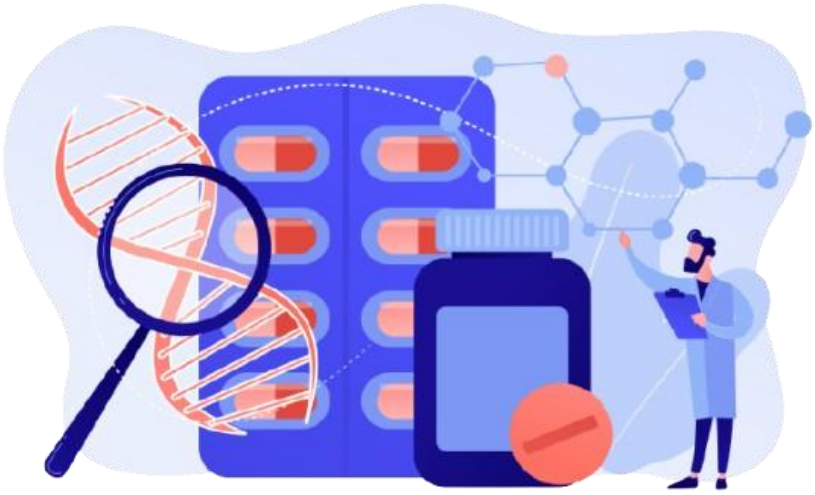
ACCELERATING DRUG DISCOVERY WITH A MOLECULAR FRAGMENT BASED MACHINE LEARNING BIOCHEMISTRY



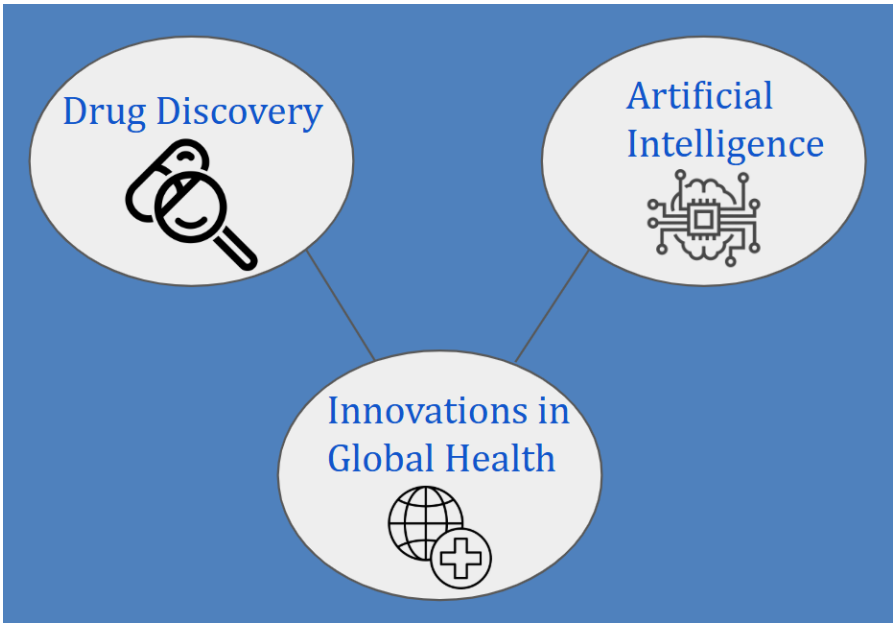


Table of Contents

Contents

Abstract

Machine-Learning is starting to emerge in modern drug discovery by enabling advanced methods to accelerate certain processes. I studied three computational strategies to illustrate how machine learning can address critical drug discovery challenges. First, I explored an unsupervised learning-based approach on SARS-CoV-2 and lung-cancer inhibitors. From 3d structural fragmentation, I created drug sets, and ranked them using a custom Transformer VAE pipeline, and compared the ranking to docking scores. My results demonstrated that unsupervised learning can effectively cluster molecules and match predicted Molecular Docking silico data. Second, I applied a ChemBERTa-based drug classification pipeline to categorize small molecules by therapeutic class (cancer, anti-inflammatory etc.). This method uses representations of SMILES strings to achieve a good classification accuracy, to accurately determine the therapeutic classes of a drug. Third, I developed a specialized a fragmentation predictor for lung -cancer inhibitors, identifying “hot” fragments from multiple known drugs used to inhibit lung cancers. Fragment Drug-Discovery is a faster method of drug discovery and is more cost effective and faster compared to traditional methods, which is why it is essential to develop “hot” fragments that when used to form drugs, they can inhibit lung cancers effectively. These strategies highlight the versatility of using machine learning for drug discovery. My findings develop the concept of computational models emerging as faster and more efficient compared to existing methods and could become widespread in the future. Finally, my findings and models are available on my website so that others can test them and build off my discoveries.

