Generative Molecular Design for Drug Discovery

Gauri Samith (gs3280), Nicole Brye (nab2217) , Arjun Bhan(AB5666), Elie Chemouni (ec3778)

Abstract

Our capstone project, conducted in collaboration with Frederick National Laboratory, aimed to enhance the Generative Molecular Design (GMD) pipeline, a tool for creating innovative cancer treatment drugs by targeting abnormal cell division. We focused on optimizing two main models within the pipeline, AutoGrow4 and JTVAE, which use genetic optimization and molecular docking to improve drug efficacy. Our analysis included the impact of various hyperparameters on drug properties like lipophilicity and fitness scores. The project involved weekly independent research by team members exploring additional modeling techniques and optimizations for potential integration. This work represents a notable advancement in drug discovery and cancer therapy.

1 Introduction

Drug research is an essential element of modern medicine. Its goal is to identify and create new therapeutic agents to prevent, treat, or cure diseases. This field faces an enormous challenge in analyzing potential drug compounds because that number ranges between 10^{33} and 10^{80} combinations, making the search for effective drugs arduous, expensive and time consuming.

Recent advancements in AI have significantly improved this process. AI software can now automate the analysis of a vast number of chemical compounds, creating new drugs with high accuracy. This automated process facilitates a much faster and cost-effective pathway from concept to clinical trial, reducing the time and labor requirements.

Frederick National Laboratory has been at the forefront of these advancements. Their FNL-GMD pipeline is able to generate chemical compounds efficiently and effectively. This pipeline uses two cutting edge models: the Junction Tree Variational Autoencoder and AutoGrow4. These models are able to develop drug compounds that target and disrupt abnormal cell division potentially leading to groundbreaking treatments for various forms of cancer. This AI-based approach not only bolsters the efficiency and effectiveness of drug creation but opens the door to revolutionary new cancer therapies.

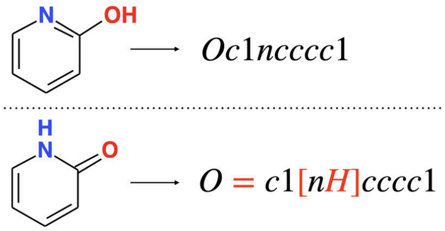
During this project, we have had the opportunity to collaborate with the research team at Frederick National Laboratory on enhancing the FNL-GMD pipeline. Our work involved integrating new features into the pipeline and conducting thorough analyses of its outputs. We have gained a comprehensive understanding of the model from both a theoretical and applied perspective. Our modifications and insights have greatly improved the FNL-GMD pipeline, making it a more robust system.

2 Datasets

For our project, we rely on two primary databases: the ZINC database and the OPERA database. The ZINC database serves as a comprehensive resource for drug discovery, offering a vast collection of commercially available compounds. It presents a pivotal resource for structure-based virtual screening, addressing a critical barrier in accessibility to purchasable compound libraries. Currently housing a library of 727,842 molecules, each with a detailed 3D structure, the database continues to expand in size. These molecules are sourced from catalogs of compounds from vendors and are meticulously curated with biologically relevant protonation states and annotations, including molecular weight, calculated LogP, and the number of rotatable bonds. Accessible for free in multiple file formats including SMILES, mol2, 3D SDF, and DOCK flexibase format, the database offers a user-friendly query tool equipped with a molecular drawing interface for seamless searching, browsing, and subset creation.

On the other hand, the OPERA database specializes in natural product-inspired compound libraries, offering a curated collection of molecules derived from natural sources. These compounds are particularly valuable for their diverse and complex chemical structures, often serving as inspiration for novel drug design. By leveraging both the ZINC and OPERA databases, we gain access to a rich array of molecular structures, enabling us to conduct thorough explorations and optimizations in our generative molecular design efforts.

In our project, the input data consist of SMILES strings to represent the molecules. SMILES, which stands for Simplified Molecular Input Line Entry System, is a widely used notation system for representing the structure of chemical molecules using ASCII strings. Each SMILES string provides a compact and standardized representation of a molecule's structure, capturing its connectivity and stereochemistry in a simple, human-readable format.



This notation system is particularly useful for our project as it allows for the efficient storage, sharing, and manipulation of molecular structure data. Additionally, SMILES strings can be easily processed by computational algorithms, making them ideal input data for tasks such as generative molecular design, virtual screening, and molecular similarity analysis. Overall, the use of SMILES strings streamlines our workflow, enabling seamless integration of molecular data into our computational models and facilitating the exploration of chemical space.

