From Generation to Validation: A Combined RLGAN and JTNNVAE Approach for Novel Drug Discovery

Majd Barghouthi*

Department of science and technology, IE university

Bachelor of computer science and artificial intelligence

Robert Polding

May 3rd , 2024

Contents

Abstract

This six-month research project seeks to integrate Reinforcement Learning-based Generative Adversarial Network (RLGAN) and Junction Tree Variational Autoencoder (JTNNVAE) for innovative drug discovery. The goal is to generate new drug molecules that are chemically viable and biologically active, meeting specific molecular constraints. De novo drug design, the creation of new molecular structures, is often slow and costly using traditional methods. This project proposes a novel approach using RLGAN for its adaptive control and JTNNVAE for molecular analysis, aiming to enhance efficiency and reduce the time and costs of drug discovery.

Index Terms- De novo; drug discovery; JTNNVAE; RLGAN, MolGAN

0. Introduction

Finding new drugs requires not just a significant amount of time and money, while mitigating high failure rates in the winds of uncertainty, like navigating a complex maze. To overcome the obstacles in the way, this complex task demands more innovative approaches. Conventional drug discovery approaches, while crucial in many medicinal advances, are beset by their linear approaches, large lead times and the unpredictable behavior of biological systems. Considering this, there has been growing interest in using artificial intelligence (AI) to usher in a new era of drug discovery that will be characterized by increased success rates, shorter development cycles, and greater efficiency.

In the realm of AI-driven drug discovery, promising frontiers include Reinforcement Learning (RL)-based Generative Adversarial Network (RLGAN) and Junction Tree Variational Autoencoder (JTNNVAE). The RLGAN is adept at adjusting and acquiring knowledge in dynamic environments, making it a valuable tool for creating new drug molecules that meet specific chemical and biological criteria while the JTNNVAE has become essential for examining molecular structures and accurately predicting compound properties and interactions. The objective of this research is to explore the potential of combining RLGAN and JTNNVAE in the context of novel drug discovery. By examining the current state of research, technological advancements, and the implications of these technologies, we aim to highlight their significance, challenges, and prospects in enhancing drug discovery processes. This review will provide a comprehensive overview of the integration of these technologies, underscoring their potential to revolutionize the field and contribute to the acceleration of bringing new, effective drugs to market.

1. Literature Review

The exploration of AI and machine learning (ML) has resulted in numerous advancements in the search for new drugs. As AI and ML methods navigate this voyage with the help of complex algorithms, the days of depending only on trial and error and extensive screenings are over. This significant step forward addresses long-standing issues that have plagued the drug discovery process. Specifically, the route from concept to pharmacy shelf is not only notoriously long, typically lasting more than a decade, but also incredibly expensive, with expenses exceeding $2.6 billion for just one breakthrough. This sections explores the advancement and history of the most prominent and relevant ai and ml models in the journey of drug discovery.

1.1 Advances in De Novo Drug Design

AI and ML integration have heralded a paradigm shift in drug design. Using deep learning models, such as GNNs, has played a key role in the domain and have drastically contributed to a new paradigm allowing efficient traversal of huge chemical state spaces—estimated to have a size of up to 10^{60} molecules that can possibly be drug-like. This area presented an insurmountable challenge for conventional methodologies, where the traditional methods used failed to improve the efficiency of the drug discovery process. In contrast, GNNs improved the efficiency of the whole process and simultaneously discovered previously unknown molecular structures; in this case, those related to halicin, a powerful antibiotic.

The combination of such models with deep learning-based ones, like the Variational Autoencoders (VAE) with Generative Adversarial Networks (GAN) and RL, has moved the field towards inventive methodologies for drug discovery. Large libraries allow such models to be proficient in learning the distribution of biochemical properties in training sets as they are fit for the specific targets. While they offer great potential, challenges remain numerous, from improving scaffold diversity to overcoming the

large computation of deep learning-based models. This, of course, has the potential to revolutionize the process of discovering drugs in a way that goes far beyond their computational capacity. Among some latest developments in natural language processing, there lie Chemical Language Models (CLMs), that contribute to improving this operation of generating new drug molecules at a fast pace and, in this way, help speed up the very first steps of drug discovery. Such integration brings these disciplines ever closer together and further improves model interpretability and validation.

While AI and ML have been outstanding tools in the field, it has also dealt with some very important setbacks, especially on issues of computational power and synthetic possibility of generated compounds. Future lines of development will focus on the support in scaffold diversity, reduction of computational demand, and practical applicability of the generated compound in real conditions. While high marks don't precisely translate to the actual synthetic feasibility of the desired compounds or biological activity, it does represent a big leap in the translation of computational models into concrete drug development. Another hurdle at hand is ensuring that the desired AI-generated molecules can indeed be synthesized in a real laboratory and demonstrate the desired effect on biological targets with no unexpected symptoms. Strategies used to bridge the gap between computational predictions and experimental observation focus on integrating AI predictions with experimental validation, where computational insights would convert into actual outcomes. AI and ML technology incorporation in drug design ensures a high leap in the search for new therapeutics. AI and ML have addressed the complications and challenges that traditional drug discovery methods pose and, therefore, opened new ways to innovate, be efficient, and effective in the pharmaceutical world.

1.2 Deep Learning in Drug Discovery

Deep Learning (DL) models have offered a powerful toolset for drug discovery to extract useful information from big datasets and discover complex patterns, which may not be easily spotted by human experts or even using traditional computational approaches. These models are particularly notable for their generative capabilities, hence having significantly facilitated and paved the way for the creation of new compounds.

It is not surprising that DL models have found their entrance into drug discovery, with Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Reinforcement Learning (RL) models taking up the rein of this transformative wave. GANs have become architects of novelty, crafting molecular structures that mimic reality but remain novel. This is executed through an interplay with a generator, which generates new molecules, and a discriminator, which acts as the de facto judge for those same molecules. The role of the VAE is to master the art of encoding data, including the very essence of molecules, into a latent space from which future discoveries can be decoded. On the other hand, RL models take a far more strategic view: they modify molecular designs so that each change is evident and advantageous, which are done through the iterative nature of such models.

That said, the road was not free of bumps. The data-hungriness, computational demand, and enigmatic decision processes of DL models give birth to huge challenges. The leap that researchers needed to fill ranged from the digital blueprints of drugs to the synthesis of such drugs, thus demanding innovations which would bridge the transition of virtual creations to tangible hypotheses.

1.3 Generative Chemistry

Generative chemistry represents a new transformation in drug discovery through a change created using deep learning techniques for the invention of how novel compounds are conceived, optimized, and evaluated. This revolution lies at the heart of the generative models that show exceptional efficiency in

the complex, multi-dimensional space of chemical structures to propose new molecules that might display therapeutic activity.

1.3.1 Deep Learning Techniques in Generative Chemistry

These are further extended with the utilization of deep learning in generative chemistry using models like GANs, VAEs, and RL algorithms. These models learn from existing molecular datasets to generate novel molecular structures, often predicting their properties and biological activities of never-seen-before compounds.

• *GANs in Generative Chemistry*: GANs have become powerful in generating new molecules via a competitive framework between two neural networks that simulate a process like natural selection. This approach supports a novel generation of molecules that, at high levels, are optimized for the desired properties along with them.

• *Molecular Generation with VAEs*: VAEs provide an alternative paradigm that encodes molecules into a latent space where not only efficient sampling occurs but manipulations as well. Thus, the exploration of chemical spaces that were infeasible via traditional methods are now accessible and paves the way to finding the novel structure with the desired pharmacological profiles.

• *RL in Optimizing Compound Design*: RL algorithms stand out with their intrinsic properties allowing iterative refinement of molecular designs towards a specific goal, such as maximizing drug-likeness or binding affinity to a target protein. This process follows the same looping drug design process, paving the way for systematic improvements in compound properties.

1.3.2 Generative Chemistry Breakthroughs

The impact of DL on generative chemistry has produced some major breakthroughs.

- *Fast discovery of new antibiotics*: This has allowed rapid progress in the discovery of new antibiotics, such as the discovery of Halicin: observed from a large chemical library screening which demonstrated activity against bacteria with multi-drug resistance.

- *Designing Targeted Therapeutics*: Deep-learning-driven generative chemistry has enabled the design of targeted therapeutics by making it possible to produce molecules with high specificity and affinity to be targeted on disease-relevant targets, sparing off-targets, therefore decreasing side effects, and eventually optimizing the therapeutic effect.
- *Drug Repurposing*: It predicts the enhanced ability of existing molecules for new indications and fills the gap in the drug arsenal at lower costs.