Acknowledgements

I would like to thank my parents and teacher(s) for their wonderful help and support.

Introduction

Currently, pharmaceutical research relies on high-throughput screening (HTS), in which millions of potential drug molecules are tested against a target of interest. Even though this process has been very effective at yielding many medicines, it remains very costly and resource intensive requiring certain professionals and significant capital investment. Recognizing these challenges emerged more targeted approaches, particularly Machine Learning methods to simplify the screening process. One such process of this is Fragment Based Drug Discovery (FBDD) in which rather than searching massive libraries of fully developed molecules, FBDD focuses on smaller “fragments” of these molecules which can be merged or expanded into overall drugs. When combining this process with Machine Learning, it is far cheaper, faster, and in most cases much more accurate and efficient than High Throughput Screening.

My project leverages these processes where Machine Learning can be combined with Drug Discovery to answer the question: “Can Machine Learning drug discovery approaches mimic existing methods and literature such as high-throughput-screening?” My project assesses 3 distinct areas of this.

1. How well can unsupervised Machine Learning group together a dataset of compounds to rank them from most effective of inhibition to least effective of inhibition?
2. How well can Machine Learning determine the therapeutic class of drugs?
3. How well can Machine Learning develop fragment leads from existing drugs which can be used to create potential inhibitors?

I hypothesize that when integrating machine learning to these 3 points, the models will be able to compete with existing methods.

Review of Literature

On average, drug discovery and development for a single small-molecule drug takes about 15 years and approximately 2 billion U.S. dollars in which the resource-intensive process involves the identification and development of new therapeutic drugs (Sadybekov 23). Traditional methods such as high-throughput screening (HTS) evaluate vast libraries to identify potential drug candidates. Though effective, these methods are often time-consuming, expensive, and can result in the oversight of promising candidates (Aldewachi 21).

In order to achieve more efficient and accurate processes of drug discovery, approaches that include Machine Learning in relation to existing approaches have been effective. These include of using supervised learning and unsupervised learning in order to train datasets and achieve familiarity with drug candidates in order to predict and generate drugs that can be suitable for different diseases. Approaches like Clustering drugs, fragmenting drugs, and identifying use cases for drugs have been especially useful for these kinds of practices and will be some common methods that I test.

Fragment-Based Drug Discovery (FBDD) is a useful alternative approach that has gained traction in recent years, particularly for its efficiency and cost-effectiveness compared to existing methods(Bon 22). Rather than creating an entirely new therapeutic compound or drug, FBDD focuses on the creation of small chemical fragments, simplifying the identification of drug candidates compared to traditional methods (Fusco 17). These fragments yield a wider array of chemical structures generating the creation of drug interactions (Chen 09).

To create effective fragments, fragments must follow a set of rules. For example, fragments must adhere to the “Rule of Three”, a set of guidelines describing the ideal physicochemical properties of molecules in FBDD. This guideline defines that ideal fragments should have a molecular weight of less than 300 daltons, a ClogP value, a measure of lipophilicity below three, fewer than three hydrogen bond donors and acceptors, and fewer than three rotatable bonds (Jhoti 13). These limitations ensure that fragments are small enough to be utilized but diverse, while still maintaining desirable properties required for binding and interaction.

Statement of Problem

Do Unsupervised and Supervised Machine Learning models trained on drug datasets compare and efficiently outperform existing drug discovery processes like High Throughput Screening(HTS) .

Hypothesis

My project leverages these processes where Machine Learning can be combined with Drug Discovery to answer the question: “Can Machine Learning drug discovery approaches mimic existing methods and literature such as high-throughput-screening?” My project assesses 3 distinct areas of this.

