural Products (NPs) extracted from plants used in traditional African medicine, some of which have already been approved by the FDA, such as artemimisin, an antimalarial drug.

Therefore, this work aims to explore the potential of Generative Artificial Intelligence to revolutionize drug discovery especially in Africa. Our research question is how to use Generative AI to modernize the drug discovery process in Africa?

To answer this question, our approach is based on a designed model called DrugGEN, to effectively generate molecules against AKT1 also known as protein kinase B (PKB) and cyclin-dependent kinases (CDK) proteins, which are critically important in developing treatments for various types of cancer (Ünlü et al., 2023). For the experimentations, we will use the DrugGEN model on Afro-

[1] https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people%27s-livelihoods-their-health-and-our-food-systems(10/05/2025)

[2] https://african-compounds.org/anpdb/(10/05/2025)

Cancer database (Namba-Nzanguim et al., 2022) containing compounds from African medicinal plants that have shown in vitro and/or in vivo anticancer, cytotoxic, and antiproliferative.

1.1 Objective of the study

The objectives of this study are:

- to understand the concept of Generative Artificial Intelligence (Gen AI),
- to explore the mathematics behind Gen AI techniques,
- to study the current state of Drug Discovvery in Africa,
- to show how Gen AI can be used to generate drug against cancer based on African database.

1.2 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 presents all the preliminaries we need to know to understand the topic. Chapter 3 discusses the State of the Art of Drug discovery in Africa and using Generative AI and chapter 4 provides the Methodology used for our experimentations. Chapter 5 contains the results of our experimentations and discussion. Finally, chapter 6 is dedicated to the summary, future works and conclusions.

2. Preliminaries

In this section, we aim to provide the elements needed to understand the rest of the work. To do this, we first explore the drug discovery process and then explore Artificial Intelligence, focusing on Generative AI.

2.1 Drug Discovery

Drug discovery can be described as the process of obtaining new drugs by identifying chemical entities that have the potential to become therapeutic agents, with the goal to safely treat and manage disease. This process can take up to 12 years and cost more than billion because it follows several steps, each one important (Deore et al., 2019). The steps[1] are shown on Figure 2.1:

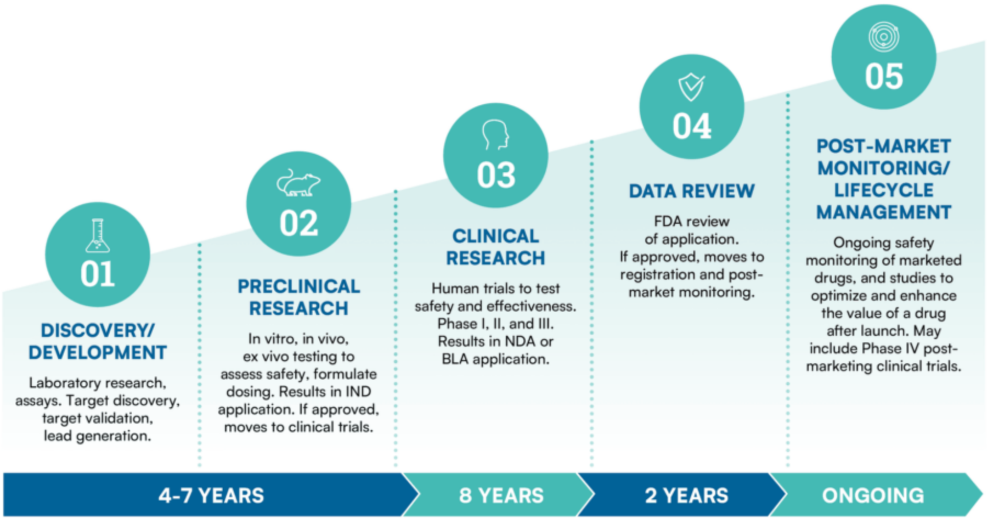
01
DISCOVERY/ DEVELOPMENT
Laboratory research, assays. Target discovery, target validation, lead generation.
02
PRECLINICAL RESEARCH
In vitro, in vivo, ex vivo testing to assess safety, formulate dosing. Results in IND application. If approved, moves to clinical trials.
03
CLINICAL RESEARCH
Human trials to test safety and effectiveness. Phase I, II, and III. Results in NDA or BLA application.
04
DATA REVIEW
FDA review of application. If approved, moves to registration and post-market monitoring.
05
POST-MARKET MONITORING/ LIFECYCLE MANAGEMENT
Ongoing safety monitoring of marketed drugs, and studies to optimize and enhance the value of a drug after launch. May include Phase IV post-marketing clinical trials.
4-7 YEARS
8 YEARS
2 YEARS
ONGOING

Figure 2.1: Stages in the drug development process Source

2.1.1 Discovery and development or De novo design. It is the first stage of the drug discovery process which involves identifying and validating the biological target involved in the disease. The goal is to find the biological origin of the disease and enumerate the possible interventions. Thus, the intervention target is selected based on many criteria like efficacity, safety and druggability (Deore et al., 2019). Following that, an expected molecule or drug candidate is validated function of the ability to reproduce or improve target specificity, selectivity, pharmacodynamic/pharmacokinetic parameters, and toxicological properties of existing solutions. This stage is long, complex, and expensive, typically taking 12-15 years and costing up to $2 billion.

[1] https://www.patheon.com/us/en/insights-resources/blog/drug-development-phases.html(15/05/2025)

2.1.2 Preclinical research. To ensure the drug's safety and efficacy, drugs are tested on animal species to predict the potential human outcomes. These trials assess the drug's pharmacokinetic (absorption, distribution, metabolism, and excretion) and pharmacodynamic parameters, which are essential for determining safety and efficacy, half-life, and distribution mechanisms. Also, toxicology studies are done to evaluate toxicological effects using in-vitro and in-vivo tests, such as assessing effects on cell proliferation or performing quantitative determination of toxicological effects (Deore et al., 2019). Regulatory authorities must approve preclinical trials, ensuring they are conducted safely and ethically. Data from these preclinical and toxicity studies are important in the drug development process.

2.1.3 Clinical research. After the preclinical step has been done and the safety and efficacy confirm, tests are now done on humans. Clinical research is typically divided into four phases: the first phase involves trials on a small group of healthy volunteers or individuals with the disease. The second phase consists of extending the trials to a group up to one hundred of patients to evaluate the drug's efficacy and gather additional safety data. The third one is conducted on a larger groups of patients to assess the efficacy and gather additional safety data (Deore et al., 2019). The last one provide most of the safety data, which are crucial for detecting less common, long-term, or uncommon side effects due to the large number of participants and longer duration.

2.1.4 Regulatory Review and Approval. If the results of the previous steps have been concluant, drug developers must submit a New Drug Application(NDA) including all preclinical and clinical data, labeling, safety updates, patent info, and usage directions. A team of experts including doctors, chemists, statisticians, microbiologists, and pharmacologists review the drug compound's safety and efficacy based on the NDA (Deore et al., 2019). If the result of the expertise are positive, the regulatory authorities (e.g., the FDA in the United States, EMA in Europe) can give approval to manufacture, market, and distribute the drug. Developers can appeal FDA decisions or choose further development upon request.

2.1.5 Post-Market Surveillance. Even after a drug or device has been approved by the FDA, it's possible that new concerns may arise in the general population after its approval. Several programs are created to allow healthcare professionals and consumers to report serious problems with medical products or drug's reaction. Moreover, inspections are made on manufacturing industries to ensure that the production follow the true protocol (Deore et al., 2019). This continuous monitoring ensures long-term safety and efficacy for the public.

Representations of molecules: There exist different ways to represent molecules as shown in Figure 2.2 such that they can be handle by an AI-driven drug discovery model (Rajan et al., 2021) (Noutahi et al., 2023).

- **SMILES(Simplified Molecular Input Line Entry System)**: Linear notation using ASCII strings to represent the 2D structure of chemical molecules. The special feature is that atoms are represented by their atomic symbols (e.g. C, N, O). The main advantage of this method is that it is concise and relatively human-readable for simple molecules. Although a single molecule can have several valid SMILES representations, this can lead to confusion.

- **DeepSMILES**: An extension to SMILES designed to improve the performance of recurrent neural networks (RNNs) in molecule generation. It aims to reduce the "long-range

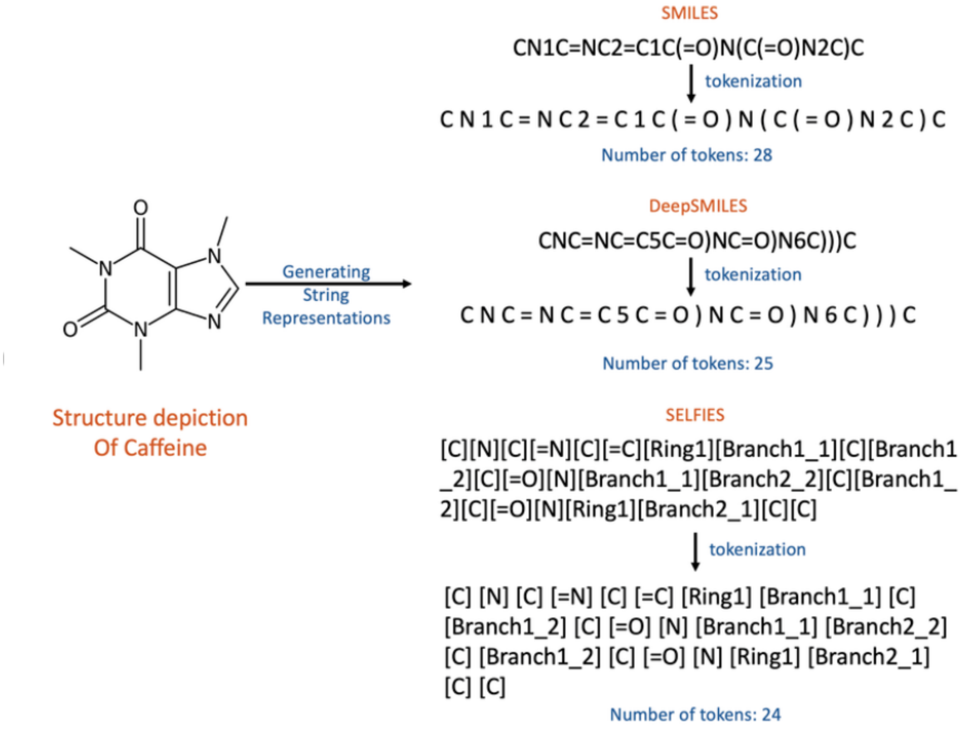
SMILES
CN1C=NC2=C1C(=O)N(C(=O)N2C)C
tokenization
C N 1 C = N C 2 = C 1 C (= O) N (C (= O) N 2 C) C
Number of tokens: 28
DeepSMILES
CNC=NC=C5C=O)NC=O)N6C)))C
tokenization
C N C = N C = C 5 C = O) N C = O) N 6 C))) C
Number of tokens: 25
SELFIES
[C][N][C][=N][C][=C][Ring1][Branch1_1][C][Branch1_2][C][=O][N][Branch1_1][Branch2_2][C][Branch1_2][C][=O][N][Ring1][Branch2_1][C][C]
tokenization
[C] [N] [C] [=N] [C] [=C] [Ring1] [Branch1_1] [C] [Branch1_2] [C] [=O] [N] [Branch1_1] [Branch2_2] [C] [Branch1_2] [C] [=O] [N] [Ring1] [Branch2_1] [C] [C]
Number of tokens: 24
Generating String Representations
Structure depiction Of Caffeine
O
N
N
O
N
N

Figure 2.2: Differents molecules representations (Rajan et al., 2021)

dependency" problem in SMILES. It introduces new characters to provide more local information on the structure of the molecule. DeepSMILES aims to improve the "learnability" of molecular representations, but is less human-readable than SMILES.