3 Related Work

Our team engages in a weekly journal club aimed at delving deeper into the realm of Generative Molecular Design (GMD). During these sessions, a team member presents a research or review paper pertaining to GMD, facilitating ongoing learning and knowledge expansion within our group. This initiative serves to enrich our comprehension of recent advancements and bolsters our capacity to integrate state-of-the-art techniques and insights into our project. Throughout these discussions, we explore various research papers, such as those briefly presented below, which have provided valuable insights into the methodologies and principles underpinning GMD.

In their review, Bian and Xie (2021) explore the advancements in drug discovery facilitated by deep learning models, emphasizing the shift from traditional screening methods to sophisticated approaches like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). This backdrop highlights the significance of the Junction Tree Variational Autoencoder (JTVAE) in the generative chemistry landscape. JTVAE innovatively extends VAEs to molecular graphs, enabling the generation of chemically valid molecules by treating them as assemblies of predefined subgraphs. The insights from Bian and Xie underscore the transformative potential of JTVAE within AI-driven drug discovery, marking a significant evolution in leveraging deep learning for computational chemistry.

Spiegel and Durrant's 2020 work introduces AutoGrow4, an open-source genetic algorithm for automated drug design and lead optimization. AutoGrow4 innovates by using genetic algorithms to evolve drug-like molecules, not limited by pre-enumerated compound libraries. Its enhancements in speed, stability, modularity, and compatibility with new docking programs and filters mark a significant advancement. Demonstrated through applications to the PARP-1 protein, AutoGrow4 effectively generates compounds with superior predicted binding affinities compared to existing FDA-approved inhibitors, showcasing its potential in facilitating computational drug discovery processes.

Mcloughlin et. al’s work (2023) discusses the ATOM-GMD framework, and how it poses a unique, efficient solution for exploring large chemical spaces in search of compounds that bind well to Histamine H1 receptors (but not to M2 receptors). The framework utilizes autoconders to project the compounds into latent space, predictive models, and an optimization framework to explore a very broad chemical space and preserve diversity in the compounds produced. Overall, this framework addresses the concerns that created molecular structures must simultaneously satisfy multiple objectives, including but not limited to efficacy, safety, and developability properties.

In their review, Ulrich, Goss, and Ebert (2021) investigate the utilization of deep learning techniques and data augmentation to explore the octanol–water partition coefficient dataset. They highlight the increasing availability of large datasets and the growing relevance of deep neural networks (DNNs) in computational chemistry. Focusing on the prediction of chemical properties from chemical structures, they specifically examine the octanol-water partition coefficient (log P), crucial in environmental chemistry, toxicology, and chemical analysis. Their developed DNN exhibits promising predictive performance. Notably, they employ data augmentation, considering all potential tautomeric forms of the chemicals, to enhance the DNN's training. Additionally, Ulrich et al. demonstrate the utility of DNN models in curating the log P dataset by identifying potential errors and addressing dataset limitations.

4 Methodology

4.1 Genetic Optimization

Genetic optimization works on a population of potential solutions to a problem, treating each solution as an individual in that population. Each individual is typically represented as a string of values, known as chromosomes or genomes, which can be manipulated algorithmically. The process mimics natural evolutionary processes such as selection, crossover (recombination), and mutation. The key components are detailed below:

- Population**:** A set of potential solutions to the problem. The diversity within this population is key to the success of genetic optimization, as it increases the chances of finding a globally optimal solution.
- Fitness Function: A function that evaluates and assigns a fitness score to each individual in the population based on how well it solves the problem. The fitness score determines the likelihood of an individual being selected for reproduction.
- Selection: A process that simulates natural selection where individuals are chosen from the current population to be the parents of the next generation. Selection is often based on fitness, with fitter individuals having a higher probability of being chosen.
- Crossover: A genetic operator used to combine the genetic information of two parents to generate new offspring. This mimics biological reproduction and is vital for sharing and mixing genetic traits within the population.
- Mutation: A genetic operator that introduces random changes to individual genomes. This is crucial for maintaining genetic diversity within the population and for exploring new parts of the solution space that might not be reachable by crossover alone.

4.2 AutoGrow4

AutoGrow4 is an open-source software tool designed for the discovery and optimization of novel chemical compounds in drug development. It utilizes molecular docking techniques to model how various chemical compounds affect specific biological targets. This method allows for a deeper understanding of the drug-target interactions, enhancing the ability to assess and optimize the effectiveness of potential drug compounds.

AutoGrow4 does not “train” in the traditional sense associated with machine learning models. Instead, it utilizes evolutionary principles to develop and refine drug candidates through multiple iterations. This process relies heavily on random sampling, filtration and modification to mimic natural selection. AutoGrow4 starts with an initial population of given compounds. Each compound undergoes molecular docking, which simulates the interactions between a target protein and potential drug compound.