1.3.3 Generative Chemistry Problems

Despite these advancements, the application of deep learning in drug discovery is not without challenges. There are large datasets required for an effective learning process, substantial computation resources to train models, and the "black box" nature of DL models, which is a challenge in understanding how predictions are made. Moreover, a big challenge remains to be the ability to synthesize the compounds predicted by DL in the laboratory, which will greatly close the gap between predictive computational work and experimental work. This has opened numerous avenues for rapid development, such as the discovery of new antibiotics and the design of targeted therapeutics that have high specificity and affinity toward disease-related targets. However, some of the roadblocks, that generative chemistry still needs to go through, include data quality, data availability, synthetic accessibility of the generated molecules, regulatory issues, and developing a better interpretability of DL models for making truthful predictions. With solutions to these challenges and further improvement of DL methods, the area of drug discovery is opening new perspectives in the development of effective and targeted therapy, thus confirming the huge potential of AI to change pharmaceutical R&D.

1.4 Chemical Language Models in Drug Design

Chemical Language Models are the merger of NLP methods with chemical informatics, a novel approach that conceptualizes molecules at the linguistic plane considering molecules as sentences and the atom or constituent groups as words. This perspective allows for the use of NLP methods for predictions of chemical attributes and interactions in the crafting of new drug entities. Adapting NLP techniques for use in a chemical language model includes:

• *Sequence-to-Sequence Models (Seq2Seq):* Seq2Seq models used to predict reactions or convert one chemical entity into another exploit chemical representations (like SMILES strings) (Weininger,1988) that are sequential and make it possible to perform translation, as in the case of text-to-text tasks.

• *Transformers and mechanisms of attention:* Inspired by language translation, transformers use attention mechanisms to weigh the parts of molecules when predicting their properties or creating a new compound. It is not surprising that such methods become increasingly useful in the sense of acquiring insights into the relationships and dependencies that exist among the features of complex molecular structures.

• *Graph Convolutional Networks (GCNs):* Although not the NLP approach to itself, GCNs model molecules as graphs (i.e., sentences with the linking of words) and have been used in conjunction with NLP models for the better 3D molecular and biological function representation.

The impact of these models is particularly evident in the generation of novel drug molecules. For example, using models, such as BERT and GPT, adapted to chemical sequences in such prediction makes it possible to suggest existing drugs for reuse in cases of new therapeutic uses (Niu, Y.,2020). These models were invaluable to quickly repurpose the existing potential drug among other applications in view of breaking health emergencies like the spreading of the coronavirus and the ensuing COVID-19 pandemic. In addition, the very marked use of chemical language models in drug design de novo has sped

the early steps of drug discovery, allowing the ability to design molecules that have optimized pharmacokinetic profiles or low toxicity from scratch.

Despite the development of NLP techniques in the case of chemical language models, few key challenges do remain open for further research in drug design. The challenge of representing the vast spread of complex biological information of molecules so that these can be understood and processed by the models is the current problem. Certainly, that is very promising research into graph-based models and hybrid approaches with structural data along with NLP.

1.5 Target Selectivity in Drug Design

In the world of drug design, zeroing in on the right target with precision is more than just a step; it's a cornerstone for crafting treatments that strike the perfect balance between potency and safety.

1.5.1 The Essence of Selectivity in Drug Design

Imagine a drug that knows exactly where to go and what to do, with minimal detours. That's the beauty of selectivity. It's about a drug's finesse in homing in on its intended biological target while steering clear of others. This discernment is pivotal for a few reasons: It slashes the chances of unwanted side effects by avoiding unintended interactions, ensuring the drug does its job without causing unnecessary harm. It amplifies the drug's therapeutic impact, concentrating its action where it counts and making it more effective in smaller doses. It polishes the drug's safety profile, reducing toxicity risks by limiting its interaction sphere.

1.5.2 Crafting Selectivity:

Getting to that level of selectivity is a very complex task that involves the latest technology, the deepest biological insight, with a little touch of creative flair.

- *Structure-Based Drug Design (SBDD):* It's like creating a key for a lock. Knowing the 3D structure of a target protein enables scientists to play with molecules that fit perfectly, perhaps decreasing chances they will latch onto unintended targets.

- *High-Throughput Screening (HTS):* Imagine going through a treasure trove of compounds, at breakneck speed, to spot those that hit the mark against a particular target, while making sure they don't bind somewhere else by mistake.

- *Pharmacophore Modeling:* This is the marking of essential features that a molecule must provide to interact with a target. The designed guide compounds are such that these critical attributes are fully endowed and at the same time selective.

- *Molecular Dynamics simulation:* Virtual rehearsals representing how a drug candidate and its target would interact over time to help predict and refine selectivity.

- *Ligand Efficiency Indices:* Efficient binding trends to be carried out with selectivity, meaning the binders are able to focus on the effects where needed with much more precision.

- *Genetic and cellular approaches:* Tools like CRISPR-Cas9 and RNAi have made it possible not only for the target validation in a disease but also to predict the impact of the modulation of a certain target and aid in building selective drugs.

Achieving selectivity is not only a technical victory but also a fundamental step toward the procurement of potent drugs that are safe. It guides the arsenal to act with precision, without collateral damage and to the maximum benefit for health. By combining technology, biology, and innovation, the

next generation of targeted drugs is being designed, with all the steps followed by the purpose of selectivity.

1.6 GANs and VAEs in Drug Discovery: Pivotal Roles and State-of-the-Art Applications

As mentioned previously, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have significantly impacted the field of drug discovery, offering innovative approaches to design novel compounds with desired properties. Both models leverage deep learning techniques to generate new molecular structures, but they operate under different principles and are suited for distinct aspects of the drug discovery process.

1.6.1 GANs: Enhancing Molecular Diversity in Drug Discovery

GANs utilize a dual neural network system—a generator and a discriminator—in a game-theoretic framework. The generator produces new molecular structures while the discriminator evaluates their quality or authenticity. This method proves particularly effective in generating chemically diverse and realistic molecular structures. At Insilico Medicine, for example, researchers have developed a "Molecular GAN"[1] that accelerates the design of novel molecules with desired characteristics, substantially reducing the time needed to discover viable drug candidates. Furthermore, conditional GANs have been used to optimize the pharmacokinetic properties of molecules, such as solubility and permeability, demonstrating their utility in enhancing drug-like characteristics without compromising molecular activity.

[1] (Medicine, 2023)

1.6.2 VAEs: Probabilistic Modeling in Molecular Design

Variational Autoencoders employ a probabilistic approach where they encode input data into a latent space and subsequently decode it to generate new molecules. This technique is primarily utilized for generating molecules and predicting molecular properties. For instance, a study in the Journal of Medicinal Chemistry[2] has demonstrated the use of a VAE to predict ADME (Absorption, Distribution, Metabolism, and Excretion) properties from the representations in the latent space. In another application, VAEs have been employed in iterative refinement processes where an initial set of molecule structures is progressively improved based on their predicted properties.

1.6.3 Comparative Insights and Hybrid Models

While GANs are adept at generating novel and diverse molecular structures, VAEs excel in the precise control and optimization of molecular features. The synergy between these models is evident in hybrid approaches, which enhance the interpretability of latent spaces through VAEs and leverage the robust generative capabilities of GANs for efficient exploration and optimization of chemical spaces.

1.6.4 State of art and Chosen models

When it comes to the state of art of such models in drug discovery, it is apparent that more than one approach has been taken and with those approaches come their advantages and limitations. This section will talk about the GAN and VAE models that outperformed and stood out from the others, the idea behind them and the proposed hybrid model. This summary dives into the specifics, advantages, and limitations of these models, focusing on four distinct studies: Mol-Zero-GAN (Aphikulvanich & Pornputtapong, 2023), GANsDTA (Zhao et al., 2020), MoVAE(Society for Industrial and Applied Mathematics eBooks, 2023), and Disentangle VAE(2022, Disentangle VAE for Molecular Generation). And how they compare JTVAE(Jin, 2019) and MOLGAN(De Cao & Kipf, 2022) which are the models chosen to further research.

[2] (Dara et al., 2022)

Mol-Zero-GAN exemplifies the strategic application of GANs in generating drug candidates tailored to specific protein targets. This model employs Bayesian optimization for weight optimization using singular value decomposition, optimizing properties like druglikeness, synthesis accessibility, and binding affinity. The ability of GANs to generate novel data points makes them particularly suitable for exploring uncharted chemical spaces, thereby enhancing the diversity of molecular structures.

