**DruGen**

AI-Driven Molecular Generation and Property Prediction for Drug Discovery

**Introduction**

Brief Background

Drug discovery is one of the most intensive and time-consuming processes in healthcare, often requiring over a decade and billions of dollars to bring a single drug to market. Artificial Intelligence (AI) has shown tremendous promise in transforming this space by enabling *de novo* drug design, and molecular property prediction.

One of the most powerful chemical representations used in AI-driven drug discovery is the SMILES (Simplified Molecular Input Line Entry System) notation, which encodes chemical structures into text strings. Combined with graph-based generative models and property predictors, AI can design novel molecules with desired properties in a fraction of the traditional time.

Motivation

The motivation of this project is to leverage generative models to generate novel molecules and then evaluate them using a SMILES Property Predictor for critical properties such as solubility, toxicity, drug-likeness (QED), and more. This pipeline helps in identifying promising candidates early, thereby reducing the experimental search space in drug discovery.

This work bridges molecule generation and property prediction, providing an end-to-end AI system for *rational drug design*.

**Objectives**

The primary objectives of this project are:

1. Develop a Conditional Molecule Generator that can generate novel molecules based on given data
2. Integrate a SMILES Property Predictor to evaluate generated molecules for 12 molecular properties including logP, TPSA, QED, solubility, and toxicity.
3. Provide a scoring mechanism to assess how "promising" a molecule is for drug discovery.
4. Demonstrate the utility of AI in healthcare by automating an early-stage drug discovery workflow.

**Dataset Description**

* Source of Data:  
  The dataset is obtained from the official repository of *Multi-Objective De Novo Drug Design with Conditional Graph Generative Model*:  
  [Datasets.tar.gz](https://github.com/kevinid/molecule_generator/releases/download/1.0/datasets.tar.gz)
* Nature of the Data:  
  The dataset contains molecular structures in SMILES format, ESOL and Tox21 data is annotated with various properties such as QED (Quantitative Estimate of Drug-likeness), logP (lipophilicity), and synthetic accessibility (SAscore).
* Time Period Covered:  
  Not time-based, as molecules are static chemical entities.
* Number of Data Points:  
  The dataset contains hundreds of thousands of molecules, ensuring more than sufficient data points (>500 required) for meaningful analysis.

**Team Members and Responsibilities**

* Team Member 2: *[Shreeya Pandey]* – Property Prediction and Exploratory Data Analysis
* Team Member 3: *[Aditya Raj]* – Model Training and Molecule Generation.

**Expected Outcomes**

Through this project, we aim to achieve:

* A conditional molecule generator capable of producing molecules which can be a possible drug candidate.
* A property predictor that evaluates 12 key molecular descriptors and ML-predicted properties (solubility and toxicity).
* A scoring mechanism that evaluated molecules based on their “drug-likeness” and healthcare applicability.
* Insights into how AI can accelerate drug discovery pipelines and reduce dependence on costly wet-lab experiments.

**Timeline**

| Week | Task | Deliverable |
| --- | --- | --- |
| Week 1 | Dataset preprocessing & setup | Clean SMILES dataset |
| Week 2 | Implement Molecule Generator | Generator code ready |
| Week 3 | Implement Property Predictor (RDKit + ML models) | Predictor ready |
| Week 4 | Integrate generator + predictor | End-to-end pipeline |
| Week 5 | Testing and analysis | Results with promising molecules |
| Week 6 | Write final report & presentation | Final report |

**References**

1. You, J., Liu, B., Ying, R., Pande, V., & Leskovec, J. (2018). *Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation*. NeurIPS.  
    <https://arxiv.org/abs/1806.02473>
2. Li, Y., Vinyals, O., Dyer, C., Pascanu, R., & Battaglia, P. (2018). *Learning deep generative models of graphs*. ICML.  
   <https://arxiv.org/abs/1803.03324>
3. Jin, W., Barzilay, R., & Jaakkola, T. (2018). *Junction Tree Variational Autoencoder for Molecular Graph Generation*. ICML.  
    <https://arxiv.org/abs/1802.04364>
4. RDKit: Open-Source Cheminformatics Software.  
    <http://www.rdkit.org/>
5. DeepChem: Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry, Materials Science and Biology.  
    <https://deepchem.io/>
6. Dataset: *Multi-Objective De Novo Drug Design with Conditional Graph Generative Model*.  
    <https://github.com/kevinid/molecule_generator>
7. Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). *Molecular de novo design through deep reinforcement learning*. J. Cheminformatics.  
    <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0235-x>

**Conclusion**

This project demonstrates how AI can transform drug discovery by combining graph-based molecule generation with property prediction. By generating molecules and evaluating them for solubility, toxicity, and other critical features, we create a powerful tool for identifying promising drug candidates.

The challenges include handling the vast chemical search space, ensuring robust predictive models, and validating generated molecules against real-world pharmacological constraints. Despite these challenges, the project provides a concrete example of how AI can significantly accelerate and optimize early drug discovery.