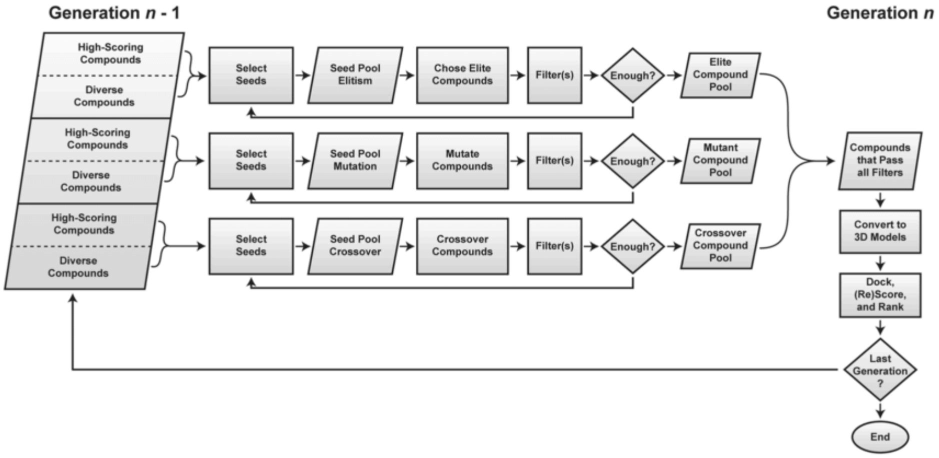
4.2.1 Docking and Diversity Scores

This software assesses and scores the binding affinity of each drug to the target protein. A higher docking score indicates a greater stability and strength of the interaction. To ensure effective scoring, the software represents the molecular compound in 3-dimensional space. This representation closely mirrors the actual binding that occurs and allows the model to more accurately assess how effectively the drug compound binds to the target protein. In situations where the structure of the target protein is unknown, AutoGrow4 uses a Ligand-Based Design (LBD) approach to analyze the target protein. This method uses data from other molecules known to bind effectively in order to create new drug compounds. Through the analysis of the chemical and spatial configurations of these successful compounds, AutoGrow4 can measure the docking score without having direct information on the target protein.

Besides docking scores, AutoGrow4 uses another scoring system called Diversity Score. This scoring metric measures the diversity of the initial compound population. By having a more diverse initial population it can ensure a more broad exploration of the chemical space. This can lead to the discovery of more varied and effective drug compounds.

4.2.2 Role Genetic Optimization

In order to correctly find the most optimal drug compound, AutoGrow4 uses the process of genetic optimization. After each iteration, the software uses a filtration system to eliminate candidates which fail to meet certain criteria such as having too low of a docker score. In each iteration, Autogrow4 introduces a new population of drug compounds based on the output of genetic optimization.



4.3 Junction-Tree Variational Auto-Encoders (JTVAE)

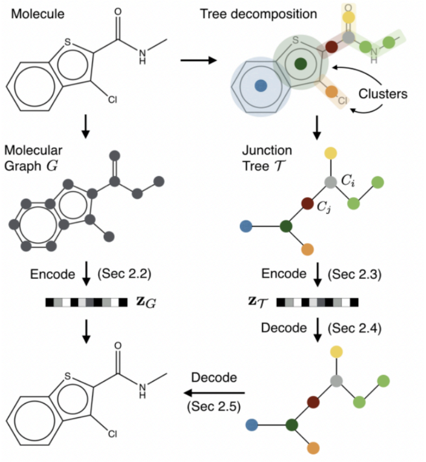
The Junction Tree Variational Autoencoder (JT-VAE) is a specialized framework designed to encode and decode molecular structures, which are inherently graph-based and complex. The architecture of JT-VAE is bifurcated into two primary components: the encoder and the decoder, both of which are augmented by the junction tree representation of molecules.

4.3.1 Encoder

The encoder consists of two key subcomponents—the tree encoder and the graph encoder. The tree encoder processes the junction tree of the molecule, which abstracts the molecule into clusters of atoms (or chemical substructures), facilitating the capture of the molecule's topological hierarchy. Concurrently, the graph encoder processes the molecular graph, focusing on the atomic connectivity. Both encoders utilize neural networks, typically convolutional neural networks or recurrent neural networks, to map the input structures into a continuous latent space. This latent space encodes complex molecular information as fixed-size vectors, which represent probabilistic distributions defined by learned parameters for the mean and variance.

4.3.2 Decoder

The decoder reverses the encoding process, aiming to reconstruct the molecular graph from the latent representation. It uses the probabilistic latent space vectors to generate both the junction tree and the molecular graph. The reconstruction process ensures that the output molecules are chemically valid, leveraging the structured representation provided by the junction trees to maintain molecular integrity and validity.