The model's dependency on the quality of existing pharmaceutical data limits its applicability in scenarios with inadequate data. The complexity introduced by advanced optimization techniques such as SVD requires precise calibration to maintain generalizability and avoid overfitting. Furthermore, the high computational demands may restrict its use in less resource-endowed settings.

GANsDTA introduces a semi-supervised learning model using dual GANs integrated with a regression network to predict drug-target interactions. This model efficiently utilizes both labeled and unlabeled data, enhancing its ability to operate effectively even with sparse datasets. The use of GANs to mimic real data distributions is crucial in domains where actual interaction patterns are partially unknown. The performance of GANsDTA heavily depends on the quality of input data. Its complex architecture, combining dual GANs and a regression network, poses challenges in tuning and scalability, particularly when adapting to larger or diverse datasets.

MoVAE leverages the strengths of VAEs in learning latent manifold representations, crucial for accurately modeling molecular structures. This model enhances the traditional VAE setup by individually encoding and decoding nodes and edges, which simplifies the handling of molecular graphs and facilitates the generation of chemically accurate structures through adversarial training. Despite its advancements, MoVAE struggles with the complete validity of generated molecules and could become computationally intensive as dataset complexity increases. The model's efficacy is also tightly coupled with the quality of the training data.

Disentangle-VAE emphasizes controlled generation of molecules, incorporating an extractor module within the VAE framework to manipulate specific molecular properties. This approach enhances the model's ability to tailor molecule generation to desired characteristics, making it highly relevant for targeted drug design. The integration of an extractor module increases the training complexity, requiring sophisticated optimization techniques. Similar to other models, the performance of the Disentangle-VAE is dependent on the quality and diversity of training data, with scalability remaining a challenge. When comparing these models, GANs demonstrate a strong capability for generating novel and diverse molecular structures, making them ideal for applications requiring high variability. On the other hand, VAEs offer more control and precision in the generation process, suitable for tasks demanding specific molecular properties.

Both model types face challenges related to data dependency and computational demands, reflecting a common issue in applying advanced machine learning techniques to drug discovery. However, their ability to push the boundaries of traditional drug discovery methods by enabling rapid screening and optimization of molecular candidates underscores their transformative potential in pharmaceutical research. As these models evolve, their integration into drug discovery workflows promises to enhance both the efficiency and creativity of the drug development process.

1.7 Comparative Analysis of MolGAN and JT-VAE Against Previous Generative Models in Drug Discovery

In the evolving field of drug discovery, the development and refinement of generative models, like MolGAN and JT-VAE, have set new standards for molecular generation. These models are particularly notable for their innovative approaches to handling molecular graphs and ensuring chemical validity, aspects that can sometimes be challenging in other models such as Mol-Zero-GAN, GANsDTA, MoVAE, and Disentangle VAE.

1.7.1 MolGAN: An Implicit Generative Model for Small Molecular Graphs

MolGAN introduces a significant innovation by generating small molecular graphs directly without the need for likelihood estimation, a complex and computationally intensive process. This direct approach to handling graph-structured data allows MolGAN to operate efficiently in generating small molecules, an area where previous models may require conversions from SMILES strings or other vector representations. Figure 1 illustrates the dataflow and processes of the MolGAN pipeline.

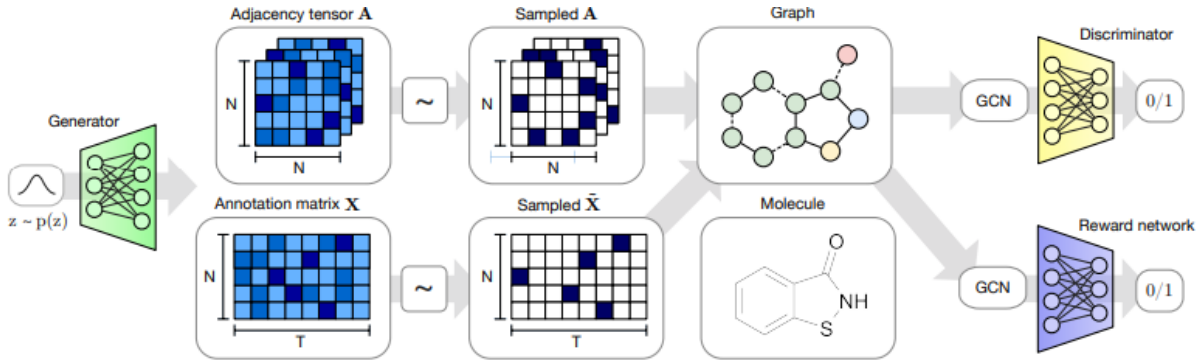

Figure 1: Outline of MolGAN. From left: the generator takes a sample from a prior distribution and generates a dense adjacency tensor A and an annotation matrix X. Subsequently, sparse and discrete A˜ and X˜ are obtained from A and X respectively via categorical sampling

Moreover, the integration of reinforcement learning to enhance the practical utility of generated molecules for chemical synthesis represents a step forward in aligning model outputs with real-world applicability. This aspect of MolGAN is particularly beneficial over models like Mol-Zero-GAN, which, while effective in targeting specific proteins, do not inherently optimize for synthesis feasibility. A more simplified representation of the MolGAN can be seen below in figure 3 where the iterative nature of the GAN coupled with RL show how the continuous improvement of generated molecules occurred while maintaining the validity, novelty and uniqueness of the molecules. Its ability to produce nearly 100% valid compounds as demonstrated on the QM9 chemical database makes it exceptionally reliable for generating feasible molecular structures. As can be seen in the figure below, the MolGAN model's performance exceeds the usual complex data hungry models.

Algorithm	Valid	Unique	Novel
CharacterVAE	10.3	67.5	90.0
GrammarVAE	60.2	9.3	80.9
GraphVAE	55.7	**76.0**	61.6
GraphVAE/imp	56.2	42.0	75.8
GraphVAE NoGM	81.0	24.1	61.0
MolGAN	**98.1**	10.4	**94.2**

Figure 2: Comparison with different algorithms on QM9.

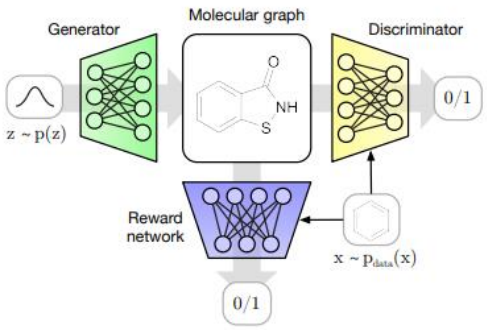

Figure 3: simplified representation of MolGAN

1.7.2 JT-VAE: Junction Tree Variational Autoencoder for Molecular Graph Generation

The JT-VAE's methodological breakthrough lies in its two-phase molecular graph generation process which guarantees the chemical validity of each molecule throughout the generation process. Unlike traditional VAEs such as MoVAE and Disentangle VAE, which may generate invalid molecular structures, JT-VAE ensures that every generated structure adheres to chemical rules by constructing molecules from valid substructures. This model's approach to encoding and decoding complex molecular graphs, shown in figure 4, also offers an improvement over methods that rely on simpler, less structured data representations. By employing a graph message passing network, JT-VAE effectively captures and

reconstructs detailed molecular structures, which is a significant advancement over simpler VAE or GAN-based models.

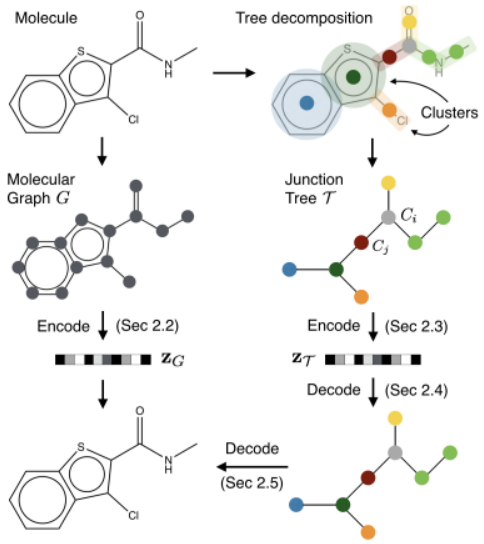

Figure 4: Overview of the JTNNVAE method

Method	Reconstruction	Validity
CVAE	44.6%	0.7%
GVAE	53.7%	7.2%
SD-VAE	76.2%	43.5%
GraphVAE	-	13.5%
Atom-by-Atom LSTM	-	89.2%
JT-VAE	**76.7%**	**100.0%**

Figure 5: JTVAE scores based on such benchmarks and against other models.