- **SELFIES(Self-Referencing Embedded Strings)**: A robust, sequence-based molecular string representation because every SELFIES string, even if randomly generated by a model, corresponds to a chemically valid molecule. It uses a vocabulary of "symbols" that represent atoms, bonds, rings, and branching. These symbols are enclosed in square brackets (e.g., $[C], [= C], [Ring1]$). The main advantage is the fact that any valid SELFIES string decodes to a valid molecule. Although SELFIES representation can lead to the generation of valid but strange or unrealistic structures).

- **AtomInSmiles**: A molecular string representation that incorporates atomic properties directly into the SMILES string. This allows generative models to learn and control these properties during molecule generation. It extends the SMILES notation by adding information about atomic properties within square brackets after the atom symbol. As an asset to use that representation, it provides more semantic information per token than character-level SMILES, potentially allowing Transformer models to learn chemical patterns more effectively. However, it requires calculating or predicting atomic properties, which can add computational overhead.

- **SAFE(SMILES Arbitrary Fragment Entry)**: A molecular string representation designed to facilitate fragment-based molecular design. It allows for the easy incorporation of predefined fragments into generated molecules. It emphasizes representing atoms within their local chemical context in a way that captures the "flow" or connectivity. The representation is more chemically intuitive tokenization than raw characters and like AtomInSMILES, it increases vocabulary size.

Metrics: When generating drugs, it is crucial to evaluate certain parameters to ensure the validity of the generated molecules. The most commonly used parameters include atom stability, which represents the percentage of atoms with the correct valency, and molecule stability, defined as the percentage of molecules whose atoms are all stable. Validity is measured as the percentage of stable molecules considered valid, often evaluated using RDKit. Uniqueness refers to the percentage of valid molecules that are unique, meaning they are not duplicates. Novelty is assessed by the percentage of molecules that are not present in the training dataset. Finally, the Quantitative Estimate of Drug-Likeness (QED) provides a formulaic combination of various molecular properties to estimate how likely a molecule is to be suitable for pharmaceutical purposes.

Drug development is a highly complex, costly and time-consuming process. It can take up to 15 years and cost around one billion euros. One of the difficulties lies in the initial stage of drug discovery process, where the objective is to design and develop a molecule that will achieve the desired outcome, even after several attempts. Drug development is sometimes halted due to a lack of funding, failures in pre-clinical and clinical trials, or other factors. Artificial intelligence (AI) is being increasingly integrated into this process due to its ability to rapidly analyse large sets of biological data in order to discover new targets and predict their therapeutic potential. AI can also optimise formulation and administration by predicting optimal physicochemical properties and drug release profiles.

2.2 Generative Artificial Intelligence

Generative Artificial Intelligence refers to algorithms and deep-learning models that can generate realistic and novel data. The content generated can be text, images, drugs, and other content based on the data they were trained on (Bandi et al., 2023). Generative AI is applied in various domain such as computer vision, natural language processing, and creative arts to address challenges including data augmentation, anomaly detection and drug discovery.

Real-world data has complex structures, but deep learning models with their many layers are able to learn complex structures and pattern inside the data like natural language or realistic images. Since we have to learn deeply the structure of the existing data in order to generate new content. Generative AI architecture are mainly based on Deep-Learning. According to the technique used, Gen AI can be classify as shown in figure 2.3 :

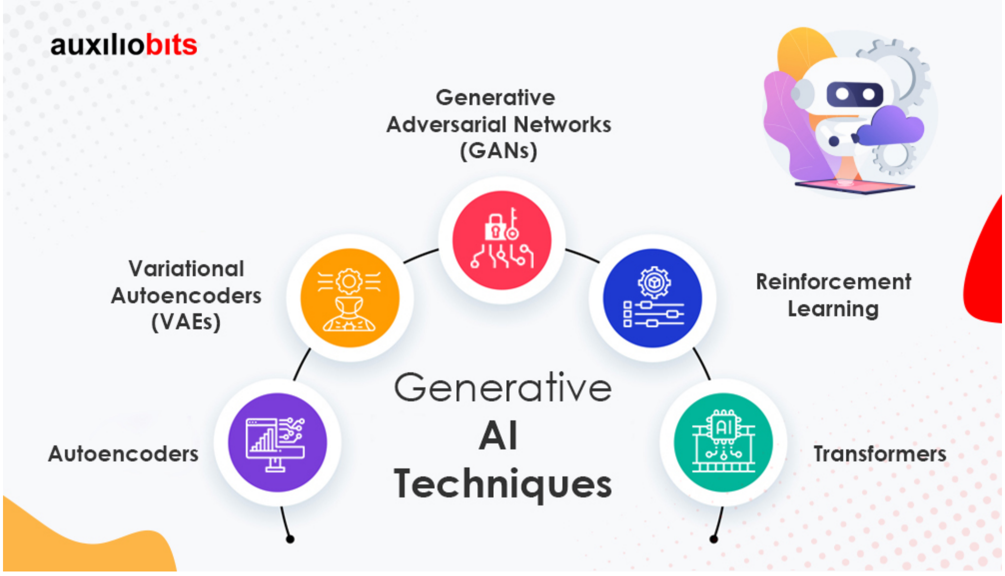

Figure 2.3: Generative AI Techniques Source

2.2.1 Variational Autoencoders (VAEs). An autoencoder is an algorithm designed to learn an informative representation of data, enabling various applications such as dimensionality reduction, feature extraction, and denoising, while effectively reconstructing input observations (Kingma and Welling, 2019). Its structure includes an encoder, which is a neural network that transforms the input data into a lower-dimensional latent representation, capturing the essential features of the data. The bottleneck layer holds this latent code, which serves as a compressed version of the input, facilitating tasks like visualization and classification. The decoder, another neural network, attempts to reconstruct the original data from this latent representation, thereby learning to minimize reconstruction error. Variational Autoencoders (VAEs) are a specific type of autoencoder that incorporate Kullback-Leibler (KL) divergence to regularize the latent space, encouraging the learned distribution to closely align with a prior distribution, typically a standard normal distribution. This regularization not only enhances the quality of the generated samples but also allows for more meaningful interpolation in the latent space, making VAEs particularly useful for generative tasks.

During the training, the VAE aims to balance between well reconstruction and well latent representations, which can be expressed as Kingma and Welling (2013):

$$\max_{\theta} \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - D_{KL} \left(q_\phi(z|x) \| p(z)\right) \tag{2.2.1}$$

where $\mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right]$ is the reconstruction loss; and D_{KL} is the KL divergence, which ensured that the learned distribution is close to the prior distribution. VAEs generate new molecules by learning latent representations of known compounds and modifying the values in the latent space to generate new content.

2.2.2 Generative Adversarial Networks (GANs). The generative adversarial network (GAN) was first introduced by (Goodfellow et al., 2014) and is a powerful machine learning model that consists of two neural networks: the Generator and the Discriminator. The Generator (G) is trained to deceive the Discriminator by taking random noise zz sampled from a simple prior distribution (such as uniform or Gaussian) and transforming it into a data sample G(z)G(z), with the goal of producing samples that are indistinguishable from real data. This process involves iterative training, where the Generator improves its ability to create realistic data by continuously adapting based on feedback from the Discriminator. Conversely, the Discriminator (D) receives a data sample xx, which can either be a real sample from the dataset or a fake sample generated by GG, and outputs a scalar value D(x)D(x) representing the probability that xx originates from the real data distribution. The Discriminator's objective is to accurately classify real samples as real (output close to 1) and fake samples as fake (output close to 0). This adversarial training process fosters a dynamic equilibrium where both networks enhance their capabilities: the Generator becomes increasingly adept at creating high-quality data, while the Discriminator sharpens its skills in distinguishing real from generated samples. GANs have been successfully applied in various fields, including image generation, video synthesis, and data augmentation, highlighting their versatility and effectiveness in generating realistic outputs.

The optimization problem consists of maximizing the probability of the discriminator misclassifying the generated data as real and minimize the probability of the generator to produce unrealistic data. We can formulate it as (Saxena and Cao, 2021):

$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.2.2}$$

where $p_{data}(x)$ is the distribution of real data; $p_z(z)$ is the distribution of the input noise (often Gaussian); $\mathbb{E}_{x \sim p_{data}(x)}[\log D(x)]$ represents the expectation of the log probability that the discriminator correctly identifies real data x drawn from the true data distribution $p_{data}(x)$; $\mathbb{E}_{z \sim p_z(z)}[\log(1-D(G(z)))]$ represents the expectation of the log probability that the discriminator correctly identifies generated data $G(z)$ (where z is sampled from the latent space distribution $p_z(z)$ as fake.

2.2.3 Diffusion Models. Diffusion models are a family of probabilistic generative models that progressively destruct data by injecting noise and then learn to reverse this process for sample generation (Yang et al., 2023). Inspired by non-equilibrium thermodynamics, diffusion models involve two main processes: the forward process and the reverse process.

In the Forward Process (Diffusion/Noising), the model begins with a real data sample (e.g., an image x_0) and gradually adds a small amount of Gaussian noise over a large number of discrete time steps T. The objective is to transform the data x_T into a state that is practically indistinguishable from pure Gaussian noise. Notably, this noising process is fixed and not learned, providing a structured way to degrade the data.

The Reverse Process (Denoising/Generation) involves the model learning to reverse the noising process through denoising. By iteratively applying learned denoising steps T times, the model transitions backward from $t = T$ to $t = 1$, ultimately generating a clean data sample x_0. This generation process relies on the model's ability to iteratively refine the noisy input, transforming it into a coherent and realistic output. Diffusion models have gained attention for their effectiveness in generating high-quality samples, and they are increasingly being applied in various domains, including image synthesis and audio generation, showcasing their versatility and potential in generative tasks.

The optimization problem is to minimize a loss function that quantifies how well the model predicts the noise added during the forward diffusion process (Ho et al., 2020).

$$\min_{\theta} \mathbb{E}_{x_0,\epsilon,t} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right] \tag{2.2.3}$$

where ϵ is the noise and $\epsilon_\theta(x_t, t)$ is the predicted noise by the model.