4.3.3 Role of Genetic Optimization

Genetic optimization is applied within the latent space of the JT-VAE to explore and discover novel molecular structures with optimized properties. This process is articulated through several stages:

- Initialization: A population of latent vectors is initialized, either by sampling from the distributions represented in the latent space of known molecules or by generating random vectors within the bounds of the latent space.
- Fitness Function: A fitness function is crucial in this methodology as it quantitatively evaluates the molecular structures decoded from the latent vectors. The fitness criteria can be diverse, including but not limited to, pharmacokinetic properties, synthetic

accessibility, or specific biological activities. The evaluation guides the selection process in the genetic algorithm, driving the evolution toward optimal molecular configurations.

Each generation involves decoding the manipulated latent vectors back into molecular structures, which are then evaluated using the fitness function. This iterative process continues until a defined convergence criterion is met, such as a set number of generations or a threshold fitness value.

In terms of deciding the last generation, both models enable the user to select how many generations the algorithm should run. The user is able to adjust several key parameters such as population size, mutation rate, crossover rate and elitism. The population size dictates the number of molecules used in each generation. The rate of crossover and mutation can be changed with the mutation rate and crossover rate respectively. Adjusting the elitism amount changes the number of elite values carried over from the prior generation. By adjusting these hyperparameters these allow us to optimize the discover and development of effective drug compounds.

5 Experiments & Results

In order to evaluate the efficacy of both the JT-VAE and Autogrow 4 models described in detail above, we first analyzed chemical compound metric distributions for the initial population of molecules. The initial population of molecules is composed of all molecules from both the ZINC and OPERA datasets, and the molecular properties being analyzed in these experiments are lipophilicity (logP), synthetic accessibility (SA) and cycle score. The following experiments also look largely at a molecule’s overall fitness score, which can be defined as:

$$Fitness = Normalized\ logP + Normalized\ SA + Normalized\ Cycle\ Score$$

The initial population distributions for each of these metrics are displayed as histograms below:

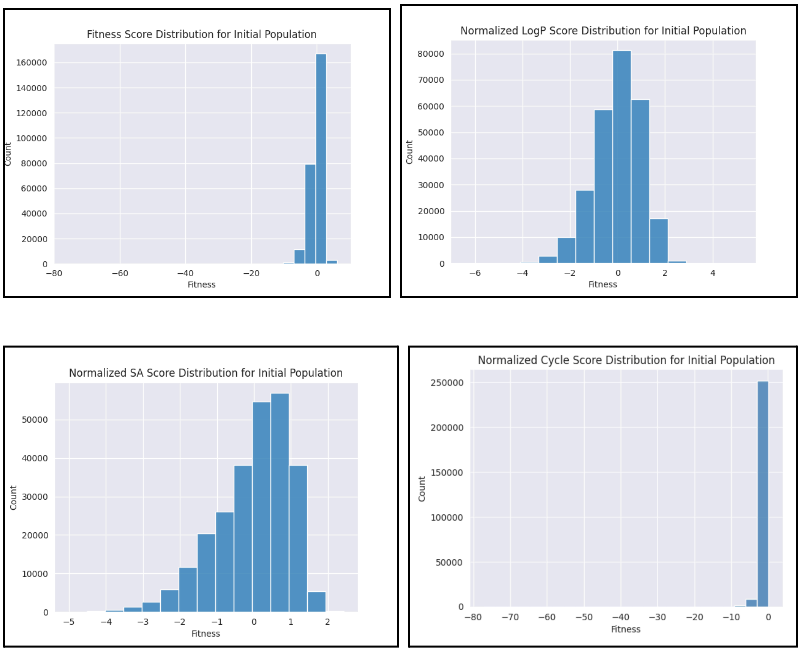


Figure 1: Initial population distributions for overall fitness (upper left), normalized logP (upper right), normalized SA (lower left), and normalized cycle score (lower right).

As can be seen above, the overall fitness score distribution is very skewed left and highly variable. The goal of employing the JT-VAE and Autogrow4 models is to generate compounds that exhibit desirable/targeted properties (e.g. less variable fitness scores), therefore we have performed optimization parameter manipulation and initial population fine tuning to achieve this task.