JT-VAE is the ideal choice for projects that require the generation of complex molecules where chemical validity is non-negotiable. Its structured approach to generating valid subgraphs before assembling them into complete molecules provides a robust framework for ensuring the feasibility and efficacy of the generated compounds. This makes JT-VAE particularly suitable for large-scale drug discovery projects that aim to explore new chemical spaces or optimize molecules with specific properties.

Both the MolGAN and the JT-VAE offer distinct advantages over models like Mol-Zero-GAN, GANsDTA, MoVAE, and Disentangle-VAE, primarily through their innovative handling of molecular graphs and enhanced focus on chemical validity. Choosing between these models largely depends on the

specific requirements of the drug discovery project, such as the need for high throughput, specific target properties, or computational resource constraints. MolGAN is preferable for projects focused on small molecules and high validity, while JT-VAE excels in projects requiring the generation of complex, large-scale molecular structures with guaranteed chemical accuracy. These attributes position both models as highly effective tools in the next generation of computational drug discovery. With the integration of both models, we can say that a pareto front[3] (Navon et al., 2021) effect might be introduced, meaning that the trade-offs that are considered for both of these models might be reduced if we integrate them into each other.

[3] (Navon et al., 2021)

2. Methodology

Choosing the combination of Reinforced Generative adversarial Network (RLGAN) and Junction Tree Variational Autoencoders (JTNNVAE) for drug discovery is a very compelling choice given the capabilities and results demonstrated by RLGAN, in addition to, at the same time, the inherent strengths of JTNNVAEs in modeling the molecular structure. Integration of these technologies brings about comprehensive coverage to overcome the limitations seen in other domains of drug discovery while increasing target selectivity and embedding detailed chemical which will be discussed further later in this paper.

2.1 The Integration of RLGAN and JTNNVAE in Drug Discovery

The architecture is a new type of deep neural network (DNN) structure under the Generative Adversarial Network (GAN) paradigm combined with Reinforcement Learning and the Junction Tree Variational Autoencoder. Its generator employs the differentiable neural computer (DNC) and a JTNNVAE to augment the generating capabilities. This enables the RLGAN to efficiently come up with new chemical structures for small-molecule drugs compared to the generation of new, unique structures that will pass medicinal chemistry filters (MCFs), be of desirable chemical property, and cover the distribution range of chemical descriptors within the training set.

JTNNVAEs, on the other hand, are fantastic for the modeling of data naturally represented as graphs and hence find good applications in chemical informatics, where graphs represent molecules and their constitutive atoms (nodes), connected by bonds (edges). JTNNVAEs are extremely suitable to capture complex relationships among atoms and properties associated with molecular behavior, making it ideal for predicting the behavior of molecules and designing new compounds with some characteristics prescribed in advance. So, with the integration of such models comes its advantages and limitations, however from an objective perspective its advantages far exceed its limitations which is further discussed below.

Enhanced Generative Capabilities: The RLGAN architecture demonstrated excellent performance not only in generating but also in predicting new small-molecule organic structures when using a differentiable neural computer (DNC). The second major improvement is that RLGAN outperforms its predecessors in finding unique structures passing medicinal chemistry filters (MCF), Muegge criteria, having high Quantitative Estimate of Drug-likeness (QED) scores. A capability that directly answers the need for innovative compounds in drug discovery with the diversity and novelty of chemical space exploration. So, the addition of technology like the JTNNVAE would improve such architecture drastically.

Accurate Representation of Molecular Features: The ability of RLGAN-JTNNVAE in the generation of compounds, represented exactly in the distributions of key chemical features and descriptors such as molecular weight (MW), logP, Topological Polar Surface Area (TPSA), and lengths of the SMILES strings in the training dataset, indicate the potential capability of the compound in creating viable drug candidates. This precision in mimicking training data distributions is important for developing molecules with properties of interest. JTNNVAEs do well to capture the relations in the molecule deemed graph based. This structural understanding complements the generative capability of RLGAN to ensure the produced molecules will be novel and structurally sound, with biological relevance. Utilizing JTNNVAE would further enrich the integration of information on the chemical structure and biological information, making drug design more informed and effective.

Addressing Drug Discovery Disciplines: To the unaided eye, these look like 'Drosophila Experiments.' They are the early phases of today's in silico modeling and combinatorial organic synthesis-driven modern drug design and development, evaluated using high-throughput biological screening (HT Together with its in silico modeling capability, RLGAN directly serves these critical scientific disciplines in their relentless pursuit of discovery of diverse small-molecule drug molecules with high structural diversity. The JTNNVAE and RLGAN combination promises an improvement in target selectivity. To model accurate drug-target interactions, JTNNVAE is the tool of choice, while RLGAN has generative capabilities in creating optimized molecules. This synergy could yield the development of drugs with

better activity, selectivity, and a desirable pharmacokinetic profile; therefore, this would solve a fundamental challenge in drug discovery.

Integration of RLGAN and JTNNVAE in drug discovery is not the theoretical improvement here but a very practical answer based on empirically collected facts. Taken in conjunction with the modelling strengths of JTNNVAE's structural and relational capabilities with RLGAN, demonstrates a powerful innovative way for generating novel, diverse, and biologically relevant compounds. This new integration allows one to push the boundaries of the current state-of-the-art models by offering a robust framework for the development of new therapeutics with optimized target selectivity and comprehensive embedding of chemical and biological information. Therefore, this means that the synergy of RLGAN with JTNNVAE could actually put up a new standard relating to the efficiency, accuracy, and efficacy of drug discovery processes.

2.2 Data Collection and Preprocessing: Building the Foundation

In the landscape of modern drug discovery, the first stage of the pipeline, data capturing and preprocessing, could be regarded as the most crucial. Translating abundant molecular structures and biological data into novel antibiotics is the field of innovative computational methodologies. To this end, the Comprehensive Antibiotic Resistance Database (CARD)[4] and ZINC15[5] are utilized as the invaluable basis for the sourcing of these necessary data. CARD is widely known for its extensive library comprising antibiotic resistance genes, small molecules and the relevant molecular structures that shed light on bacteria’s ability to elude the existing drug mechanisms. Conversely, ZINC15 represents a free database for virtual screening offering more than 230 million commercially available compounds in compound, ready-to-dock, 3D formats. That database could, thus, be interpreted as an abundance of molecular structures transcribed in Simplified Molecular Input Line Entry System (SMILES) strings.

It is quite evident that the importance of acquiring data from either CARD or ZINC15 cannot be overstated. The tunnel vision of CARD concerning resistance mechanisms opposed to the expansive

[4] Alcock et al. 2023. CARD 2023

[5] Sterling and Irwin, J. Chem. Inf. Model, 2015

landscape of chemical entities availed by ZINC15 is inherently different but considered together go a long way in providing a thorough view of the opponent (pathogenic bacteria) and its possible weapons (drug molecules).

However, the integration of relevant biological data points concerning these molecules, even though positioned much later in the model architecture, is still vital for ensuring the resulting computational forecast follows ultimate biological resonance, and thus is theoretically sound as well as practically implementable. Lastly, the most complicated part of the system workflow is characteristic of the elemental conversion from SMILES strings to graph and tree representations. Consequently, the conversion from SMILES into a graph and then a tree is conducted very carefully using RDKit as an open-source cheminformatics toolkit that will be used with graph convolutional networks (GCNs). RDKit can be used to de-serialize SMILES strings into a more computationally analytically useful format such as a molecular graph, where atoms are described as nodes, and bonds between them are edges.

The graph is then interpreted on the structures by use of GCNs, as they are especially used to capture the vast and very complex relationships and patterns in molecular structures, an ability that would be very much needed when trying to decipher the biological activity pre-imposed by the molecules. This is not in any way a technicality but the essential way of ensuring that the comprehensiveness and complexity of molecular data are preserved and “enhanced”. This means that the molecular structure graph and tree data representations of these types of information will be open to very sophisticated computational analysis and the application of very modern machine learning models. Thus, the dataset gets “enhanced” in a way that its original data becomes transformed from a mere heap of strings into a dynamic, linked network of chemical entities. This, with much more detail on their actual interactions, forms good stepping-stones to advance drug discovery. Giving attention to detail to ensure good data quality through processing will make sure that the analysis that follows is, thus increasing the odds of novel and potent antibiotic discovery.