2.2.4 Transformers. Transformer is an architecture for transforming one sequence into another with the help of two parts (Encoder and Decoder), but it differs from the previously described/existing sequence-to-sequence models because it does not imply any Recurrent Networks (GRU, LSTM, etc.) (Vaswani et al., 2023). Unlike VAEs or Diffusion Models, which are primarily generative frameworks, the transformer is a specific neural network architecture. In the sense that it adds the attention mechanism to weigh the importance of different parts of the input data, allowing them to handle long-range dependencies more effectively than previous architectures.

The goal of training a transformer model is to minimize the difference between the predicted output and the actual output during training. The optimization problem can be formulated as (Aurpa and al., 2025):

$$\min_{\theta} \mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P(y_t | y_{<t}, x; \theta) \tag{2.2.4}$$

where y_t is the target token at time step t; $y_{<t}$ is the sequence of tokens generated up to time $t-1$; x is the input sequence.

The particularity of the transformer is the self-attention mechanism. This enables the model to capture long-range dependencies, analyze both local and global contexts simultaneously, and resolve ambiguities by attending to informative parts of the sentence[2]. Mathematically, we can efine attention by (Vaswani et al., 2023):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{2.2.5}$$

where $Q = XW_Q$, $K = XW_K$, $V = XW_V$, and W_Q, W_K, W_V are learnable parameter matrices.

2.2.5 Reinforcement Learning. : Reinforcement learning (RL) is a machine learning (ML) technique that trains software to make decisions to achieve the most optimal results (Sutton and Barto, 1998). It mimics the trial-and-error learning process that humans use to achieve their goals. Software actions that work towards your goal are reinforced, while actions that detract from the goal are ignored.

RL algorithms use a reward-and-punishment paradigm as they process data. They learn from the feedback of each action and self-discover the best processing paths to achieve final outcomes. The

[2] https://www.datacamp.com/blog/attention-mechanism-in-llms-intuition(12/05/2025)

algorithms are also capable of delayed gratification. The best overall strategy may require short-term sacrifices, so the best approach they discover may include some punishments or backtracking along the way. RL is a powerful method to help artificial intelligence (AI) systems achieve optimal outcomes in unseen environments.

In the context of generating new things, the "thing" could be anything from text and images to molecules and game strategies. In generation context, the environment represents the space of possible things that can be generated. It also includes the rules and constraints that govern how the agent can interact with this space. The agent is the RL algorithm that learns to generate new things by interacting with the environment. The agent has a policy that determines its actions in each state of the environment. He uses the rewards to update its policy, so that it is more likely to take actions that lead to high rewards. Once the agent is trained, it can be used to generate new things by starting in an initial state and repeatedly taking actions according to its learned policy. The process continues until a termination condition is met (e.g., a sentence is complete, an image is fully generated, a molecule reaches a desired size). RL allows you to explicitly define a goal for the generation process through the reward function. However, RL algorithms can be sample-inefficient, meaning that they require a lot of interaction with the environment to learn a good policy.

In this section, it was question for us to give the theoritical background necessary to understand the ideas of the work. In the next part, we are going to talk about some works on Drug Discovery using Generative AI and the current state of drug ddiscovery in Africa.

3. State of the Art

This thesis focuses on Drug Discovery and on this chapter have to goal to give a general idea on the current state of drug discovery using Generative AI and also the actual work done in Africa.

3.1 Generative AI for Drug Discovery

Generative AI is increasingly used in the drug discovery process, particularly in the initial phases, due to its ability to create new molecular structures that could lead to new drug candidates not easily identified by traditional methods. To speed up the costly and time-consuming process of drug development, researchers are increasingly turning to generative AI using differents architectures such as:

3.1.1 Variational Auto-encoders. "Benchmarking Deep Graph Generative Models for Optimizing New Drug Molecules for COVID-19" (Ward et al., 2021)

Current Projects Drug search and screening actively uses machine learning. Drug design has made use of deep learning. Deep reinforcement learning (RL) variants, variational autoencoders (VAEs), and graph convolutional policy networks are some of the graph-generative models investigated for target-driven optimization. Examples that have been specifically mentioned include the CogMol generative framework, the Deep Purpose Toolkit for Drug Repurposing, ML-based screening for binding affinity, and the Pac-cMann RL approach modified for viral targets. In particular, a graph-based deep reinforcement learning technique (DQN) and a junction-tree based variational autoencoder (JT-VAE) are evaluated in this work. Another pertinent method is message-passing neural networks (MPNNs). One well-known technique for VAEs is Bayesian optimization (BO). DQN uses Morgan fingerprints as a molecular representation. There are Drug-Target Binding Affinity (DTBA) techniques based on machine learning. To design new drug compounds with target properties, specifically for COVID-19, (Ward et al., 2021) has proposed a model based on Variational Autoencoders. The goal of the Junction-Tree Variational Autoencoder (JT-VAE) model is to produce molecular graphs that take into account important bioactivity properties like $pIC50$ and satisfy a user-specified threshold for a property optimization function. A graph-based generative model called the Junction-Tree Variational Autoencoder (JT-VAE) is intended to generate chemically valid molecules by simulating their substructures and interrelationships. Its main concept is to depict molecules as two-level graphs: the corresponding molecular graph made up of atoms and bonds, and a junction tree that captures the arrangement of substructures (like rings and functional groups). It uses a surrogate prediction model for pIC50 to guide Bayesian optimization. After that, a message-passing network is used to train the encoder to encode into a vector representation (ZT). By maximizing the likelihood function, the decoder component learns to produce the same junction tree structure from the latent vector ZT. Following training, molecules are produced in two stages: first, a junction tree is created by sampling from the latent space, and then the junction tree is mapped to a molecular graph that maximizes target properties using Bayesian optimization (BO). For BO, a sparse Gaussian process (SGP) is employed. A database of recognized drug molecules with pIC50 activity associated with SARS was used to train the JT-VAE. JT-VAE builds from pre-defined substructures (also known as "vocabulary") to produce chemically valid molecules by default. Functional groups that are typical of the training set are incorporated. Even when these are not specifically optimized for, they frequently yield molecules with consistently better druglikeness (QED) and synthesizability (SA). If the training set has high-quality examples, the structural similarity of the generated molecules to those in the training set may be advantageous. JT-VAE produced new molecules structurally similar to indinavir, some of which had higher predicted pIC50. The DBTA classifier test was passed by four of the top JT-VAE molecules. The limitations of JT-VAE include its dependence on pre-existing structural motifs and training data biases, which hinder its ability to explore truly novel chemical space, balance multiple properties, and ensure practical drug-likeness beyond chemical validity, even though it efficiently produces chemically valid and structurally similar molecules.

3.1.2 Reinforcement Learning . "ACEGEN: Reinforcement learning of generative chemical agents for drug discovery " (Bou et al., 2024)

(Bou et al., 2024) proposes a model based on reinforcement learning (RL). Indeed, reinforcement learning is increasingly used by researchers to find new drugs. However, existing RL implementations for drug discovery rely heavily on custom code, resulting in redundancy, complexity and limited effectiveness. To overcome these limitations, the authors present ACEGEN, a new toolbox designed to streamline and improve the generative drug design process using RL. ACEGEN is a new open-source, modular and reliable toolbox for generative drug design, based on TorchRL. What makes it special is that it facilitates algorithmic research by encapsulating new ideas in seamlessly integrated components. ACEGEN is based on the formalization of reinforcement learning (RL) tasks as Markov decision processes (MDP), described by the quintuple **(S, A, R, P, ρ_0)** where S is the set of all possible states representing a partially constructed molecule, A represents the set of valid actions available to the agent, corresponding to molecular modifications or tokens in a molecular chain representation such as SMILES. $R : S \times A \times S \longrightarrow R$ is the reward function that assigns a numerical value to the transition from one state to another depending on the action performed, and ρ_0 represents the distribution of the initial state. One of the main advantages of the ACEGEN model is its support for various grammar environments, including SMILES, DeepSMILES, SELFIES, AtomInSmiles and SAFE. Here, molecular generation is realized as an automatic natural language processing (ANLP) problem. Policy models, called chemical language models (CLMs), generate molecules sequentially as strings, with actions as tokens. CLMs are usually unsupervisedly pre-trained on large datasets (such as ChEMBL or ZINC) to learn how to generate valid molecules, often using the teacher application method. This pre-trained CLM serves as an initial π_θ policy, which is then refined by RL. The previous policy can also be used as an anchor to prevent excessive deviation. ACEGEN supplies pre-trained models and ready-to-use architectures (GRU, LSTM, GPT2, Llama2, Mamba), enabling users to integrate customized architectures. The authors use numerous metrics to evaluate the performance of different reinforcement learning algorithms and generative models in the context of drug design. These metrics aim to assess not only the ability to optimize specific goals, but also other crucial aspects such as efficiency, exploration, chemical quality and sample diversity. However, it can be very difficult to customize the ACEGEN code, making it difficult to integrate into a variety of solutions.

3.1.3 Transformers and GAN. "Target-specific de Novo Design of Drug Candidate Molecules with Graphic Transformer-Based Generative Adverserial Networks" (Ünlü et al., 2023)

The existing generative models often are not able to produce drug like molecules that are large enough and structurally relevant for practical drug development. They may generate molecules that, while valid, are not optimized for interaction with specific biological targets, limiting their utility in real-world applications. (Ünlü et al., 2023) propose DrugGEN an end-to-end framework for de novo design of target-centric, drug-like molecules. It based on GAN architecture with graph Transformer layers incorporating graph representation learning. The goal is to generate drug-sized molecules and learn long-range dependencies between atoms.

Molecules are on SMILES string formats and represented as graphs, with:

$$G = (A, X)$$

where $A \in \{0,1\}^{N \times N}$ is the adjacency matrix encoding bonds between atoms and $X \in \mathbb{R}^{N \times F}$ contains atom features (annotations). The architecture of DrugGEN consists of two main components: the Generator (G_θ) and the Discriminator (D_ϕ). The Generator is a graph transformer encoder that takes randomly selected real drug-like molecules from the training data as input, rather than random noise, allowing it to effectively manage the high complexity and sparsity of large molecular graphs containing 38 to 45 heavy atoms. It incorporates graph transformer layers that utilize self-attention over atom embeddings to learn long-range dependencies. The Generator also accepts a random noise vector z (sampled from a prior distribution, such as Gaussian) along with the target representation t, aiming to produce a molecular graph $\hat{G}$ that closely resembles real, target-specific molecules. The Discriminator, which also features graph transformer layers, compares the generated molecules with real inhibitors of the target to evaluate their structural plausibility.

To evaluate the model, the authors have used may metrics that can be categorized into three main groups: the first group is the Fundamental Molecular Generation Metrics (Structural and Diversity Properties) to ensure the basic efficiency and quality of the generated molecules by evaluating the validity, the uniqueness and the novelty. The second is the Physicochemical and Drug-Likeness Metrics evaluating how well the generated molecules match with the desirables properties for the drug candidates. There are Quantitative Estimate of Drug-likeness (QED) and Synthetic Accessibility (SA). And the last is the Target-Centric and Binding Affinity Metrics to assess if the generated molecules interact with a specific protein target (AKT1 in this study) by computing the Molecular Docking Scores.