4. How well can unsupervised Machine Learning group together a dataset of compounds to rank them from most effective of inhibition to least effective of inhibition?
5. How well can Machine Learning determine the therapeutic class of drugs?
6. How well can Machine Learning develop fragment leads from existing drugs which can be used to create potential inhibitors?

I hypothesize that when integrating machine learning to these 3 points, the models will be able to compete with existing methods.

Materials

Hardware/Computer:

A personal Laptop and Desktop (Custom Built Desktop with Windows 10 and an Nvidia graphics card) in order to build, run, test, and analyze the Machine Learning models.

Machine Learning Software & Libraries:

Python (version 3.12)

RDKit(SMILES parsing, canonicalization, fragmentation).

PyTorch and Tensorflow for building the Machine Learning model

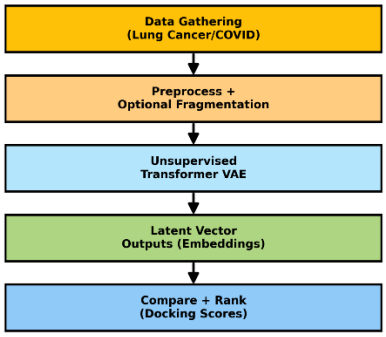
Pandas, NumPy, matplotlib for data handling

Dataset:

A CSV file containing known cancer and covid inhibitors, their SMILES structures, and possibly metadata such as Indications and Targets. This will be created through Publicly available data from ChemBl and BindingDB.

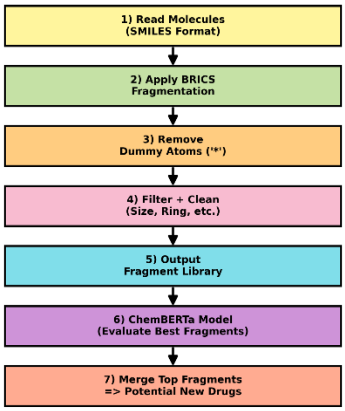
Procedure

Sorting Procedure:



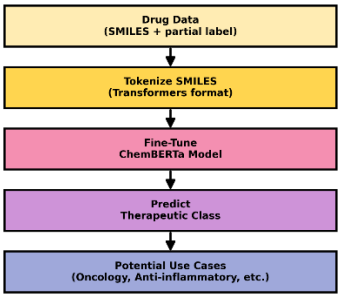
In this pipeline, molecular datasets for both Cancer and COVID are collected in SMILES format. Each molecule's SMILES string is fed into a Transformer Variational Autoencoder(unsupervised machine learning model), which learns to generate and reconstruct chemical fragments without direct supervision. By adjusting network parameters to minimize reconstruction error and a KL divergence penalty, the VAE's latent space captures different patterns of molecular structure. After training, generated fragments or full molecules can be ranked for potential binding affinity, then compared to in silico docking data to validate their predicted viability.

Fragmentation Procedure:



Starting from a curated set of drug molecules (lung cancer inhibitors), a BRICS algorithm in RDKit splits them into chemically meaningful fragments. These fragments are then filtered to remove dummy atoms, enforce size constraints, and remove unstable substructures. Once a clean “fragment library” is obtained, a ChemBERTa model is used to evaluate which fragments appear most effective for a given therapeutic aim. Finally, top-ranking fragments can be recombined to propose new drug candidates that inherit potency from each of the drug fragments.

Drug Classification Procedure:



For the drug classification feature, known drugs across multiple therapeutic domains (anticancer, cardioprotective, anti-inflammatory) are represented in SMILES form. A ChemBERTa-based model is fine-tuned on these labeled examples to learn associations and structural patterns between molecular features and therapeutic class. In this process, a newly proposed molecule is tokenized and fed into the model, yielding probabilities across the learned classes. Because of this, Researchers can assign a likely usage domain (e.g., oncology vs. cardiovascular), guiding early experimental design and hypothesis testing before extensive in vitro validation.

Results and Statistical Analysis

Our machine learning model developed in this study utilizes the ChemBERTa algorithm, a sophisticated Masked Language Algorithm using a generative AI learning process like the Chat GPT generative AI model, in conjunction with RDKit Chemistry Python packages for molecular visualization and optimization.

Drug Ranking:

A batch of 300 lung cancer compounds and 3000 COVID compounds were giving to a VAE unsupervised machine learning model. The model tracked patterns, correlations, and ranked these compounds from most effective to least effective determined by how well the drugs would inhibit cancer proteins and the SARS protein.