2.3 JT-VAE Encoder: The Heart of Molecular Innovation

The Junction Tree Variational Autoencoder is a major computational drug discovery advancement. Essentially, the JT-VAE receptor's role is to simplify and condense vast amounts of molecular knowledge. This simplification is critical since it converts the rich structural information of chemical elements encoded as documents into a condensed vector mat. In this perspective, Graph Neural Networks are critical because they are the sole purpose of these visits to seize and process this extensive molecular character's capacity, including inherently graph data. The JTNNVAE examines these compositions by considering the nodes (atoms) and the corners (bonds), effectively learning molecular topology and characteristics directly from the graph edition.

However, JT-VAE's integral validity goes far beyond mere data compression. Encoding molecular graphs into a latent space facilitates a more efficient exploration of the chemical universe. The latent representation serves as a compact, yet information-rich summary of the molecule's structure and properties, enabling the generation of new molecular structures through the decoder component of the autoencoder. This ability to explore vast chemical spaces efficiently is particularly crucial in the search for novel antibiotics, where the discovery of unique structures with potential antibacterial activity is of paramount importance.

This process of encoding and the subsequent generation of novel compounds embodies the modern approach to drug discovery, where computational tools are leveraged not just for analysis but as active participants in the creative process. By enabling the exploration of uncharted chemical spaces with unprecedented efficiency, the JT-VAE encoder plays a pivotal role in the discovery of novel antibiotics, offering hope against the growing threat of AMR. Its impact extends the boundaries of traditional drug discovery methods, promising a future where the rapid, informed generation of potential drug candidates becomes the norm, significantly accelerating the pace at which new treatments become available

2.4 Latent Space Exploration: Unveiling Novel Compounds

The Junction Tree Variational Autoencoder's operation is centered on the investigation of the latent space, and it is here that the promise of revolutionary drug discovery becomes a reality. In the latent space, encoded chemical structures that have been reduced to a manageable form are dematerialized and reimagined, resulting in new compounds and possibilities. The exploration is carried out using representation learning approaches, which have made raw molecular data suitable for training. The model discovers these representations and generates novel molecule configurations, some of which already exist but have not been synthesized.

The significance of this method of investigation in the realm of drug development is enormous. The latent space is an environment for synthetic and biological invention that is free of the constraints of chemical synthesis. There, the model offers novel chemical entities with distinct features, which could lead to remarkable advances in antibiotic discovery. The ability to model these novel things, derived from an abstract space rich in information, propels the latent space to the forefront of radically revolutionary drug discovery efforts.

2.5 Decoding Molecular Structures: Bridging the Gap

The transition from the theoretical latent space exploration to the concrete reality of molecular synthesis is in the decoding phase. In this stage, the Junction Tree Variational Autoencoder (JT-VAE) assumes a pivotal role, tasked with performing the reverse of the encoding phase by translating the abstract, latent representations of potential drug molecules, back into their structured molecular systems. This process transcends mere reversal; it is a generative act that actualizes the potential discovered in the latent space, making previously theoretical possibilities concrete and testable. As it is done by having the clusters of nodes and edges, reconstructing the said latent space while maintaining the exploration part by calculating which is the next most probable cluster that connects with the existing ones, and ensuring that

it is a viable connection. After all of that, the newly constructed and explored molecule will be reconstructed back to the graph representation as seen in fig. 4.

The decoder's role is crucial in forming a bridge between the computational model and actual therapeutic solutions. By reconstructing the encoded molecular graphs and junction trees, the JT-VAE ensures that the generated molecules transition from being mere digital artifacts to viable compounds ready for synthesis and biological testing. This crucial step in the drug discovery process affirms the practical applicability of the compounds generated through the model's exploration of chemical space, confirming that these molecules possess the essential structural integrity and chemical properties to be considered as potential leads for antibiotic development.

2.6 Integrating the JTVAE into the MOLGAN (RLGAN) architecture:

Having two complex models that bring new and innovative way to explore the challenges that are existing in the realm of drug discovery, and integrating them into each other is not an easy task to fulfill. As each of those models have their own way of tackling the problems they are tasked to solve, bringing them together did come with its hardships. For instance, each model was built using different machine learning frameworks, MolGAN uses tensorflow while the JTVAE uses PyTourch. Both models are also considered to be generative models so combining them in theory would take a lot of time and maybe wont improve much, however upon further investigation, the computational cost of the MolGAN was found to be high and a suspect of that is the early training phases of the model as it only takes a noise vector and goes into the generative and iterative refinement nature of the model. The issues didn't stop there, as with many GAN applications, there is a risk of mode collapse where the model might start generating a limited variety of molecules. the JTVAE also came with its limitations as its complex architecture and the need to adapt that to the MolGAN framework, also a main problem that occurred was the scalability of the model as it was hard to extend it to a larger and a more diverse dataset like the CARD.

Nonetheless, such issues had to be solved to progress with the model integration and development, so starting with the adaptation problem, I translated the JTVAE model from pytourch to tensorflow to streamline the process and ease the flow of the models together. However, when it comes to the computational cost of the MolGAN, and this is where the proposed model differs from others, introducing the learnt latent space of the JTVAE into the generator of the MolGAN is a very sophisticated yet rewarding approach. Not only should it lessen the computational cost, but it should also steer the generated molecules from the generator of the MolGAN into a more direct and structured approach. So, the JTVAE can be considered a conditional input for the generator of the MolGAN by concatenating and matching the two vectors so we would have the structural properties of the JT and the uniqueness and the continuity of the MolGAN.

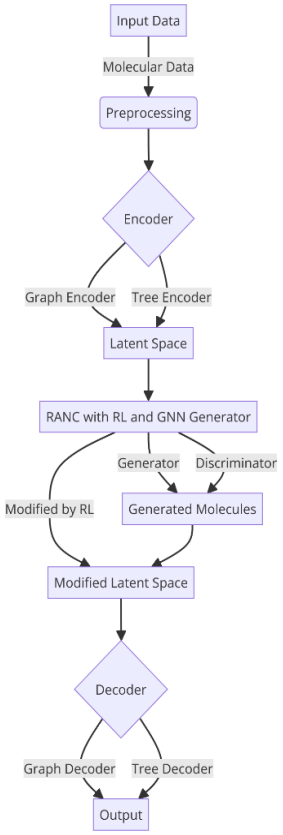

Figure 6 overflow of the integrated models

3. Evaluation and experimental setup

The comprehensive evaluation of the generated molecules is one of the most crucial steps of the drug discovery process, where the viability of these novel compounds is critically assessed. This evaluation encompasses a range of criteria essential for determining a molecule's potential as a pharmaceutical agent, including chemical validity, synthetic feasibility, and predicted efficacy against target pathogens.

3.1 Ensuring Viability

Chemical validity ensures that the molecules adhere to the fundamental principles of chemistry, while synthetic feasibility assesses whether the compounds can be realistically synthesized in a laboratory setting. Predicted efficacy against target pathogens is evaluated through various computational and experimental techniques, predicting how effectively these compounds can combat specific strains of bacteria contributing to AMR. This rigorous evaluation process is a crucial step in validating the model's output, ensuring that the compounds generated are not only innovative but also practical candidates for development into antibiotics. The model's ability to produce molecules that meet these stringent criteria demonstrates its potential to contribute valuable insights to the ongoing battle against AMR. By providing a pipeline for the rapid identification and assessment of novel antibiotic candidates, the model represents a significant advancement in the quest for new therapeutic agents. It highlights the potential of integrating advanced computational methods into the drug discovery process, offering beacon of hope in the fight against the threat of antimicrobial resistance and novel drug molecule discovery.

3.2 Technologies and tools used

In this research paper about the drug discovery using RLGAN and JTNNVAE the technical infrastructure played a vital role in determining the methods and the results. An Nvidia RTX 3050 GPU, a 12th Gen Intel(R) Core(TM) i7-12650H CPU, and 16 GB of RAM served as the hardware setup, which

was a good platform for computational tasks, although some weaknesses were noted when compared to research lab setups which are typically more powerful.

Nvidia RTX 3050 is part of Nvidia's Ampere architecture and noticeably improves matrix multiplication and parallel processing, which are core components of the generative models used in the current study. These tensor cores, AI-specific, with their increased speed of training and mixed-precision computing capability, are highly efficient in AI tasks. This GPU support was the key to drastically decreasing the time needed for model training, a frequent factor limiting the progress of AI-driven drug discovery projects.