Overall, the results indicate that DrugGEN is an effective system for target-centric de novo molecule design, capable of generating novel, drug-like molecules with predicted high binding potential to the intended target like AKT1 and CDK. However, training such graph transformer-based GANs requires significant computational capacity, which may limit accessibility for some researchers.

3.2 Drug Discovery in Africa

Africa is incredibly rich in biodiversity, with a vast number of plants, microorganisms, and marine life that have been used in traditional medicine since many centuries. Many of these natural products contain bioactive compounds with proven or potential therapeutic properties. Global pharmacology has already benefitted from African natural products, such as Prunus africana compounds for prostate issues and Madagascan periwinkle vincristine / vinblastine for cancer. However, drugs used in Africa are made outside from the continent but factors like nutrition, co-infections and environmental exposures common in some African contexts can influence drug efficacy and safety. Increasingly, some African companies, research institutions, universities, and even governmental initiatives are focusing on modernizing the drug discovery process.

We have for example in Ghana (Amewu et al., 2022) some institutions like the Noguchi Memorial Institute for Medical Research and universities such as the University of Ghana, KNUST, and UHAS have established laboratories capable of pathogen screening, biosafety level 3 facilities, and some research on herbal medicines. The focus includes infectious diseases prevalent in Ghana, such as malaria, tuberculosis, leishmaniasis, schistosomiasis, and non-communicable diseases such as cancer, which require tailor-made diagnostic and treatment research. Much work attempts to exploit small molecules from natural sources such as plants, marine invertebrates, fungi and bacteria. Some of this research is guided by traditional medicine, with the objective to identify the active constituents. These investigations have yielded structurally unique compound classes with promising anti-infective activities. However, training is needed in the specialized aspects of drug discovery, including understanding the stages, the integration of disciplines, drug metabolism and pharmacokinetics (DMPK), data interpretation, and collaborative decision-making. Also, an expertise is needed in robotics, data processing software, computational chemistry techniques, machine learning, and artificial intelligence.

Increasingly, researchers are turning to artificial intelligence to assist in the drug discovery process. We have (Namba-Nzanguim et al., 2022) providing a perspective on the benefits, and limitations of AI/ML tools in antiviral drug discovery specifically in low and middle income countries (LMICs), which include many countries in Africa. There is a huge need for new drugs targeting infectious diseases prevalent in LMICs, like viral, bacterial, and fungal infections. However, existing AI/ML tools are largely biased towards non-communicable diseases (like cancer) due to more data availability and research investment in High-Income Countries (HICs). The intrinsic cost and risk of drug discovery accentuate the need to invest in diseases affecting LMICs. AI/ML could potentially have the greatest impact in settings where experimental costs are prohibitive. A major barrier is the shortage of skills and training in data science, computer science, chemoinformatics, and bioinformatics in LMICs. Also, the access to good quality, task-specific data is essential for successful AI/ML modelling. However, there exists some Natural Product (NP) databases, particularly from African species (like AfroDB, ANPDB, AfroMalariaDB), are highlighted as an untapped resource of novel chemical structures for antiviral drug discovery in LMICs. These databases can be used by AI/ML tools offer the solutions with competitive performance in low-resourced settings. In that sense, the emerging Center for Drug Discovery (UB-CeDD) at the University of Buea in Cameroon as a model for Central Africa aims to discover plant-based antivirals, combining AI/ML and physics-based methods for virtual screening. They are generating data from screening natural and synthetic compounds in antiviral assays to build robust AI/ML models, which also serves as training for graduate students and postdocs in implementing AI/ML
.

The Holistic Drug Discovery and Development Centre (H3D) at the University of Cape Town

(UCT) in South Africa has set out to become a source of healthcare innovation, creating and developing an ecosystem of innovative pharmaceuticals, as reported by (Singh et al., 2022). Founded in 2010 as a UCT-accredited research center and officially launched in April 2011, H3D is the first and only integrated drug discovery platform of its kind on the African continent. Its vision is to be a leading drug discovery and development organization, with a mission to discover and develop innovative, life-saving medicines for diseases that primarily affect African patients. H3D also focuses on developing Africa-specific models to improve treatment outcomes, and training skilled African scientists in drug discovery. H3D collaborates with a global network of partners, including industry, academia, product development partnerships (PDPs) such as Medicines for Malaria Venture (MMV) and TB Alliance, philanthropic organizations and the South African government. The projects are carried out in Africa and led by H3D, an organization considered important and advantageous for several reasons: due to the high burden of disease, African scientists should drive drug discovery efforts; conducting work as close as possible to African patients helps to understand and address their health needs, given the interplay between genetics, socio-economic and physical environment; and it helps to build sustainable drug discovery capacity. However, the drawbacks are limited access to infrastructure, technology and experience, and a limited pool of qualified scientists due to the brain drain. With regard to malaria, H3D's work has contributed to the global pipeline of therapeutic interventions against malaria. They have used whole-cell high-throughput phenotypic screening (HTS) of a library of small molecules against drug-sensitive and drug-resistant strains of Plasmodium falciparum (Pf). This has led to the identification of over 200 scores with selective antiplasmodial activity, including a series of 3,5-diaryl-2-aminopyridines exemplified by compounds 1-4. H3D has also developed a portfolio of products for tuberculosis, including H2L and LO programs, using a medium-throughput screening platform for phenotypic and targeted screening. Whole-cell screening has proved more effective in providing starting points. Hits are developed by cellular medicinal chemistry to improve potency and PK/PD properties while minimizing toxicity.

3.3 Application of Generative AI for Drug Discovery in Africa

Based on the previous section, the current state of drug discovery research in Africa is characterized by significant ambition and ongoing effort, but also by substantial challenges. One of the biggest challenges is access to data, although there is the African Natural Products Database (ANPDB)[1] platform, which provides researchers with databases of natural products Currently, the ANPDB comprises the North African Natural Products Database (NANPDB) and the East African Natural Products Database (EANPDB). NP databases for other African regions are still under construction and/or not yet integrated into the ANPDB. The data content currently comprises data sources (covering the period from 1962 to 2019) and is derived from literature collected in leading natural products journals, as well as master's and doctoral theses from selected African university libraries and selective searches in local African journals. The data covers 6,515 compounds isolated mainly from 1,042 source organisms; mainly plants, with contributions from

[1] https://african-compounds.org/anpdb/(10/05/2025)

micro-organisms, animals and marine sources (Namba-Nzanguim et al., 2022). However, the ANPDB is still little-known, but it represents a major step forward in modernizing the drug discovery process, as it enables us to build models based on our own resources, adapted to our own context and conditions. Also, access to this data will enable us to turn to generative AI to speed up the drug discovery process. The potential of GenAI can be used to design novel molecular structures with predicted activity against specific biological targets relevant to these diseases. This could accelerate the identification of potential drug candidates tailored to the local health burden, where there is currently "inadequate research in drug discovery into these diseases locally". Such insights can then inform the generative process to design more promising candidates. AI-based tools can also help characterise and optimise experimental conditions in areas like plant in vitro culture and predict phytochemical potential, improving the bio-relevance of assays. However, implementing Gen AI also requires specialized training in computational chemistry, data science, AI model development, deployment and interpretation. Gen AI requires significant computing power and data storage, necessitating investment in a robust computing infrastructure, likely to encounter similar procurement and maintenance difficulties. Effective AI models require large, high-quality datasets for training. While a natural products database is currently under development, expanding and standardizing the collection of data from screening, compound characterization and biological studies across institutions would be essential to train successful Gen AI models.

The application of Gen AI could significantly improve and accelerate research into the diseases concerned, support existing efforts in natural products and synthetic chemistry, and help achieve the goal of producing preclinical candidates. However, this requires substantial investment in targeted training programs, consideration of infrastructure and equipment needs, and the development of robust data management systems, while harnessing the existing spirit of collaboration. In the remainder of the work we have decided to focus on the generation of molecules to catalyze cancer-causing proteins.

4. Problem Statement and Methodology

In this chapter, we discuss the methodology used from generation to selection of the best generated molecules. The aim is to generate molecules against cancer using resources from African flower species based DrugGEN model.

4.1 Problem Statement

Generative AI has developed in recent years, with impacts such as content creation, increased data on and its ability to collaborate with humans. The beauty of generative AI lies not only in the generation of new content but rather in the ability to apply changes to achieve better results. It is also increasingly used in the first stage of the drug discovery process to speed up the process and reduce costs. A great deal of work has been done on generative AI for drug discovery. Among the existing models, the DrugGEN model has been shown to generate good molecules, but one of the great features is that we can generate molecules to catalyze a specific molecule. Given that drug discovery in Africa still encounters certain difficulties, generative AI appears to be a solution. Thus, in this work, we focus on the study of Generative AI to generate molecules against cancer-causing proteins using compounds from African resources.

4.2 Methodology

4.2.1 Dataset. According to the fact that DrugGEN have been trained to inhibite protein responsible of cancer such as AKT and CDK, we have decided to use the AfroCancer database available on the ANPDB (African Natural Products Database). The ANPDB is a platform focused on providing access to a comprehensive collection of natural compounds from Africa (Namba-Nzanguim et al., 2022). For each database we have three differents formats: SMILES, Structure Data File format for 2D chemical structures and 3D chemical structures. In fact, (Ntie-Kang et al., 2014) study 400 compounds from African medicinal plants that have shown in vitro and/or in vivo anticancer, cytotoxic, and antiproliferative activities has been explored. It contains around 1000 compounds isolated from approximately 200 African medicinal plant species (and associated microorganisms/fungi). The data was primarily compiled through extensive searching and manual curation of scientific literature (journals, theses, conference proceedings) reporting natural products from African sources with anti-cancer activity. To verify potential binding to anticancer drug targets, the interactions between the compounds and 14 selected targets have been analyzed by in silico modeling. Docking and binding affinity calculations were carried out, in comparison with known anticancer agents comprising 1 500 published naturally occurring plant-based compounds from around the world. The results reveal that African medicinal plants could represent a good starting point for the discovery of anticancer drugs. AfroCancer contains then three different databases: The first containing compounds from North African medicinal plants (smiles_unique_NANPDB.smi), the second compounds from East African medicinal plants (smiles_unique_EANPDB.smi) and the last combining compounds from both regions (smiles_unique_all.smi).

4.2.2 Data Preprocessing. Before using our databases, we need to perform several preprocessing steps to prepare structured, standardized graph representations suitable for training the graph transformer generator in the DrugGEN model. First, we convert the SMILES files from the AfroCancer dataset into the appropriate format expected by the model. Next, we transform the molecules from their SMILES representations into graph structures using RDKit, which generates annotation matrices for atom features and adjacency matrices for bond information. This conversion allows the models to process the molecules as structured graph data. We then extract atomic features, such as atom types, from the graph structures and store them in annotation matrices using one-hot encoding. Bond types (single, double, triple, aromatic, or no bond) are represented in the adjacency matrices with specific one-hot encodings for each bond type. To ensure consistency across the dataset, we standardize these features. Additionally, to address molecule size variability, we filter the molecules based on heavy atom counts, setting a threshold of 45 heavy atoms determined through distribution analysis. Ultimately, the dimensions of the annotation and adjacency matrices (45x9 and 45x45x5, respectively) reflect this maximum length of 45 heavy atoms.