Fragmentation:

A batch of 15,000 lung cancer compounds were broken into fragments which were trained on a ChemBERTa model. The model looks at the SMILES strings of the drugs, best performing fragment, and other 3d/pharmacokinetic properties in order to correlate specific fragments with better inhibition. This makes it easier to identify certain “hot” fragments associated with a drug. In our machine learning algorithm, we trained our ChemBERTa model on the dataset of docked fragments. The training process included feature extraction of chemical and structural information on the original drug and the best-performing fragment. Our training involved 10 epochs and used machines with 108 core CPUs, leveraging CUDA machine learning processes through the implementation of 18 Nvidia RTX 4090 GPUs. The model’s performance was assessed using Tanimoto similarity, which measures the structural similarity between two chemical compounds, calculated by comparing their molecular fingerprints and determining the ratio of the intersection to the union of the bits set in their binary representations (PubChem, 2019). A Tanimoto similarity of 0.85 is desired as a high-performing and high-accuracy model.

In addition to fragment prediction, I implemented a method to combine two fragments into one molecule using RDKit. This approach utilizes ester linkers or other appropriate linkers based on the structure of the fragments to ensure structural

integrity and functionality (Enamine, 2021). The process involves identifying possible attachment points on the fragments and generating multiple combined structures. This method enhances the diversity and potential effectiveness of the resulting drug candidates.

Drug Classifcation:

A batch of 100,000 drugs with the SMILES string of the drug, its use case(anti infective, antibacterial, cardiovascular, cancer, etc.), as well as 3d/pharmacokinetic properties were trained on a ChemBERTa model. The model was able to associate a certain drug and correlated it with a use case effectively.

Website for Visualization and Interaction:

To facilitate the dissemination and interaction with our findings, I created a website that displays the complete dataset, predicted fragments, docking scores, and combined fragment structures. The website includes an interactive interface allowing users to input a drug molecule and obtain predicted fragments and their docking scores. Additionally, users can visualize the combined fragments and their predicted binding poses within the target proteins.

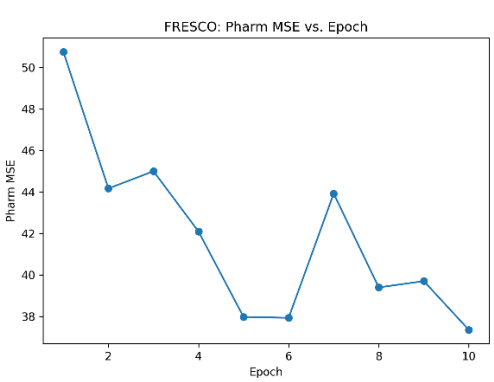
Website:

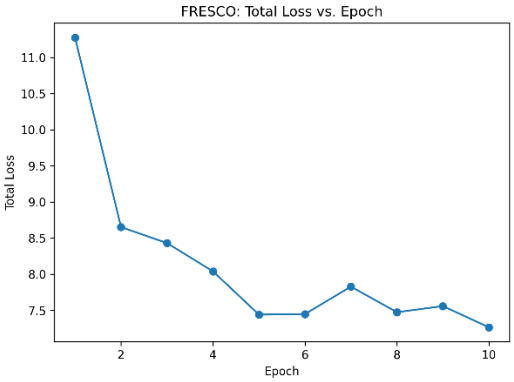
https://nisargrhino.github.io/SmileBERTa-portal/

My website showcases all the Machine Learning models discussed in this project, and users can test and visualize these models.

Analysis

Drug Ranking:





A batch of 300 lung cancer compounds and 3000 COVID compounds were giving to a VAE unsupervised machine learning model. The model tracked patterns, correlations, and ranked these compounds from most effective to least effective determined by how well the drugs would inhibit cancer proteins and the SARS protein. **The Mean Squared error of this model was 37 and the loss after 10 epochs was 7.5.**