Intel(R) CPU Core(TM) i7-12650H of the 12th Gen complements the GPU with 10 efficient and powerful cores that manage performance and efficiency. This type of software architecture allows for the execution of several computational tasks concurrently, including data preprocessing and model validation, which are among the major components of the machine learning cycle. The ability to seamlessly carry out these concurrent tasks demonstrates that the CPU is a powerful equal to the GPU and, though it might not be as advanced as the CPUs used in highly specialized research facilities, it still provides significant computational power.

On the contrary, the system's 16 GB of RAM, although enough for most apps, has some limitations in the case of deep learning projects. Since AI research in advanced levels is mostly about handling big datasets and using complex neural networks, it has been viewed as a must to use more extensive memory capacity (32 GB or more). A drawback in this process could have been that the small RAM size did not allow for bigger batch sizes and for the amount of data processed simultaneously to be increased, which, in its turn, could have caused the model iteration and testing to become slower.

Besides the fact that the RTX 3050 offers some advantages, it does not match the performance of the more expensive GPUs, which are used in dedicated AI research almost all the time, such as the Nvidia Tesla or Quadro series. The upgraded GPUs with more power can deal with bigger models and datasets more accurately, thus making it possible to conduct the experiments more quickly and the model more effectively. GPU power inequality could have been an obstacle for our research in terms of the

complexity of models which can be trained and the speed of the experiments performed. Thus, our research could have been less competitive than research conducted with more advanced hardware.

Lastly, the machine learning tasks performed in this study were adequately run on the hardware that was used, but it could not fully take advantage of the high-end resources available in more specialized settings. Probably, the disparity led to the low efficiency and scope of research which show one of the common problems that researchers who work in academic or less resourceful institutes face. Nevertheless, this research shows the possibility of AI-empowered studies within reach of public technology, as well as the correlation between price and computing power in academia.

3.3 Proposed Integration Strategy for JTNNVAE and MolGAN:

JTNNVAE's incremental approach to generating chemically valid scaffolds can significantly enhance MolGAN's generative capabilities. The integration strategy involves using JTNNVAE to first create a molecular scaffold—this includes the core structure and key functional groups of the molecule. This scaffold then serves as the starting point for the MolGAN framework, which adds atoms and bonds to further develop the molecule, focusing on optimizing specific properties such as bioactivity or solubility. JTNNVAE's capability to ensure the chemical validity of generated structures will be crucial in ensuring that all produced scaffolds are chemically valid, while MolGAN can then be adapted to generate a large quantity of new valid and novel structures. This hybrid approach leverages JTNNVAE's strengths in maintaining chemical validity and structural complexity, while utilizing MolGAN's efficiency in rapid generation and property optimization through its adversarial and reinforcement learning techniques.

3.3.1 Modifications Needed in MolGAN to Accept JTNNVAE Output:

To successfully integrate JTNNVAE scaffolds into MolGAN, several modifications are necessary which include the following:

MolGAN must be adapted to accept partial structures or scaffolds as input rather than generating a molecule from scratch. This would involve adjusting the input layer of the network to handle variable-sized inputs, as the scaffolds generated by JTNNVAE can vary in size and complexity.-

MolGAN's generator must align with the JTNNVAE scaffold context. It should generate additional atoms and bonds while maintaining coherence with the existing scaffold structure. Enhancements might be needed in the model architecture to include elements representing the scaffold's spatial and electronic properties. The reinforcement learning component guiding MolGAN towards desired properties needs recalibration considering the predefined scaffold. The reward system should focus on optimizing scaffold extensions rather than building them from scratch, ensuring that the final molecules are both desirable and chemically valid.

4. Results and findings

Among the top models in the domain, MolGAN, JTNNVAE, and RANC significantly contribute to the developments in computational drug discovery. MolGAN combines generative adversarial networks (GANs) and reinforcement learning to allow the optimization of molecule generation in terms of desired chemical properties. JTNNVAE uses the junction tree method to generate molecular graphs that remain chemically valid throughout the generation process. RANC implements the reinforced adversarial network method to generate diversified and high-quality molecules, which are critical for identifying new drug targets.

4.1 Performance and Features

MolGAN demonstrates an impressive capability in generating valid compounds, achieving near 100% validity in tests using the QM9 chemical database. This is a significant achievement compared to many state-of-the-art models that struggle with validity rates (De Cao & Kipf, 2018). A limitation noted in MolGAN is its tendency to generate disconnected graphs as molecular size increases. This was addressed in a later variant, L-MolGAN, which incorporated checks during training to penalize disconnected graphs, thus enhancing the connectivity and size of generated molecular graphs (Tsujimoto et al., 2021).

JTNNVAE excels in expanding molecules incrementally while maintaining chemical validity, a significant advantage in generating complex molecules (Jin, Barzilay, & Jaakkola, 2018). Across multiple tasks, from molecular generation to optimization, JTNNVAE has shown to outperform previous state-of-the-art methods significantly. This is largely due to its innovative approach to managing the molecular graph's generation through a structured scaffold.

MoFlow is an invertible flow model that provides exact likelihood training, efficient embedding and generation, and a guarantee of chemical validity. MoFlow (Zang & Wang, 2020) excels in tasks like

property optimization and constrained property optimization, showing high performance and generalization capability. Inspired by Transformer models, MolGPT (Bagal et al., 2021), uses natural language processing techniques for molecular design. It performs well in generating valid, unique, and novel molecules and is conditionally trainable to control multiple properties of generated molecules.

The generation of molecular graphs has significant implications in the field of drug discovery, offering a streamlined path for identifying and synthesizing new compounds with desired properties. Molecular graph generation models, such as MolGAN, JTNNVAE, and RANC, leverage advanced machine learning techniques to predict and construct molecule structures directly from learned data representations. Additionally, JTNNVAE has been applied successfully in targeted molecule generation tasks, where the model is used to generate molecules with desired properties, such as optimal binding affinity or specific pharmacokinetic characteristics. In these applications, JTNNVAE not only provides a high success rate in generating valid molecules but also significantly speeds up the design process compared to traditional drug discovery methods.

4.2 MolGAN vs JTNNVAE Comparison:

In the field of molecular graph generation, MolGAN and JTNNVAE stand out, each employing its unique approach. This comparison focuses on each model's strengths and weaknesses, evaluating their generative benefits and discussing computational efficiency and scalability. Additionally, benchmarks and performance metrics from studies using datasets like QM9 illustrate their overall performance. MolGAN and JTNNVAE both aim to generate novel molecular structures, but their approaches differ significantly. MolGAN uses a generative adversarial network combined with reinforcement learning, targeting molecules with higher solubility or suitable bonding characteristics. This process allows MolGAN to effectively generate molecules that not only meet specific requirements but also learn and create as needed. JTNNVAE employs a variation of autoencoders that construct molecules step-by-step, ensuring attainable results. This approach maintains the integrity of molecular structures, making it optimal for

designs where chemical correctness is crucial, and the creation process must be meticulously controlled. While MolGAN can generate a larger quantity of molecules quickly, many may not be chemically feasible. In contrast, JTNNVAE may produce fewer molecules, but a higher proportion are chemically viable, leading to more efficient outcomes. MolGAN benefits significantly from using GANs, enhancing its capability for rapid molecule generation. However, this advantage is tempered by challenges in training stability and sensitivity to hyperparameter settings, which can affect model performance. Conversely, JTNNVAE offers a more stable training regime thanks to its VAE framework, but the complexity of its incremental generation process can slow it down, especially when scaling to very large datasets or when rapid generation of numerous molecules is required.

The QM9 dataset, which includes about 134,000 molecules with up to nine non-hydrogen atoms, is a standard benchmark for evaluating molecular generation models like MolGAN and JTNNVAE. These models have been tested on this dataset to determine their effectiveness in generating valid, novel, and diverse molecules. Studies indicate that JTNNVAE consistently produces a higher percentage of chemically valid molecules—over 90%—compared to MolGAN, which achieves validity rates varying from 75% to 85%, depending on its configuration and training details. This highlights JTNNVAE's strength in ensuring chemical correctness. Conversely, MolGAN tends to generate a wider variety of molecular structures due to its adversarial training, which encourages the generator to explore more of the chemical space, as reflected in diversity scores and the count of unique molecules on QM9.