4.2.3 Molecule Generation with Trained Models. In this work, we are using DrugGEN, a new de novo small molecule design system, an end-to-end framework, that generates target-centric drug-like molecules. The core idea of the DrugGEN system is to incorporate graph transformer layers in the architecture of the Generative Adversarial Network (GAN) as shown on figure 4.1. Thus, the architecture consists of two main components:

The generator(G) module employs transformer encoder blocks that operate on graph-based data. Unlike typical GANs that use random noise as input, DrugGEN's generator takes randomly selected real drug-like molecules from the training dataset as input. This approach helps to manage the high complexity and sparsity of large molecular graphs. The input annotation and adjacency matrices are first processed through individual multi-layer perceptrons (MLPs) to create embeddings. Both MLPs consist of four layers. The embeddings for both matrices have a dimension size of 128 (the default dk dimension of the transformer encoder module). Within the graph transformer, the self-attention mechanism calculates attention weights by multiplying the adjacency matrix (A_m) with the scaled dot product of the query (Q_m) with key (K_m) representations of the annotation matrix (V_m) and d_k the dimension of the transformer encoder module and it is used to scale the attention weights. The attention calculation is formally described as:

$$\text{Attention}(Q_m, K_m, V_m) = \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d_k}} A_m\right) V_m \tag{4.2.1}$$

After the attention layer, the resulting annotation and adjacency matrices undergo an addition and layer normalization process:

$$\hat{X}^{l+1} = \text{LayerNorm}(X^l + \bar{X}^{l+1}), \quad \hat{A}^{l+1} = \text{LayerNorm}(A^l + \bar{A}^{l+1}) \tag{4.2.2}$$

X and A correspond to annotation and adjacency matrices, respectively, LayerNorm represents the layer normalisation, and l is the layer number. $\hat{X}^{l+1}$ represents the intermediate product after the layer normalisation. X^l is the annotation matrix before attention and $\bar{X}^{l+1}$ is the product of the attention mechanism. The same annotations are used for A (adjacency matrix).

They are then processed by a feed-forward network (FFN) to introduce non-linearity and learn complex representations. The output of the FFN is added back via a residual connection, and the final matrices are passed through layer normalization :

$$X^{l+1} = \mathrm{LayerNorm}(\hat{X}^{l+1} + \mathrm{FFN}(\hat{X}^{l+1})), \quad A^{l+1} = \mathrm{LayerNorm}(\hat{A}^{l+1} + \mathrm{FFN}(\hat{A}^{l+1})) \quad (4.2.3)$$

The Discriminator (D) from his part compare molecules generated by the generator (synthetic/fake) with real molecule data and classify them as either "fake" or "real". Similar to the generator, DrugGEN's discriminator also incorporates graph transformer encoder blocks. It starts by processing the input annotation and adjacency matrices through linear layers to obtain embeddings. These embeddings are then fed into the graph transformer encoder blocks for further representation transformation. The output node representations from the transformer are processed by an MLP prediction head. This MLP has layers with decreasing neuron sizes to produce the final classification scores for the real/fake evaluation.

DrugGEN aims to learn the distributions of physicochemical properties and topological attributes of drugs and drug candidate molecules from the given data to generate drug-like small molecules that are valid and novel. During the training we are trying to minimize Wasserstein Generative Adversarial Network (WGAN) (Arjovsky et al., 2017) loss given by (4.2.4):

$$L = \mathbb{E}_{x \sim p_r(x)}[D(x)] - \mathbb{E}_{z \sim p_g(z)}[D(G(z))] \quad (4.2.4)$$

where x denotes real molecules, which are experimentally validated inhibitors of the target of interest; z denotes the input distribution of the generator; p_r denotes real data distribution and p_g the generated data distribution. However, in the literature it has been shown that using gradient penalty (GP) improves the performance of WGAN. So, the authors integrated the loss GP formulated as (4.2.5):

$$L_{GP} = \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}(\hat{x})} \left[\left(\left\| \nabla_{\hat{x}} \tilde{D}(\hat{x}) \right\|_2 - 1 \right)^2 \right] \quad (4.2.5)$$

where λ denotes a penalty coefficient; $\hat{x}$ denotes generated samples coming from x (real data); $p_{\hat{x}}(\hat{x})$ refers to sampling uniformly along straight lines between pairs of points from the data distribution p_r and generator distribution p_g. We obtained our finalised loss function (4.2.6) as via combining the losses we got:

$$L_{\text{total}} = L + L_{GP} \quad (4.2.6)$$

After training, the DrugGEN model is used in inference mode to generate novel molecules, which are then evaluated using various metrics. Validity measures the fraction of valid molecules in the generated set, with higher values indicating a greater proportion of chemically feasible compounds, essential for drug development. Uniqueness assesses the fraction of unique molecules, maximizing diversity within the library of compounds, which is crucial for identifying novel drug candidates. Novelty evaluates the fraction of generated molecules that are not present in the reference sets, highlighting the discovery of new therapeutic agents that do not overlap with existing drugs. Internal diversity is measured by the average Tanimoto similarity between all pairs of generated molecules; higher internal diversity (lower similarity) suggests a broader range of chemical structures, increasing the likelihood of finding effective drugs. The Quantitative Estimate of Drug-likeness (QED) calculates the average QED score of the generated molecules, ensuring they possess favorable characteristics for drug development. Synthetic Accessibility (SA) measures the average SA score, with lower scores being more desirable, as they indicate easier-to-synthesize compounds. Overall, this evaluation process results in a set of newly generated molecules, which are then validated and selected based on their potential activity against the target protein.

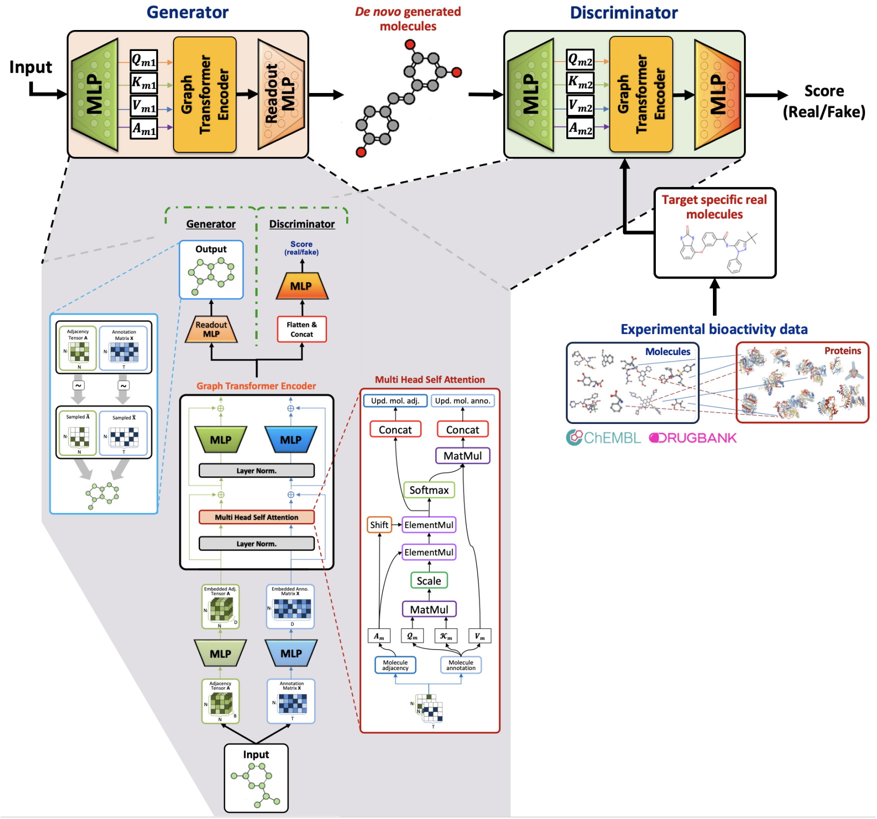

Figure 4.1: Architecture of the DrugGEN (Ünlü et al., 2023)

4.2.4 Deep Learning-based Bioactivity Prediction. After the generation step by the DrugGEN model, the produced molecules are evaluated using a deep learning model called DEEPScreen (Rifaioglu et al., 2020), which predicts bioactivity to assess potential interactions with the target protein AKT1. DEEPScreen is a supervised discriminative deep learning system specifically designed to classify molecules as either active (interacting) or inactive (non-interacting) against a specific protein target. It uniquely takes as input 2-D structural representations (images) of compounds, typically sized 200-by-200 pixels or 300-by-300 pixels, generated from standard SMILES representations using tools like RDKit. This image-based input approach allows for greater coverage of compound features compared to conventional featurization methods like fingerprints. The architecture as shown on figure 5.1 is built on Deep Convolutional Neural Networks (DCNNs), which are well-suited for processing image data, with each target protein having its own individual prediction model optimized with specific hyperparameters. DEEPScreen models are trained using curated experimental bioactivity data, such as that from the ChEMBL database, enabling the CNN to recognize structural features associated with activity by analyzing numerous labeled images. Once trained, the model predicts the likelihood that a new molecule interacts with the target by processing its 2-D image and outputting a probability score; higher scores indicate greater confidence in activity, with molecules above a certain threshold classified as "active" and those below as "inactive." DEEPScreen operates independently from the DrugGEN system in terms of its modeling approach, datasets, and outputs, providing a robust bioactivity prediction mechanism through its tailored target-specific training and image-based approach.

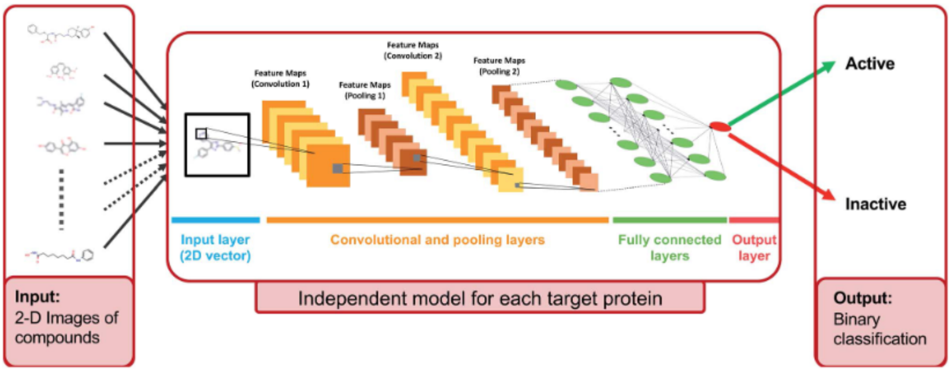

Figure 4.2: Architecture of DEEPScreen (Rifaioglu et al., 2020)

4.2.5 Final Candidate Molecules Selection. The de novo molecules generated by DrugGEN are subjected to a downstream analysis pipeline to identify the most promising candidates for further investigation. This acts as an early selection process Molecules predicted as "active" by DEEPScreen, especially those with high confidence scores and good molecular docking results, are then considered for final selection as promising drug candidates through expert curation. This provides an independent validation of predicted bioactivity based on structural features learned from experimental data. The independence of DEEPScreen from the DrugGEN generative model is highlighted as ensuring an objective evaluation of the generated molecules. These selected molecules are intended to be directed towards the next steps, such as synthesis and wet-lab validation.