As stated above, we have performed a series of optimization parameter manipulation experiments to fine tune both models. For the JT-VAE model, we modified the mutate and mate probabilities to determine whether or not these hyperparameters had a significant effect on the generated compounds' overall fitness score. Below are boxplots displaying the distributions of the resulting compounds after we hold the mate and mutate probabilities constant while running the model for 5 epochs on each combination of hyperparameters:

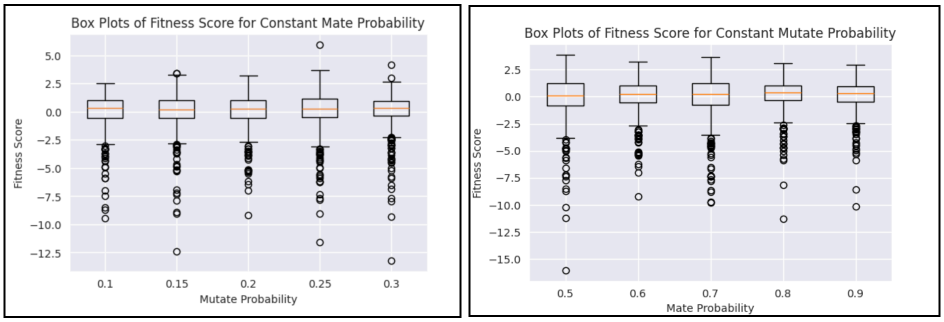


Figure 2: Overall fitness score distributions for running the JT-VAE model on different combinations of mate and mutate probabilities. Boxplot of distributions while holding the mate probability constant (left) and boxplot of distributions while holding the mutate probability constant (right).

As can be seen above, the center of all distributions remains at around 0. Hyperparamter tuning yields very small changes in the distributions of fitness scores however, with regards to the initial distribution of overall fitness scores (Figure 1) the distributions above are far less variable and have a much less prominent left skew. In order to investigate hyperparameter tuning for the JT-VAE model even further, we ran the model for 25 different combinations of mate and mutate probabilities. The resulting heatmap is displayed below:

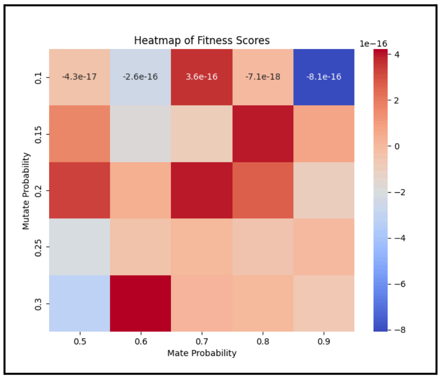


Figure 3: Heatmap of mean overall fitness score for different combinations of mate and mutate probabilities.

As can be seen above, more extreme values of mate and mutate probabilities (eg. 0.9 mate probability and 0.1 mutate probability) yield lower mean fitness score values of the output population, while less extreme values of mate and mutate probabilities (eg. 0.8 mate probability and 0.15 mutate probability) yield higher mean fitness score values of the output population.

Although we are not aiming for a particular value of fitness in the scope of this project, these details are important to note when, for example, higher or lower fitness scores are ideal to design a drug with the desired properties.

With regards to the Autogrow4 model, we have also conducted a variety of experiments to determine which components of the model have the largest impact on the resulting compounds' overall fitness score. Firstly, we analyzed the distribution of overall fitness at each epoch (generation), and also analyzed the distributions of overall fitness score for each source of genetic optimization (mutation, crossover, or elitism). The resulting distributions are displayed as boxplots below:

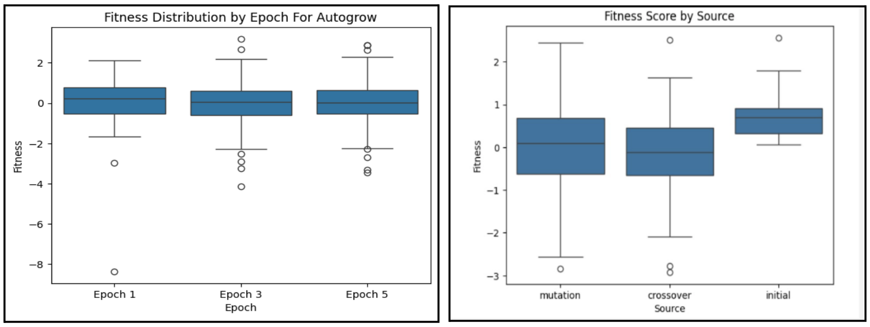


Figure 4: Autogrow 4 Distributions of overall fitness after each epoch (left) and distributions of overall fitness from each genetic optimization source (right)

As shown above, it is evident that changes in epoch have minor effects on the overall fitness score of the resulting compounds. The distributions are all roughly centered at a fitness score of 0, but are far less variable and far less skewed left than the initial population distribution of overall fitness (Figure 1). However, it is important to note that the source of genetic optimization appears to have some effect on the distribution of overall fitness score from the resulting population. When using mutation and crossover to generate new compounds, it appears that the fitness scores are centered roughly at 0, and there is a larger spread of fitness scores than with elitism. When using elitism to generate new compounds, it appears that overall fitness scores increase in the resulting population, and this distribution of fitness scores is less variable.