4.3 Expected Outcomes and Performance Improvements:

Integrating JTNNVAE scaffolds into MolGAN is expected to significantly enhance the model's performance, predicting improvements in key metrics but not only that it also streamlined the training of the MolGAN by introducing a structural concepts instead of the random noise. With chemically valid scaffolds from JTNNVAE, the proportion of valid molecules generated did increase, avoiding unfavorable or unstable substructures. Utilizing a chemically feasible starting point allows for more

efficient optimization cycles and higher success rates in property achievement. It also can reduce the time required to generate optimized molecules as it simplifies the task. The integration aims to maintain or increase the diversity of molecules. JTNNVAE ensures chemical validity of scaffolds, allowing MolGAN's adversarial network to explore a broader chemical space. With JTNNVAE ensuring the foundational structure, the overall chemical validity of the molecules as said before did improve, assessable through drug-likeness and synthetic feasibility tests.

In summary, the proposed integration of JTNNVAE and MolGAN represents a promising advancement in the field of computational molecular design. By combining the strengths of these two models, the hybrid approach aims to achieve higher efficiency, enhanced molecule diversity, and improved chemical validity, thereby accelerating the discovery and optimization of novel drug candidates. Further empirical studies and simulations will be essential to refine the integration strategy and fully realize its potential benefits.

5. Challenges and limitations

The integration of MolGAN and JTNNVAE offers a promising method to enhance molecular graph generation but introduces several technical and computational challenges. These challenges stem from the distinct architectural designs and operational dynamics of the two models. Addressing these challenges is crucial to ensure seamless integration that leverages the strengths of both models without compromising performance or scalability.

5.1 Challenges in Integrating MolGAN and JTNNVAE

Combining two complex models escalates overall computational complexity. Each model brings its set of parameters and computational needs; integrating them demands additional resources for data handling, processing, and storage. This complexity significantly increases the computational load, necessitating more powerful hardware or more efficient algorithms. Seamless integration of JTNNVAE scaffolds into MolGAN's framework for further molecule elaboration involves overhead. This overhead includes data transformation and synchronization processes that must be efficiently managed to avoid bottlenecks.

JTNNVAE and MolGAN use different learning paradigms (variational autoencoding vs. adversarial training combined with reinforcement learning). Harmonizing these approaches posed significant challenges in training and convergence. There's a risk that the integrated model might not effectively learn from the combined dataset, leading to poor convergence and suboptimal performance. Errors or biases in the scaffolds generated by JTNNVAE did at some times propagate through MolGAN, leading to the generation of invalid or undesirable molecules. This propagation could compound through training iterations, complicating stable convergence.

5.2 Overcoming Challenges: Proposed Solutions for Integration Issues

Unified Training Framework: Developing a unified training framework that accommodates each model's unique characteristics could resolve many integration issues. This would involve establishing a meta-learning environment where both models are trained simultaneously, adhering to their individual learning dynamics. This synchronization would help ensure more stable convergence. Moreover, Using advanced data preprocessing techniques to normalize and refine JTNNVAE's scaffold data for MolGAN input could reduce integration overhead. Methods like feature scaling, embedding transformations, and dimensionality reduction could ensure that the scaffold data integrates seamlessly into MolGAN without significant overhead. Implementing error detection and correction mechanisms could address error propagation issues from JTNNVAE to MolGAN. Outlier detection algorithms could be used to identify and rectify anomalous scaffold data before MolGAN processes it.

5.3 Role of Advanced Machine Learning Techniques in Mitigating Problems:

I. Transfer Learning: Employing transfer learning could facilitate better integration by using knowledge from scaffold generation (JTNNVAE) to enhance molecule elaboration (MolGAN). This could help harmonize the learning processes of the two models, improving convergence and performance.

II. Adaptive Learning Rates: Adaptive learning rate algorithms could dynamically manage learning rates based on feedback during training, addressing issues of incompatible learning dynamics. This would optimize each model's learning rate to maintain balance and support effective convergence.

III. Ensemble Techniques: Using ensemble methods, where multiple models make decisions in conjunction, could enhance the robustness of the integrated system. This approach could be particularly effective in managing the diversity of generated molecules and ensuring the chemical validity of outputs by averaging out biases or errors from individual models.

In conclusion, the integration of MolGAN and JTNNVAE poses several challenges, especially in computational demands and training dynamics. However, by employing advanced machine learning techniques such as unified training frameworks, adaptive learning rates, and ensemble methods, these challenges can be effectively mitigated. These solutions not only enhance the model's performance but also ensure its scalability and applicability in real-world drug discovery scenarios.

6. Conclusion and future work

This paper's research serves as an important landmark in the field of drug discovery and uses the potentials of RLGAN together with JTNNVAE. These technologies are revolutionary milestones that help with the process of discovering new drug molecules by drastically reducing the time and financial costs associated with the same, and thus, they tackle the foremost issues in pharmaceutical research.

The integration of RLGAN and JTNNVAE leverages the strengths of both approaches: RLGAN's ability to adapt to the dynamic environment and JTNNVAE's precision in molecular analysis. Together, they become a highly potent tool for the development of novel drug molecules that are chemically sound, as well as that are specifically aimed at certain biological targets. In contrast, this approach entirely changes the traditional drug discovery process, which is in most cases inefficient and slow as it is based on trial and error and high-throughput screening.

The results of the research demonstrate the creative power of artificial intelligence and machine learning for redefining the conventional approaches in drug development. Through automation and refinement of the drug design process by RLGAN and JTNNVAE the drug predictions and insights are generated at a speed and accuracy that far exceed what can be achieved using traditional methods. The significance of that is considerable, as it opens the way to the faster market introduction of drugs, thus saving patients' time.

Another important area of further research is discovering other ways of combining RLGAN and JTNNVAE. The present research was based on a particular integration approach, however the complexity of these models implies that other methods could lead to different or better outcomes. For instance, changing the architecture to enable more dynamic interaction between the models and by adjusting the training protocol to match the models' strengths in the generation of novel drug molecules. Investigating these alternatives could provide a deeper understanding of how RLGAN and JTNNVAE can be used together through their complementary strengths.

Furthermore, it should be noted that the experiment had its limitations ranging from hardware limitations to the intrinsic complexities of the models involved. The Nvidia RTX 3050 GPU and Intel i7 CPU, although sufficient, failed to provide the computational power needed to fully utilize the capabilities of the models. This limitation may have hindered the complexity of the experiments and the size of the datasets that could be processed effectively. Additionally, the models themselves are complex and their integration is a substantial problem that requires a lot of computing resources and expertise for the optimization.

While keeping this in mind, future works should also attempt other hardware and computational issues. Upgrading to computers with more power may help to solve some of the computational bottlenecks that occur during this research. Additionally, the models can be simplified or optimized in their architectures without compromising their effectiveness in order to make them more accessible and viable for wider applications.

This research should not be seen as the final product or the end-goal of our aspirations but rather as a part of an ongoing process of discovery and improvement. The potential for these technologies to revolutionize drug discovery is immense, and with continued refinement and exploration, we can move closer to realizing their full potential. Continuing to explore, refine, and challenge the integration of RLGAN and JTNNVAE will be essential in advancing our capabilities in drug discovery and opening new frontiers in the creation of effective solutions.

REFERENCES

[1] Kieuvongngam, V., Tan, B., & Niu, Y. (2020). Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2. ArXiv, abs/2006.01997.

[2] Bian, Y., & Xie, X. (2021). Generative chemistry: drug discovery with deep learning generative models. Journal of Molecular Modeling, 27(3). https://doi.org/10.1007/s00894-021-04674-8

[3] Fischer, T., Gazzola, S., & Riedl, R. (2019). Approaching target selectivity by de novo drug design. Expert Opinion on Drug Discovery, 14(8), 791–803. https://doi.org/10.1080/17460441.2019.1615435

[4] Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Regep, C., Liu, G., Hayter, J. B. R., Vickers, R., Roberts, C. S., Tang, J., Roblin, D., Blundell, T. L., Bronstein, M. M., & Taylor-King, J. P. (2021). Utilizing graph machine learning within drug discovery and development. Briefings in Bioinformatics, 22(6). https://doi.org/10.1093/bib/bbab159

[5] Grisoni, F. (2023). Chemical language models for de novo drug design: Challenges and opportunities. Current Opinion in Structural Biology, 79, 102527. https://doi.org/10.1016/j.sbi.2023.102527

[6] Grosdidier, A. (n.d.). SwissDock - The online docking web server of the Swiss Institute of Bioinformatics - TargetProteins. http://www.swissdock.ch/target_proteins

[7] Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative Structure–Activity relationships. Journal of Chemical Information and Modeling, 55(2), 263–274. https://doi.org/10.1021/ci500747n

[8] Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A. G., Aidinis, V., Lynch, I., Greco, D., & Melagraki, G. (2021). Advances in de novo drug design: from conventional to machine learning methods. International Journal of Molecular Sciences, 22(4), 1676. https://doi.org/10.3390/ijms22041676

[9] Stokes, J., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T., Barzilay, R., & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.e13. https://doi.org/10.1016/j.cell.2020.01.021

[10] Wang, M., Sun, H., Wang, J., Pang, J., Chai, X., Xu, L., Li, H., Cao, D., & Hou, T. (2021). Comprehensive assessment of deep generative architectures for de novo drug design. Briefings in Bioinformatics, 23(1). https://doi.org/10.1093/bib/bbab544

[11] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences, 28(1), 31–36. https://doi.org/10.1021/ci00057a00

[12] Sterling and Irwin, J. Chem. Inf. Model, 2015 http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00559

[13] Alcock et al. 2023. CARD 2023: Expanded Curation, Support for Machine Learning, and Resistome Prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research, 51, D690-D699.