This chapter provides a comprehensive overview of the methodology adopted in this study. We have presented the various steps involved in training and selecting promising generated molecule candidates. This information will serve as the basis for the understanding, implementation of the experiments and analysis of the results presented in the remainder of this work.

5. Results and Discussions

In this chapter, we describe the environment use for our experimentations and the results we got for each step of the methodology, starting from the training to the selection of the final candidate generated molecules.

5.1 Study Area

Globally, our experimentations have been made directly in the bash terminal and Visual Studio Code of one Dell Desktop on Debian 64 bits Operating System with Intel Core i7-10700 CPU @ 2.90GHz, 16 cores, 16 RAM memory and 512 Gb of Disk Capacity. The environment is built using Anaconda 1.14.0. The programming language is Python and we have use many librairies; these librairies are saved in the Requirements.txt. We can list :

- **PyTorch**: A deep learning framework for building and training neural networks.
- **RDKit**: A cheminformatics toolkit for handling and analyzing chemical structures, used for processing SMILES strings and molecular data.
- **Matplotlib/Seaborn**: Libraries for data visualization.

5.2 Training of the models

We have trained three different models: the first, named "DrugGEN-akt1," is designed to generate molecules that inhibit the AKT protein; the second, "DrugGEN-cdk2," targets the CDK protein, and the last, "DrugGEN-NoTarget," is trained without a specific target protein. For these experiments, the models are trained on the ChEMBL dataset, which contains a wide range of chemical compounds, and the AfroCancer dataset, which includes natural compounds from North Africa. Additionally, the training incorporates a target dataset that contains compounds specifically relevant to inhibiting the respective proteins. Another parameter established is a limit on the maximum number of atoms allowed in the molecular structures being processed. This helps control the complexity of the molecules generated or used in training.

To train DrugGEN, we used the following commands:

```
python train.py --submodel="[MODEL_TYPE]" \
            --raw_file="data/[GENERAL_DATASET].smi" \
            --drug_raw_file="data/[TARGET_DATASET].smi" \
            --max_atom=[MAX_ATOM_NUM]
```

After differents implementations, the configuration of the model giving the best results is shown in table 5.1.

Table 5.1: Model Configuration

Aspect	Details
Model Architecture	GANs utilizing Graph Transformer architecture
Batch Size	128
Training Epochs	100 epochs (with early stopping based on validation and novel metrics)
Learning Rate	1e-4; after hyperparameter optimization for all models (G and D)
Attention Heads	Evaluated sizes: 16, 32, 64; 128 was found optimal
Embedding Dimension	512 (hyperparameter convergence)
Transformer Depth	1
Training Duration	Approximately 6 days(on CHEMBL) and 1 hour(on AfroCancer)
Training Data	90% of the ChEMBL dataset and North African Afrocancer dataset

5.3 Results of the De Novo Generated Molecules by DrugGEN Models

For the experiments we have used the different pretrained models and evaluated the generated molecules using the metrics. The generation takes as parameters the test dataset in SMILES format, the train dataset in SMILES format, the target dataset in SMILES format and the maximum number of atoms.

```
python inference.py --submodel="[MODEL_TYPE]" \
                    --inference_model="experiments/models/[MODEL_NAME]" \
                    --inf_smiles="data/[TEST_DATASET].smi" \
                    --train_smiles="data/[TRAIN_DATASET].smi" \
                    --train_drug_smiles="data/[TARGET_DATASET].smi" \
                    --sample_num=[NUMBER_OF_MOLECULES] \
                    --max_atom=[MAX_ATOM_NUM]
```

5.3.1 Metrics for DrugGEN-akt1. For this experiment, we used the pretrained "DrugGEN-akt1," designed to generate novel drug candidates, possibly specifically targeting Akt1. The model's performance is evaluated under two different training conditions: once trained on the "AfroCancer North African Dataset named smiles_unique_NANPDB.smi" and once trained on the "ChEMBL Dataset." For each training condition, the generated molecules are then evaluated against two different reference sets: smiles_unique_EANPDB.smi (dataset of unique SMILES from East African natural products) and chembl_test.smi (a test set derived from the ChEMBL database). The evaluation of the generated molecules is present in table 5.2 and reveals several key metrics. Validity shows a perfect score of 1.0 when assessed against smiles_unique_EANPDB.smi, but a lower score of 0.713 for chembl_test.smi with the model trained on the AfroCancer dataset, suggesting that it learns the chemical "rules" of this dataset well but may not conform as closely to the broader chemical space represented by ChEMBL. In contrast, the ChEMBL-pretrained model demonstrates higher validity scores (0.741 for smiles_unique_EANPDB.smi and 0.92 for chembl_test.smi), indicating better performance in generating valid compounds. Uniqueness is consistently perfect at 1.0 across all datasets, confirming that all generated SMILES strings are unique. Novelty scores are high (close to 1) during training and inference, indicating that most generated compounds are novel compared to the training set, though the score for chembl_test.smi drops slightly to 0.959, showing some similarity with existing compounds. For real inhibitors, the novelty is perfect at 1.0, indicating all generated compounds are novel relative to known inhibitors. The average length of generated SMILES strings is shorter for the AfroCancer model (0.245) compared to ChEMBL (0.584 and 0.625), suggesting less complexity in the generated molecules from the AfroCancer dataset. The mean atom type scores are lower for AfroCancer (1.016 and 0.476) compared to ChEMBL (2.373), indicating a more diverse range of chemical structures in ChEMBL. Structural similarity (SNN) scores are low for AfroCancer (0.033) but moderate for ChEMBL (0.404 and 0.464), showing that ChEMBL-generated compounds are more structurally aligned with known compounds. Internal diversity scores are higher in ChEMBL (0.864 and 0.878) compared to AfroCancer (0.383), indicating a broader variety of chemical structures. The QED (Quantitative Estimate of Drug-likeness) scores are moderate, ranging from 0.456 to 0.57, suggesting that the generated compounds may not be optimally drug-like. Lastly, synthetic accessibility (SA) scores are higher for the AfroCancer dataset, indicating that the generated compounds may be more challenging to synthesize than those from ChEMBL.

Table 5.2: Metrics for DrugGEN-akt1

	AfroCancer Dataset		ChEMBL Dataset	
Metrics	smiles_unique_EANPDB.smi	chembl_test.smi	smiles_unique_EANPDB.smi	chembl_test.smi
Validity	1.0	0.713	0.741	0.92
Uniqueness	1.0	1.0	1.0	1.0
Novelty (Train)	0.983	0.985	0.969	0.991
Novelty (Inference)	0.997	0.994	0.993	0.959
Novelty (Real Inhibitors)	1.0	1.0	1.0	1.0
Average Length	0.245	0.585	0.584	0.625
Mean Atom Type	1.016	0.476	0.473	2.373
SNN (ChEMBL)	0.033	0.371	0.404	0.464
SNN (Real Inhibitors)	0.028	0.168	0.18	0.233
Internal Diversity	0.383	0.864	0.866	0.878
QED	0.57	0.456	0.469	0.518
SA	1.805	3.997	3.57	2.903

In summary, DrugGEN-akt1 is able to generate novel and unique molecules. The choice of training data (AfroCancer vs. ChEMBL) tailors the generated molecules towards different chemical spaces and property profiles. The ChEMBL-trained model seems more generally applicable for discovering AKT inhibitors due to higher similarity to known inhibitors and better performance on a general chemical test set, while the AfroCancer-trained model shows strength in generating highly drug-like and accessible compounds within a more specific natural product-like space.

5.3.2 Metrics for DrugGEN-cdk2. We have performed the same experimentation as previously, but this time the target protein is the CDK protein. The model's performance is assessed under two training regimes (using the AfroCancer Dataset or the ChEMBL Dataset) and then evaluated against two distinct chemical reference sets (smiles_unique_EANPDB.smi and chembl_test.smi). The results of this experimentations are saved in the table 5.3 and we can notice that: Validity scores for the AfroCancer model are relatively high, at 0.823 for smiles_unique_EANPDB.smi and even better at 0.914 for chembl_test.smi, indicating that most generated compounds are valid and demonstrating robust performance on known data. The ChEMBL model also achieves high validity scores (0.667 for smiles_unique_EANPDB.smi and 0.915 for chembl_test.smi), suggesting it can generate valid compounds, especially when tested on its training dataset. Uniqueness remains perfect across all datasets at 1.0, confirming that all generated SMILES strings are unique, which is crucial for drug discovery. In terms of novelty, the model displays perfect scores (1.0) for training data, indicating all generated compounds are new. However, during inference, novelty drops to 0.822 for chembl_test.smi, suggesting some overlap with existing compounds. For real inhibitors, novelty remains perfect at 1.0, which is promising for identifying new drug candidates. The average length of generated SMILES strings is consistent, ranging from 0.65 to 0.703, indicating that the model produces compounds of similar complexity across datasets. Mean atom type scores vary significantly, with higher values in the ChEMBL dataset (1.394 and 3.016) suggesting greater diversity in chemical structures, while the AfroCancer model shows lower values (1.049), indicating simpler structures. Structural similarity (SNN) scores reveal that the AfroCancer model has higher similarity (0.451 for smiles_unique_EANPDB.smi), while the ChEMBL model scores lower (0.295), suggesting that compounds generated from ChEMBL are more distinct. The internal diversity score is high for AfroCancer (0.878) and somewhat lower for ChEMBL (0.639 for chembl_test.smi), indicating a wide variety of chemical structures, particularly from the AfroCancer model. QED (Quantitative Estimate of Drug-likeness) scores are moderate, ranging from 0.455 to 0.495, suggesting that while many generated compounds are valid, they may not all meet optimal drug-likeness criteria. Lastly, synthetic accessibility (SA) scores indicate that the generated compounds may be challenging to synthesize, particularly in the ChEMBL dataset (3.554 for smiles_unique_EANPDB.smi), while lower scores in the AfroCancer dataset suggest potentially easier synthesis.