In addition to optimization parameter manipulation, we also performed a series of experiments by fine tuning the initial population of molecules. In order to do this, we defined the division molecules as follows:

- Any molecule with *logP < 0* - Assigned to the *hydrophilic* group
- Any molecule with *logP >= 0 and logP <= 3* - Assigned to the *balanced* group
- Any molecule with *logP > 3* - Assigned to the *lipophilic* group

After running both JT-VAE and Autogrow 4 for 10 epochs on the full initial population of molecules, we obtained the following breakdown of groups:

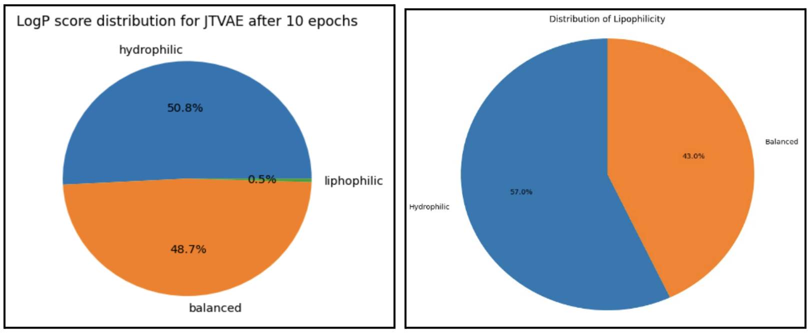


Figure 5: LogP score distribution of the initial molecule population for JT-VAE (left) and Autogrow 4 (right)

For both models, the percentages of hydrophilic and balanced compounds are at roughly 50% (with the Autogrow 4 model outputting a slightly higher proportion of balanced compounds than the JT-VAE model). However, it's important to note that only 0.5% of the resulting compounds produced by JT-VAE are lipophilic, and *none* of the resulting compounds produced by Autogrow 4 are lipophilic. In these next few experiments, we investigate whether or not fine tuning the initial population can affect this composition, and potentially provide us with more resulting compounds that are lipophilic.

We split the initial population into groups based on the criteria specified in the bullet points above, and ran both models on these fine-tuned groups. The resulting distributions are reported below:

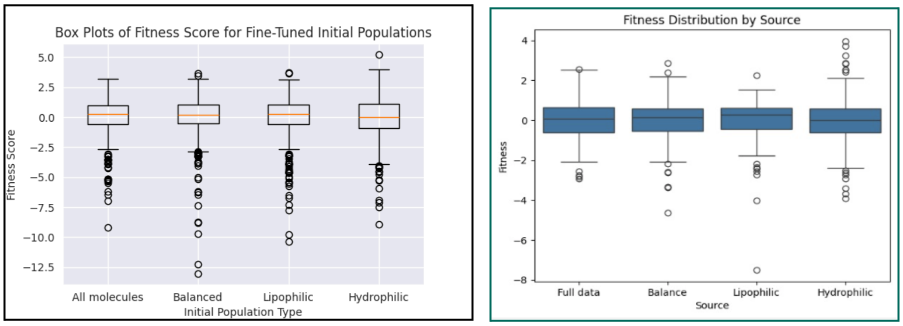


Figure 6: Comparison of JT-VAE (right) and Autogrow 4 (left) distributions of overall fitness scores after running both models on fine-tuned populations

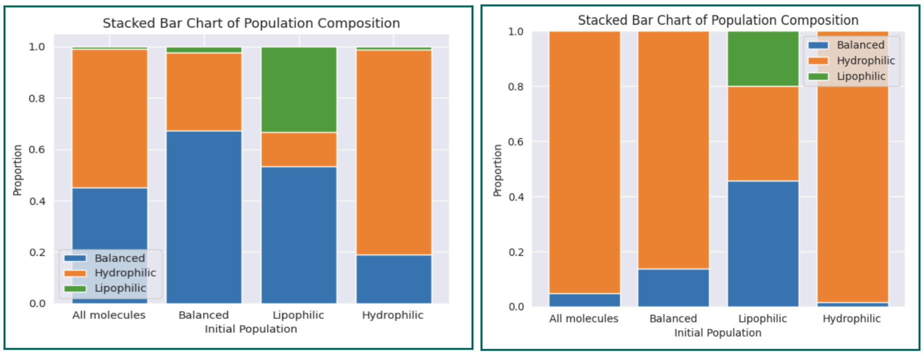


Figure 7: Comparison of JT-VAE (right) and Autogrow 4 (left) stacked bar charts of the resulting population composition for each fine tuned initial population.