[14] Aphikulvanich, R., & Pornputtapong, N. (2023, May 2). (open access) mol-zero-gan: Zero-shot adaptation of molecular generative adversarial network for specific protein targets (2023). SciSpace - Paper. https://typeset.io/papers/mol-zero-gan-zero-shot-adaptation-of-molecular-generative-1ipx4n9m

[15] arXiv (Cornell University). (2022, February 14). (open access) disentangle VAE for molecular generation (2022). SciSpace - Paper. https://typeset.io/papers/disentangle-vae-for-molecular-generation-6obpgpml

[16] Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine learning in Drug Discovery: A Review. Artificial intelligence review. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8356896/

[17] De Cao, N., & Kipf, T. (2022, September 27). Molgan: An implicit generative model for small molecular graphs. arXiv.org. https://arxiv.org/abs/1805.11973

[18] Jin, W. (2019, March 29). Junction tree variational Autoencoder for molecular graph ... https://arxiv.org/pdf/1802.04364

[19] Medicine, I. (2023, May 17). Insilico Medicine successfully explores the advantages of quantum generative adversarial networks in Generative Chemistry. United States. https://www.prnewswire.com/apac/news-releases/insilico-medicine-successfully-explores-the-advantages-of-quantum-generative-adversarial-networks-in-generative-chemistry-301827125.html

[20] Navon, A., Shamsian, A., Chechik, G., & Fetaya, E. (2021, April 26). Learning the pareto front with Hypernetworks. arXiv.org. https://arxiv.org/abs/2010.04104

[21] Society for Industrial and Applied Mathematics eBooks (Ed.). (2023, January 1). MOVAE: A variational AutoEncoder for Molecular Graph Generation (2023). SciSpace - Paper. https://typeset.io/papers/movae-a-variational-autoencoder-for-molecular-graph-3f91g14k

[22] Zhao, L., Wang, J., Pang, L., Liu, Y., & Zhang, J. (2020, January 9). (Open access) gansdta: Predicting drug-target binding affinity using gans. (2020): Lingling Zhao: 44 citations. SciSpace - Paper. https://typeset.io/papers/gansdta-predicting-drug-target-binding-affinity-using-gans-261btk2fus

[23] (open access) VGAE-MCTS: A new molecular generative model combining variational graph auto-encoder and Monte Carlo Tree Search (2023). SciSpace - Paper. (2023, June 5). https://typeset.io/papers/vgae-mcts-a-new-molecular-generative-model-combining-39fiw3yf

[24] (open access) exploring the advantages of quantum generative adversarial networks in Generative Chemistry (2022): Po-Yu Kao: 3 citations. SciSpace - Paper. (2022b, October 30). https://typeset.io/papers/exploring-the-advantages-of-quantum-generative-adversarial-35ojan69

[25] Richards, R. M., & Groener, A. M. (2022, May 1). Conditional β-vae for de Novo Molecular Generation (2022): Ryan M. Richards: 3 citations. SciSpace - Paper. https://typeset.io/papers/conditional-beta-vae-for-de-novo-molecular-generation-l9z5oqqv

[26] Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., & Zhavoronkov, A. (2017). druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico.. Molecular pharmaceutics, 14 9, 3098-3104 . https://doi.org/10.1021/acs.molpharmaceut.7b00346.

[27] Blanchard, A., Stanley, C., & Bhowmik, D. (2021). Using GANs with adaptive training data to search for new molecules. Journal of Cheminformatics, 13. https://doi.org/10.1186/s13321-021-00494-3.

[28] Li, J., Topaloglu, R., & Ghosh, S. (2021). Quantum Generative Models for Small Molecule Drug Discovery. IEEE Transactions on Quantum Engineering, 2, 1-8. https://doi.org/10.1109/tqe.2021.3104804.

[29] Xu, M., Cheng, J., Liu, Y., & Huang, W. (2021). DeepGAN: Generating Molecule for Drug Discovery Based on Generative Adversarial Network. 2021 IEEE Symposium on Computers and Communications (ISCC), 1-6. https://doi.org/10.1109/ISCC53001.2021.9631396.

[30] Putin, E., Asadulaev, A., Ivanenkov, Y., Aladinskiy, V., Sánchez-Lengeling, B., Aspuru-Guzik, A., & Zhavoronkov, A. (2018). Reinforced Adversarial Neural Computer for de Novo Molecular Design. Journal of chemical information and modeling, 58 6, 1194-1204 . https://doi.org/10.1021/acs.jcim.7b00690.

[31] Liu, K., Han, Y., Gong, Z., & Xu, H. (2023). Low-Data Drug Design with Few-Shot Generative Domain Adaptation. Bioengineering, 10. https://doi.org/10.3390/bioengineering10091104.

[32] Hussain, S., Anees, A., Das, A., Nguyen, B., Marzuki, M., Lin, S., Wright, G., & Singhal, A. (2020). High-content image generation for drug discovery using generative adversarial networks. Neural networks : the official journal of the International Neural Network Society, 132, 353-363 . https://doi.org/10.1016/j.neunet.2020.09.007.

[33] Xiong, J., Xiong, Z., Chen, K., Jiang, H., & Zheng, M. (2021). Graph neural networks for automated de novo drug design.. Drug discovery today. https://doi.org/10.1016/j.drudis.2021.02.011.

[34] Manu, D., Sheng, Y., Yang, J., Deng, J., Geng, T., Li, A., Ding, C., Jiang, W., & Yang, L. (2021). FL-DISCO: Federated Generative Adversarial Network for Graph-based Molecule Drug Discovery: Special Session Paper. 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD), 1-7. https://doi.org/10.1109/ICCAD51958.2021.9643440.

[35] Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular property prediction with graph neural networks.. Drug discovery today. Technologies, 37, 1-12 . https://doi.org/10.1016/j.ddtec.2020.11.009.

[36] Sun, R. (2022). Does GNN Pretraining Help Molecular Representation?. ArXiv, abs/2207.06010. https://doi.org/10.48550/arXiv.2207.06010.

[37] Cheung, M., & Moura, J. (2020). Graph Neural Networks for COVID-19 Drug Discovery. 2020 IEEE International Conference on Big Data (Big Data), 5646-5648. https://doi.org/10.1109/BigData50022.2020.9378164.

[38] Gupta, A., & Mukherjee, A. (2023). CardiGraphormer: Unveiling the Power of Self-Supervised Learning in Revolutionizing Drug Discovery. ArXiv, abs/2307.00859. https://doi.org/10.48550/arXiv.2307.00859.

[39] Smalley, E. (2017). AI-powered drug discovery captures pharma interest. Nature Biotechnology, 35, 604-605. https://doi.org/10.1038/nbt0717-604.

[40] Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., Gao, P., Xie, G., & Song, S. (2020). Learn molecular representations from large-scale unlabeled molecules for drug discovery. ArXiv, abs/2012.11175.

[41] Xie, Y., Shi, C., Zhou, H., Yang, Y., Zhang, W., Yu, Y., & Li, L. (2021). MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ArXiv, abs/2103.10432.

[42] Walters, W., & Barzilay, R. (2021). Critical assessment of AI in drug discovery. Expert Opinion on Drug Discovery, 16, 937 - 947. https://doi.org/10.1080/17460441.2021.1915982.

[43] Deep learning. Deep Learning. (n.d.). https://www.deeplearningbook.org/

[44] Data Center Solutions, IOT, and PC innovation. Intel. (n.d.). https://www.intel.com/content/www/us/en/homepage.html

Appendix

Attached below is the link of the proposed model's repository from Github

https://github.com/majdbar321/Capstone