Table 5.3: Metrics for DrugGEN-cdk2

	AfroCancer Dataset		ChEMBL Dataset	
Metrics	smiles_unique_EANPDB.smi	chembl_test.smi	smiles_unique_EANPDB.smi	chembl_test.smi
Validity	0.823	0.914	0.667	0.915
Uniqueness	1.0	1.0	1.0	1.0
Novelty (Train)	1.0	1.0	0.988	0.987
Novelty (Inference)	1.0	1.0	0.997	0.822
Novelty (Real Inhibitors)	1.0	1.0	1.0	1.0
Average Length	0.65	0.703	0.68	0.689
Mean Atom Type	1.049	1.394	0.968	3.016
SNN (ChEMBL)	0.451	0.242	0.295	0.451
SNN (Real Inhibitors)	0.242	0.203	0.182	0.242
Internal Diversity	0.878	0.639	0.868	0.878
QED	0.486	0.455	0.495	0.542
SA	2.164	2.248	3.554	3.118

Overall, DrugGEN-cdk2 is very effective at generating molecules that are novel compared to known CDK inhibitors, which is a critical success factor. It also produces unique and internally diverse molecular sets. However, the metrics need improvement on AfroCancer model.

5.3.3 Metrics for DrugGEN- No Target. The table 5.4 presents metrics for "DrugGEN-NoTarget," a generative model that, unlike the previous examples (DrugGEN-akt1, DrugGEN-cdk2), is likely trained without a specific biological target in mind. It is designed to generate diverse chemical structures based on the chemical space of its training data (AfroCancer Dataset or ChEMBL Dataset). Performance is evaluated against smiles_unique_EANPDB.smi and chembl_test.smi. Validity scores for the AfroCancer model are relatively high at 0.823 for smiles_unique_EANPDB.smi, indicating that most generated compounds are valid. However, validity drops to 0.61 for chembl_test.smi, suggesting that a significant portion of these compounds may not be valid. In contrast, the validity for the ChEMBL model is better, scoring 0.673 for smiles_unique_EANPDB.smi and 0.903 for chembl_test.smi, demonstrating strong performance on the dataset it was trained on. Uniqueness is high at 0.823 for smiles_unique_EANPDB.smi in the AfroCancer model, and perfect at 1.0 in other configurations, indicating that most generated SMILES strings are unique. The novelty of the AfroCancer model shows that 100% of molecules evaluated against EANPDB are not in the training set, while 11.8% overlap with the chembl_test.smi. For the ChEMBL-trained model, a high degree of novelty is observed, with 91.5% to 98.9% of generated molecules not present in the ChEMBL training set. During inference, the AfroCancer model maintains very high novelty against specific reference sets, whereas 44.6% of the molecules generated by the ChEMBL-trained model overlap with the chembl_test.smi set. Nevertheless, novelty remains perfect (1.0) for real inhibitors, which is promising for discovering new drug candidates. The average length of generated SMILES strings is consistent, ranging from 0.59 to 0.639, suggesting similar complexity across datasets. Structural similarity (SNN) scores are moderate, with the AfroCancer model showing higher similarity (0.562) for smiles_unique_EANPDB.smi, indicating that generated compounds are more similar to known compounds in this dataset than in ChEMBL. Internal diversity scores are high, around 0.865 to 0.874, indicating a wide variety of chemical structures generated in both datasets. The QED (Quantitative Estimate of Drug-likeness) scores are moderate, ranging from 0.497 to 0.575, suggesting that the generated compounds may not always meet optimal drug-likeness criteria. Finally, synthetic accessibility (SA) scores vary, with higher values in the ChEMBL dataset (4.233), indicating that the generated compounds might be more complex and potentially harder to synthesize.

Table 5.4: Metrics for DrugGEN-NoTarget

	AfroCancer Dataset		ChEMBL Dataset	
Metrics	smiles_unique_EANPDB.smi	chembl_test.smi	smiles_unique_EANPDB.smi	chembl_test.smi
Validity	0.823	0.61	0.673	0.903
Uniqueness	0.807	1.0	1.0	1.0
Novelty (Train)	1.0	0.882	0.915	0.989
Novelty (Inference)	1.0	0.993	0.993	0.554
Novelty (Real Inhibitors)	1.0	1.0	1.0	1.0
Average Length	0.639	0.59	0.59	0.627
Mean Atom Type	1.051	2.08	2.08	3.578
SNN (ChEMBL)	0.562	0.481	0.481	0.562
SNN (Real Inhibitors)	0.235	0.172	0.172	0.235
Internal Diversity	0.874	0.865	0.865	0.874
QED	0.497	0.517	0.517	0.575
SA	2.11	4.233	4.233	3.063

The ”NoTarget” model appears to be a robust general-purpose molecule generator. It could be a good starting point for generating diverse chemical matter. If the goal is to find truly novel compounds in specialized spaces, the high novelty is promising, but the synthetic accessibility and validity would need careful consideration and potential post-processing or model refinement.

If aiming for target-specific compounds, this model would likely need to be fine-tuned or used in conjunction with target-prediction models.

5.4 Results of the Bioactivity Prediction by DEEPScreen

After the generation, we verified the ability of the generated molecules to be active on the target protein. To do this, we used the pre-trained DEEPScreen on the AKT protein (Rifaioglu et al., 2020) to predict the bioactivity of the each molecules. By using the following commands:

```
# Navigate to the DEEPScreen directory
cd DEEPScreen2.1/chembl_31

# Run prediction for AKT target
python 8_Prediction.py AKT AKT
```

The bar chart 5.1 shows a right-skewed distribution with a significant number of compounds having low confidence predictions (close to 0). We notice that as the confidence increases, the number of compounds decreases dramatically. Also, a large portion of compounds (over 1000) have low prediction confidence (around 0), indicating uncertainty in their predicted activities. Fewer compounds exhibit high confidence predictions (close to 1), suggesting that while the model can confidently identify some active compounds, the majority remains uncertain.

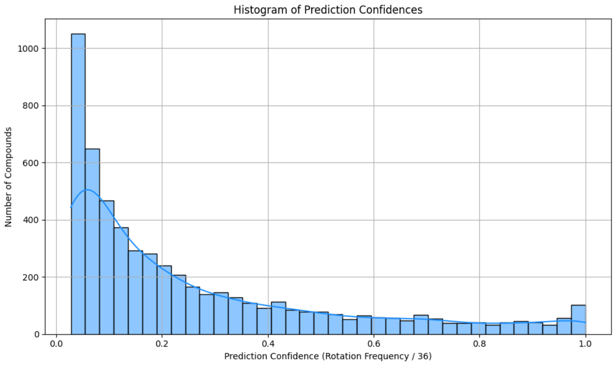

Figure 5.1: Distribution of the predicted bio-activity

5.5 Final Candidate Molecules Selection

To select the final candidate molecules, based on the article (Ünlü et al., 2023) we used 0.85 as the threshold to detect if it is an active molecule or not. We got at the end 324 final candidates. On the figure 5.2 we observed that the overwhelming majority (94.3%) of the compounds are classified as inactive, indicating that the model predicts very few compounds to be active. The small percentage of active predictions (5.7%) may suggest that the model is conservative in its predictions, only classifying a few compounds as active. This distribution may reflect the underlying dataset's characteristics or the model's performance. If the compounds in the dataset predominantly exhibit inactivity, this could be a reason for the high inactive prediction rate. However, the model may need refinement to increase its ability to identify active compounds, potentially involving further training, data augmentation, or feature enhancement. Figure 5.3 shows a sample of 20 of the final selected candidate molecules. We notice a diversity in structural characteristics, which is essential for exploring a wide range of potential interactions with the target proteins.

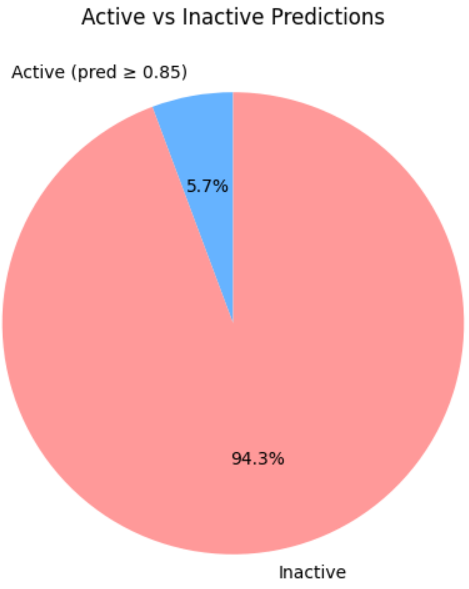

Figure 5.2: Distribution of "Active" and "Inactive" compounds

In this chapter we have presented the differents results we got. We notice that each model exhibits strengths and weaknesses across different datasets. DrugGEN-akt1 shows robust performance in validity and novelty, while DrugGEN-cdk2 performs well in uniqueness and internal diversity. DrugGEN-NoTarget has issues with validity in the ChEMBL dataset, which points to the need for targeted improvements.