As shown above in the box plots, both JT-VAE and Autogrow 4 models maintain a distribution of fitness scores centered at roughly 0 for every different initial population. When looking at the stacked bar charts, however, one can see that for both models, having an initial population consisting of only one type of compound causes the model to output more of that specific type of compound. The Autogrow 4 model appears to have a tendency to output more hydrophilic compounds overall, however, providing the Autogrow 4 with only lipophilic compounds can help guide the model to produce lipophilic compounds (when otherwise, it doesn't produce lipophilic compounds at all). The stacked bar charts above are compelling evidence to suggest that initial population fine tuning plays a large role in the composition of resulting compounds (for both JT-VAE and Autgrow 4), and that this investigation is worth pursuing further with even more fine-tuned population groups.

6 Discussion and Future Work

The exploration of generative molecular design for novel cancer drug candidates through the implementation of JT-VAE and Autogrow 4 models has yielded promising results and avenues for future research.

One noteworthy aspect of this project is its potential for scalability and applicability to a broader spectrum of drug properties beyond the limited set of molecular properties explored. While the current focus has been on properties such as lipophilicity, synthetic accessibility, and cycle score, the pipeline established here can readily accommodate additional drug properties. Incorporating a wider range of properties into the design process could enhance the robustness and versatility of the models, potentially leading to the discovery of compounds with a broader range of desirable characteristics.

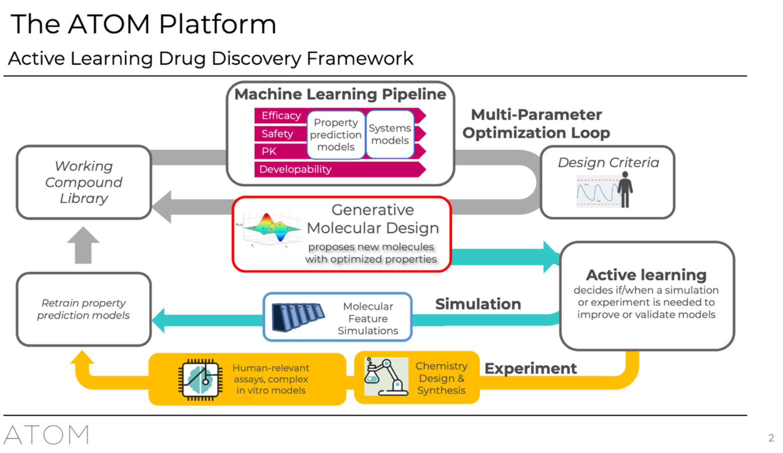
Both the JT-VAE and Autogrow 4 models have demonstrated an ability to enhance the fitness scores of the original molecular populations. Despite the complexity of molecular design, hyperparameter tuning has revealed that even subtle adjustments can have discernible impacts on the distributions of fitness scores. This underscores the importance of fine-tuning model parameters to achieve optimal performance. However, it's noteworthy that while hyperparameter tuning can refine the performance of the models, the changes in fitness score distributions remain relatively small, suggesting the need for further investigation into more sophisticated optimization strategies.

Future endeavors could involve the integration of additional scoring methods and optimization techniques to augment the current pipeline. For instance, incorporating metrics for drug novelty could help identify compounds with unique chemical structures, potentially leading to the discovery of novel therapeutic agents. Furthermore, the implementation of advanced optimization methods such as Bayesian optimization could offer more efficient exploration of the vast chemical space, facilitating the discovery of promising drug candidates.

Another avenue for future exploration lies in the extraction of more fine-tuned initial molecular populations tailored to specific drug targets. By curating initial populations based on the desired pharmacological properties of the target compounds, it may be possible to expedite the generation of compounds with the desired characteristics. This approach could potentially enhance the efficiency and effectiveness of the molecular design process, accelerating drug discovery efforts.

6.1 Integration of the AMPL score

Moreover, the integration of the AMPL score developed by the AMPL team presents an opportunity to enrich the molecular design pipeline further. Integration of the AMPL score into the GMD pipeline represents a significant advancement in the quest for novel cancer drug candidates. The incorporation of the AMPL model introduces a robust scoring function that enhances the predictive capabilities of the GMD process, thereby facilitating the identification of molecules with desired pharmacological properties.