Top 5 Cancer Drugs ranked by the model vs. Actual rank of inhibition

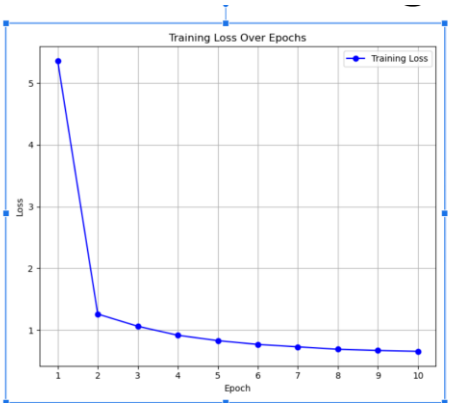
Name	Predicted Rank	Actual Rank	Model Score
Lorlatinib	1	1	30.2
Repotrectinib	2	2	25.33
Thioguanine	3	4	18.41
Mercaptopurine	4	7	17.13
Fluorouracil	5	12	16.38

The model appears to be “somewhat” accurate as the 1st and 2nd compound matched the actual rank when docking the inhibitors to lung cancer proteins.

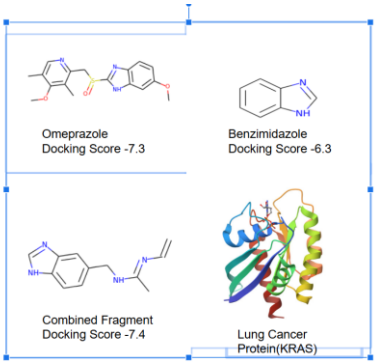
However, the 3rd-5th compounds do not match the actual rank from the Model score. However, this approach shows potential as with more training, the model exponentially increases in accuracy.

Model score was created by the unsupervised model to predict how well each compound would be able to inhibit the lung cancer/ covid proteins. It uses structural properties and creates nodes and patterns to generate this score.

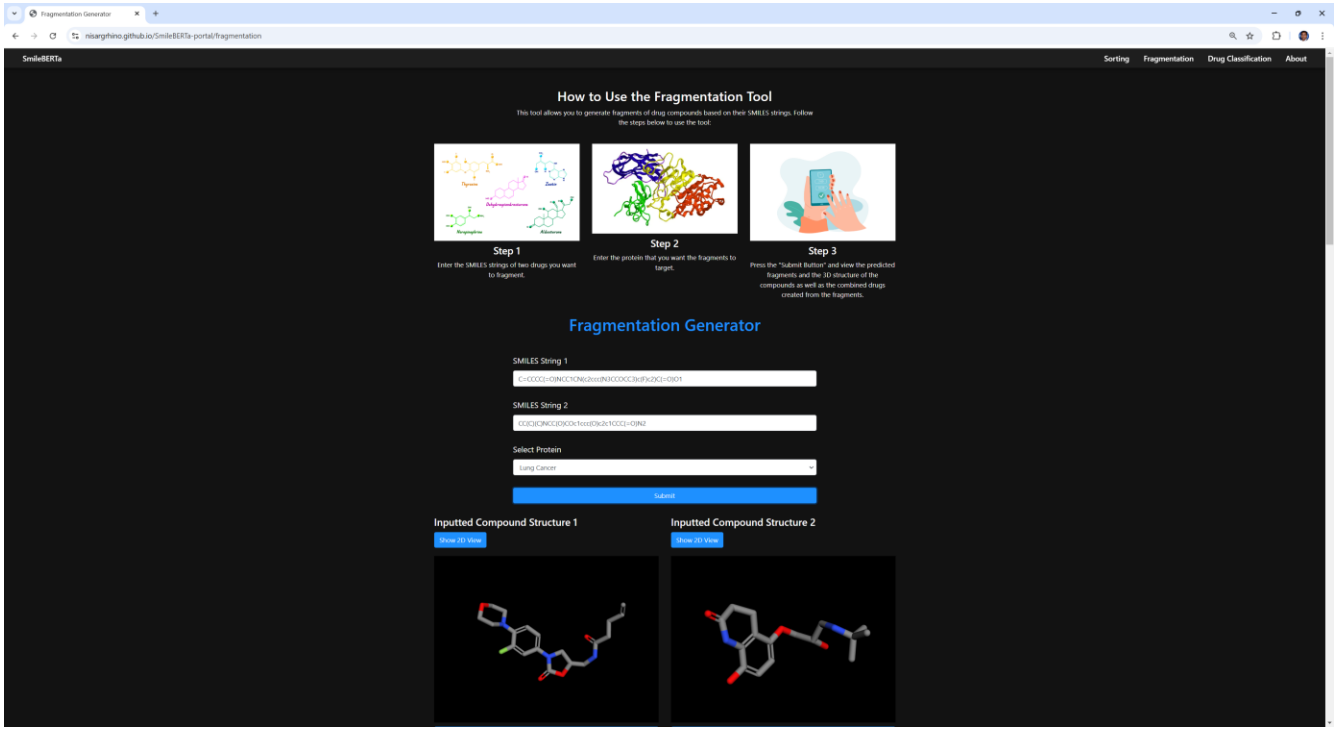
Fragmentation:



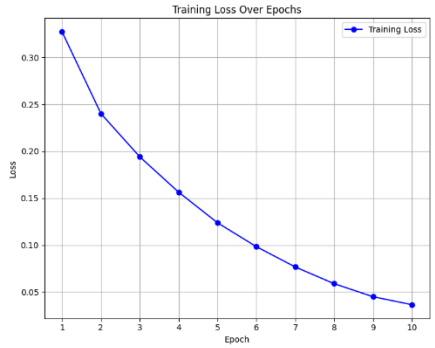
After running the Machine Learning Model for **10 epochs**, the Training loss was **0.52** and the accuracy of the model for predicting the best fragment was **60%.** The accuracy was determined through Tanimoto Similarity, in which the predicted drug fragments were compared to the original drug fragments through molecular structure. This proves an accurate model was created which identified a drugs “best” fragment for inhibition of lung cancer proteins.



My website (https://nisargrhino.github.io/SmileBERTa-portal/fragmentation): Users can test out the model and visualize the model predict the best “hot” fragments for the existing drug.

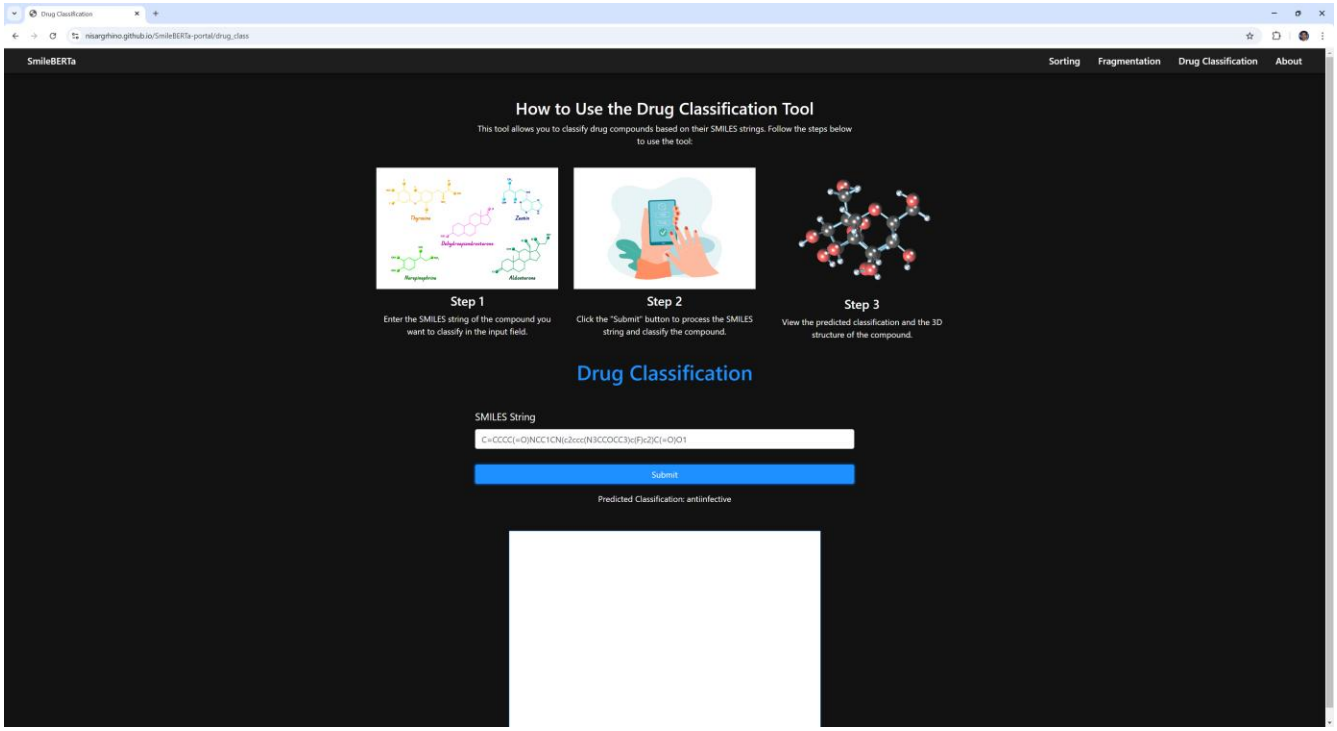


Drug Classification:



The loss of the model was **0.031** and the accuracy of determining the use case of the drug was **90%**. The accuracy was calculated by comparing the use case prediction to the actual use case of the original drug. This proves an accurate model that was created in order to effectively predict the use case for any drug.

My website (https://nisargrhino.github.io/SmileBERTa-portal/drug_class): Users can test out the model and visualize the model predict the best “hot” fragments for the existing drug.



The drug classification component fulfills a key role in evaluating novel compounds by predicting their most likely therapeutic categories. Although it was initially trained on existing molecules annotated with known clinical indications, its predictive capability extends to newly designed drugs. By inputting the molecular structure or drug into the classification model, users can see whether the compound aligns with oncology, anti-inflammatory, cardiovascular, or other therapeutic domains. This also streamlines early hypothesis testing in which people can generate new leads before in vitro testing.

Conclusion

Overall, this project successfully demonstrates how Machine Learning can be integrated into drug discovery, by offering a faster and more effective alternative to traditional approaches. Three computational approaches were explore: unsupervised clustering for drug ranking, drug classification, and a fragment based predictor for lung cancer inhibitors.

The results indicate that machine learning can effectively analyze and rank drug molecules based on inhibition potential, with a Transformer VAE model correctly predicting the top-ranked lung cancer inhibitors with moderate accuracy.