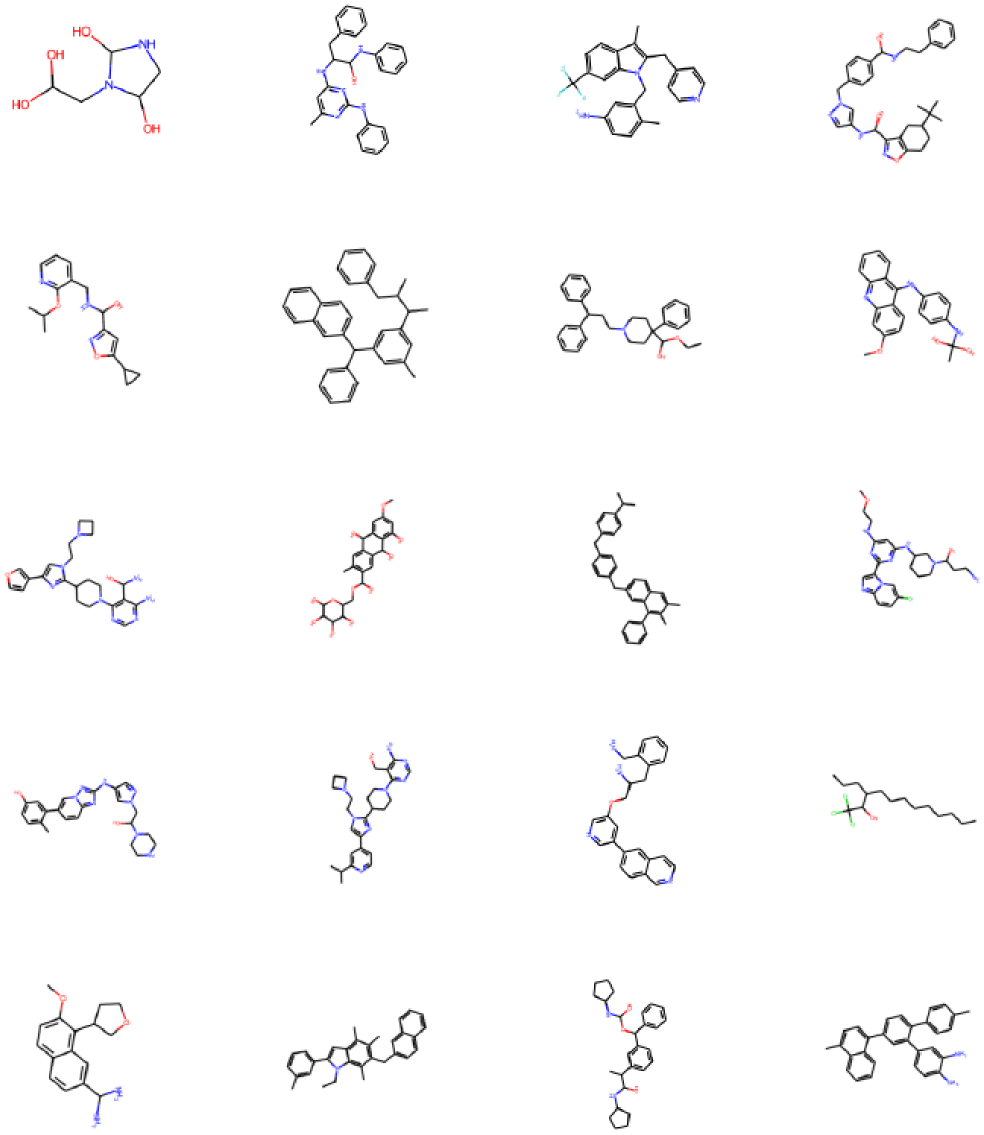

Figure 5.3: 2D Representation of the selected molecules

6. Conclusion and Future Works

In conclusion, our aim in this work is to study the use of generative AI for drug discovery. To do this, we began by defining the key concepts needed to understand the rest of the work: we defined drug discovery, focusing on the different stages of the process, molecule representations and metrics. We then reviewed the various studies carried out on the use of generative AI for drug discovery, as well as the current state of drug discovery in Africa. We then tried to propose how generative AI could be used to improve drug discovery in Africa. To implement this, we used the DrugGEN model, which is based on the Generative Adversarial Networks (GANs) architecture, with the particularity that it integrates transformer layers into its architecture. For the experiments, we used an AfroCancer database comprising 400 compounds from African medicinal plants that have shown in vitro and/or in vivo anticancer, cytotoxic, and antiproliferative activities. Firstly, the DrugGEN model is trained to generate molecules against some particular proteins(AKT and CDK) responsibles from cancer. The results indicate that while the DrugGEN models are capable of generating unique and novel compounds, there are areas for improvement, particularly in maintaining validity and novelty during inference. The models show promise for drug discovery, but ongoing refinement and optimization are essential to enhance drug-likeness, synthetic accessibility, and the generation of truly novel compounds. We then used the DEEP-Screen model to predict the bioactivity of the molecules generated against the AKT protein, selecting molecules with a bioactivity greater than 0.85% as final candidate molecules. The next crucial phase is to move from in silico (computational) predictions to real-world experimental validation. This is where the hypotheses generated by the computer models are put to the test such as Chemical Synthesis and in Vitro Biological Testing. Overall, we notice that across all three models (akt1, cdk2, NoTarget), when trained on AfroCancer, the Novelty (Inference) scores were consistently very high (0.993 to 1.0). This means they are extremely efficient at generating molecules that are new with respect to the evaluation datasets. They are less likely to "waste" computational effort by regenerating known compounds.

As perspectives, we propose to use other drug discovery generative models like ACEGEN and Junction-Tree Variational Autoencoders on the AfroCancer database data in order to get the best generated molecules. We also plan to integrate interpretability into the model as a black box, using heatmaps to show which parts of the molecules are most influential in the model's decision.

References

Richard Kwamla Amewu, Patrick Amoateng, Patrick Kobina Arthur, Prince Asare, Isaac Asiamah, Daniel Boamah, Isaac Darko Otchere, Cedric Dzidzor Amengor, Edmund Ekuadzi, Kelly Chibale, Susan Jane Farrell, Regina Appiah-Oppong, Dorcas Osei-Safo, Kevin David Read, Ian Hugh Gilbert, and Dorothy Yeboah-Manu. Drug discovery research in ghana, challenges, current efforts, and the way forward. *PLOS Neglected Tropical Diseases*, 16(9):1–9, 09 2022. doi: 10.1371/journal.pntd.0010645. URL https://doi.org/10.1371/journal.pntd.0010645.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Tanjim Taharat Aurpa and al. Deep transformer-based architecture for the recognition of mathematical equations from real-world math problems. *Medium*, 2025.

Ajay Bandi, Pydi Adapa, and Yudu Kuchi. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, 15:260, 07 2023. doi: 10.3390/fi15080260.

Leonardo Banh and Gero Strobel. Generative artificial intelligence. *Electronic Markets*, 33(1):63, December 2023. ISSN 1422-8890. doi: 10.1007/s12525-023-00680-1. URL https://doi.org/10.1007/s12525-023-00680-1.

Albert Bou, Morgan Thomas, Sebastian Dittert, Carles Navarro, Maciej Majewski, Ye Wang, Shivam Patel, Gary Tresadern, Mazen Ahmad, Vincent Moens, et al. Acegen: Reinforcement learning of generative chemical agents for drug discovery. *Journal of Chemical Information and Modeling*, 2024.

Patience Chihomvu, A. Ganesan, Simon Gibbons, Kevin Woollard, and Martin A. Hayes. Phytochemicals in drug discovery—a confluence of tradition and innovation. *International Journal of Molecular Sciences*, 25(16), 2024. ISSN 1422-0067. doi: 10.3390/ijms25168792. URL https://www.mdpi.com/1422-0067/25/16/8792.

Ericsson Coy-Barrera, Ifedayo Victor Ogungbe, and Thomas Schmidt. Natural products for drug discovery in the 21st century: Innovations for novel therapeutics. *Molecules*, 28, 04 2023. doi: 10.3390/molecules28093690.

Amol Deore, Jayprabha Dhumane, Rushikesh Wagh, and Rushikesh Sonawane. The stages of drug discovery and development process. *Asian Journal of Pharmaceutical Research and Development*, 7:62–67, 12 2019. doi: 10.22270/ajprd.v7i6.616.

Leonardo LG Ferreira, Josue de Moraes, and Adriano D Andricopulo. Approaches to advance drug discovery for neglected tropical diseases. *Drug Discovery Today*, 27(8):2278–2287, 2022.

Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Franklin Zemo Gamo, Sefirin Djiogue, Charline Florence Awounfack, Adjoffoin Chiara Nange, and Dieudonne Njamen. Neurological disorders: The use of traditional medicine in cameroon. *Traditional Medicine and Modern Medicine*, pages 1–18, 2024.

Tian Gao, Connor W Coley, et al. A survey of generative ai for de novo drug design: New frontiers in molecule and protein generation. *arXiv preprint arXiv:2401.00873*, 2024.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

A. D. Gordon. *Classification*. CRC Press, 2nd edition, 1999.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. 2013.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. doi: 10.1561/2200000056.

Cyril T Namba-Nzanguim, Gemma Turon, Conrad V Simoben, Ian Tietjen, Luis J Montaner, Simon MN Efange, Miquel Duran-Frigola, and Fidele Ntie-Kang. Artificial intelligence for antiviral drug discovery in low resourced settings: A perspective. *Frontiers in Drug Discovery*, 2:1013285, 2022.

Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf.

Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan S. C Lim, and Prudencio Tossou. Gotta be safe: A new framework for molecular design, 2023. URL https://arxiv.org/abs/2310.10773.

Fidele Ntie-Kang, Justina Ngozi Nwodo, Akachukwu Ibezim, Conrad Veranso Simoben, Berin Karaman, Valery Fuh Ngwa, Wolfgang Sippl, Michael Umale Adikwu, and Luc Meva'a Mbaze. Molecular modeling of potential anticancer agents from african medicinal plants. *Journal of Chemical Information and Modeling*, 54(9):2433–2450, 2014. doi: 10.1021/ci5003697.

Pascal Amoa Onguéné, Fidele Ntie-Kang, James Ajeck Mbah, Lydia Likowo Lifongo, Jean Claude Ndom, Wolfgang Sippl, and Luc Meva'a Mbaze. The potential of anti-malarial compounds derived from african medicinal plants, part iii: an in silico evaluation of drug metabolism and pharmacokinetics profiling. *Organic and Medicinal Chemistry Letters*, 4, 2014. doi: 10.1186/s13588-014-0006-x. URL https://doi.org/10.1186/s13588-014-0006-x.

Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. Decimer 1.0: deep learning for chemical image recognition using transformers. *Journal of Cheminformatics*, 13, 08 2021. doi: 10.1186/s13321-021-00538-8.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. URL https://arxiv.org/abs/1506.02640.

Ahmet Sureyya Rifaioglu, Esra Nalbat, Volkan Atalay, Maria Jesus Martin, Rengul Cetin-Atalay, and Tunca Doğan. Deepscreen: high performance drug–target interaction prediction with convolutional neural networks using 2-d structural compound representations. *Chemical science*, 11(9):2531–2557, 2020.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 3rd edition, 2016.

Muhammed Rabiu Sahal, Gaetan Senelle, Kevin La, Tukur Wada Panda, Dalha Wada Taura, Christophe Guyeux, Emmanuelle Cambau, and Christophe Sola. Mycobacterium tuberculosis complex drug-resistance, phylogenetics, and evolution in nigeria: Comparison with ghana and cameroon. *PLOS Neglected Tropical Diseases*, 17(10):1–22, 10 2023. doi: 10.1371/journal.pntd.0011619. URL https://doi.org/10.1371/journal.pntd.0011619.

Divya Saxena and Jiannong Cao. Generative adversarial networks (gans): Challenges, solutions, and future directions. *arXiv preprint arXiv:2101.00001*, 2021.

Vinayak Singh, Dickson Mambwe, Constance Mawunyo Korkor, and Kelly Chibale. Innovation experiences from africa-led drug discovery at the holistic drug discovery and development (h3d) centre. *ACS Medicinal Chemistry Letters*, 13(8):1221–1230, 2022.

Jair L Siqueira-Neto, Kathryn J Wicht, Kelly Chibale, Jeremy N Burrows, David A Fidock, and Elizabeth A Winzeler. Antimalarial drug discovery: progress and approaches. *Nature Reviews Drug Discovery*, 22(10):807–826, 2023.

R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, MA, 1998. ISBN 0-262-19398-1. Hardbound, $40.00.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Logan Ward, Jenna A. Bilbrey, Sutanay Choudhury, Neeraj Kumar, and Ganesh Sivaraman. Benchmarking deep graph generative models for optimizing new drug molecules for covid-19. Feb 2021. URL https://www.example.com/path/to/preprint. Preprint.

L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, and M. H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Atabey Ünlü, Elif Çevrim, Ahmet Sarıgün, Hayriye Çelikbilek, Heval Ataş Güvenilir, Altay Koyaş, Deniz Cansen Kahraman, Abdurrahman Olğaç, Ahmet Rifaioğlu, and Tunca Doğan. Target specific de novo design of drug candidate molecules with graph transformer-based generative adversarial networks, 2023.