The ATOM Platform

Active Learning Drug Discovery Framework

Machine Learning Pipeline
Efficacy
Safety
PK
Developability
Property prediction models
Systems models
Multi-Parameter Optimization Loop
Design Criteria
Working Compound Library
Generative Molecular Design
proposes new molecules with optimized properties
Active learning
decides if/when a simulation or experiment is needed to improve or validate models
Simulation
Molecular Feature Simulations
Retrain property prediction models
Human-relevant assays, complex in vitro models
Chemistry Design & Synthesis
Experiment
ATOM
2

At the heart of the GMD loop lies the iterative process of predicting molecular properties, scoring against design criteria, and optimizing molecular candidates. The integration of the AMPL model enriches this process by providing accurate predictions of crucial molecular properties essential for drug development, such as LogP values, which serve as indicators of a molecule's lipophilicity.

Additionally, the AMPL model can be extended to predict other pertinent properties related to efficacy, safety, pharmacokinetics (PK), and developability, thereby offering a comprehensive evaluation of candidate molecules. In practice, the application of the AMPL model begins with the prediction of molecular properties for the initial set of molecular structures collected at the onset of the GMD process. These predicted properties serve as the basis for subsequent scoring against predefined design criteria aligned with therapeutic goals. Molecules meeting the specified criteria are retained in the initial molecular population for further optimization rounds, while those failing to meet the criteria are discarded, ensuring that only molecules with the desired profiles progress through the design loop.

The predictive power of the AMPL model enhances the efficiency and efficacy of the GMD process by enabling the identification of molecules with favorable drug-like properties early in the design cycle. By leveraging the insights provided by the AMPL score, researchers can prioritize candidate molecules that exhibit promising pharmacological profiles, thereby accelerating the drug discovery process.

Thus, the integration of the AMPL model fosters a synergistic relationship between machine learning-driven prediction and generative model optimization within the GMD framework. This symbiosis enables continuous refinement of molecular candidates through iterative cycles of prediction, scoring, and optimization, ultimately culminating in the selection of the most promising drug candidates.

7 Conclusion

In conclusion, while this project represents a significant step forward in generative molecular design for cancer drug discovery, there remain numerous avenues for further exploration and refinement. By expanding the scope of properties considered, refining optimization strategies, and leveraging advanced scoring methods, the potential for discovering novel and efficacious cancer therapeutics can be further realized. The integration of the AMPL score into the GMD pipeline represents a pivotal advancement that empowers researchers to harness the full potential of machine learning in molecular design. By leveraging the predictive capabilities of the AMPL model, researchers can expedite the discovery of novel cancer drug candidates with enhanced therapeutic efficacy and safety profiles, thereby addressing unmet medical needs and advancing the forefront of cancer therapeutics. GMD fits into a much larger machine learning pipeline set to enhance and improve upon the current drug discovery framework. The overall framework is a helpful computational approach that needs to be combined with experimental approaches to accelerate drug discovery and development.

8 Ethical Considerations

Using generative AI in drug discovery raises significant ethical considerations including data bias, which could affect drug efficacy across diverse populations. Issues of transparency are also prevalent as AI's decision-making processes are often opaque, which could affect regulatory compliance. The safety and reliability of AI-generated treatments must be thoroughly validated, balancing the speed of discovery with thorough safety checks. Lastly, ethical clinical trial conduct, especially informed consent, must be strictly upheld to ensure that these advancements translate into fair and safe outcomes for all patients.

9 Contributions

Gauri: EDA, debugging, training and fine-tuning JTVAE model; scorer method segregation, data extraction, data curation; AMPL integration; report methodology and editing;

Nicole: EDA, debugging, training and fine-tuning JTVAE model; report results, repository organization; data curation; AMPL integration

Elie: EDA, debugging, training and fine-tuning AutoGrow4 model; report data, related works, discussion, conclusion; data curation; AMPL integration

Arjun: EDA, debugging, training and fine-tuning AutoGrow4 model; report introduction, autogrow4 process; data curation; AMPL integration

10 References

[1] Bian, Y., & Xie, X.-Q. (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

[2] Spiegel, J. O., & Durrant, J. D. (2020). Autogrow4: An open-source genetic algorithm for de novo drug design and lead optimization. Journal of Cheminformatics, 12(1). https://doi.org/10.1186/s13321-020-00429-4

[3] McLoughlin, K. S., Shi, D., Mast, J. E., Bucci, J., Williams, J. P., Jones, W. D., Miyao, D., Nam, L., Osswald, H. L., Zegelman, L., Allen, J., Bennion, B. J., Paulson, A. K., Abagyan, R., Head, M. S., & Brase, J. M. (2023). Generative Molecular Design and Experimental Validation of Selective Histamine H1 Inhibitors. https://doi.org/10.1101/2023.02.14.528391
[4] Ulrich, N., Goss, K. & Ebert, A. (2021). Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation
https://www.nature.com/articles/s42004-021-00528-9