The ChemBERTa-based drug classification model achieved 90% accuracy in categorizing drugs by therapeutic use, demonstrating its reliability for rapid screening of novel compounds.

Finally, the fragmentation model successfully identified "hot" fragments, with combined fragments outperforming individual inhibitors in docking simulations against KRAS, highlighting its potential for lead optimization.

While the models showed promising results, improvements like more advanced datasets, tweaking the training parameters, and integration with experimental validation could enhance accuracy. These findings reinforce the role of AI in accelerating drug discovery, and also pave the way for further advancements in computational pharmacology. Future work will focus on refining these models and expanding their applicability to a wider range of diseases.

Recommendations

Tweaking the parameters of the Machine Learning model, increasing the dataset size, and increasing the bandwidth of diseases, drugs, and cases we train can be improved on in order to generate a more general Machine Learning Algorithm. In addition to this, Machine Learning and drug discovery are fields that have just start gaining popularity, and as both fields progress, there will be newer and stronger models and data to use for drug discovery use cases and there is room for improvement each year.

Works Cited/Bibliography

[1] V. Sadybekov, V. Katritch. “Computational Approaches Streamlining Drug Discovery.” Nature, Nature Publishing Group, 26 Apr. 2023, https://www.nature.com/articles/s41586-023-05905-zciteas.

[2] H. Aldewachi, et al. “High-Throughput Screening Platforms in the Discovery of Novel Drugs for Neurodegenerative Diseases.” Bioengineering (Basel), U.S. National Library of Medicine, 23 Feb. 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7926814/.

[3] Zdrazil, B., et al. “The CHEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods.” Nucleic Acids Research, Oxford University Press, 2 Nov. 2023, https://academic.oup.com/nar/article/52/D1/D1180/7337608.

[4] University of Cambridge, UK. “Essential Fragment Library.” Enamine, 2023, https://enamine.net/compound-libraries/fragment- libraries/essential-library.

[5] H. Jhoti, G. Williams, D. Rees, C. Murray. “The ‘Rule of Three’ for Fragment-Based Drug Discovery: Where Are We Now?” Nature Reviews Drug Discovery, vol. 12, no. 7, July 2013, https://www.nature.com/articles/nrd3794.