PROBLEM DESCRIPTION

The exploration of potential drug molecules encompasses an estimated 10^{23} to 10^{60} unique structures, representing a vast chemical space. The problem is, that compound libraries that exist today consist of just millions of compounds, leaving a mostly unexplored domain. As such, bridging this gap between a vast total chemical space, and the relatively small subset that has been investigated, remains a central challenge in the field of drug discovery.

Virtual screening methods are commonly used in identifying small molecules by using information about existing molecules and their interactions with proteins. However, this method can often lead to the discovery of molecules that are similar to those already known, diminishing the uniqueness of the findings. Additionally, this approach depends on large databases of known small molecules, which can be a limiting factor for targets that do not have many known interacting molecules. De novo design methods aim to generate new molecules by applying theoretical principles and knowledge of molecular structures. Advances in deep learning, including generative adversarial networks and reinforcement learning, have improved these methods. Nevertheless, their effectiveness can be limited by the requirement for comprehensive structural data or specialized libraries tailored to specific target proteins.

Incorporating structural data, such as 3D protein structures, into drug discovery can be complex. Obtaining these structures is often resource-intensive, and for many proteins, the information is still unknown. Moreover, predicting protein structures computationally can be unreliable, especially for proteins with fewer known contacts within their chain or without similar structures for reference. Models like CProMG and AlphaDrug are notable for their capacity to create new, drug-like molecules, yet they depend heavily on detailed structural data and require significant computational work. These demands can limit how much these models can scale and the range of chemical space they are able to investigate.

These issues suggest a need for a methodological shift towards a more direct, sequence-based strategy in drug discovery. The focus is not on simplifying the process, but rather to enable the exploration of areas of the chemical space that were previously hard to reach and to take advantage of the rapid progress in protein sequencing. Models such as DeepTarget and others using RNN-based machine translation have already begun to adopt this approach.

Herein, we propose a pioneering latent sequence diffusion model, *'MolDetective'*, a novel algorithm that outputs a protein-targeted small molecule binder given only a protein **sequence** input.

MODELING APPROACH

Dataset Curation

The data for this effort will be sourced from BindingDB, which houses over 2.3 million records of binding data. This data includes in-depth information about protein-ligand binding strengths, supplemented with structural information from the PDB. Following a methodology similar to that used in the AlphaDrug study, our criteria for selecting data will include:

- Human proteins with high-affinity ligand interactions, specifically those with IC50, Kd, or EC50 values below 100 nM.
- Ligands that have a PubChem CID, can be represented by SMILES, have a molecular weight below 1000 Da, and are linked to protein identifiers (Uniprot ID).
- Protein sequences that are 80 to 1000 amino acids in length, and proteins that function as monomers.

Limitations and Dataset Improvements

- To improve the dataset generation for future model development, we could refine the dataset by sequence clustering at the 30% identity level using the tool Mmseqs2. This way, the datasets cover a diverse range of sequences for training, validation, and testing. However, due to time constraints, this could not be completed.
- The validation datasets should have also been carefully assembled by pairing PDB protein-ligand co-crystal structures with BindingDB entries. This can be done using specific criteria such as the maximum common substructure (MCSS) and Tanimoto similarity indices, to ensure the datasets are both comprehensive and applicable for the model's training and testing.

Model Architecture

The architecture commenced with the generation of protein and molecule embeddings. Protein sequences were processed using a pre-trained ESM-2 model, a transformer-based model adept at capturing the complexities of protein structures. For each protein sequence, embeddings were generated and averaged (excluding start and end tokens) to create a fixed-size representation. Concurrently, molecule embeddings were generated from SMILES strings using the UniMol representation tool. This involved creating 3D conformers for each molecule and extracting embeddings that encapsulate the molecular structure and chemical properties.

The latent diffusion-based model, central to learning the relationship between protein and molecule embeddings, was intricately designed with specific neural networks and layers to ensure efficient and accurate learning.

- Variance Scheduler
    - The variance scheduler, a critical component of the model, controlled noise levels throughout the diffusion process. It ensured a gradual and controlled transition of molecule embeddings from a noisy state to their denoised counterparts. This maintained the stability of the diffusion process and prevented abrupt changes in the embeddings.
- Epsilon Network
    - The epsilon network, constructed using fully connected layers, played a pivotal role in the diffusion model. It consisted of:

        - A first linear layer (fc1) with an input size of 1280 (matching the protein embedding size) and an output size of 512, was the first level of transformation.
        - A second linear layer (fc2), maintaining the dimensionality at 512, further processed the embeddings.
        - A final linear layer (fc3), also with an output size of 512, completed the transformation process.
        - Each layer was followed by a ReLU activation function, introducing non-linearity to the model, essential for capturing complex patterns in the data.
- Small Molecule Continuous Transition Module
    - The transition module, another cornerstone of the architecture, was tasked with the controlled addition and removal of noise from molecule embeddings, thereby facilitating the diffusion process. This module was conditioned on the protein embeddings and included:
        - A VarianceSchedule component dynamically adjusted the variance levels applied to the embeddings.
        - An add_noise_layer, a linear layer (512 input and output size), added calculated noise to the molecule embeddings.
        - A denoise_layer, a linear layer combined molecule embeddings (512 dimensions) and protein embeddings (1280 dimensions), for a total input size of 1792, and reduced them back to 512 dimensions. This layer was pivotal in integrating protein information into the molecule embeddings.
        - A denoise_mlp, a sequential module consisting of linear layers and ReLU activations, interspersed with dropout for regularization. This multi-layer perceptron further refined the denoising process, ensuring that the molecule embeddings were accurately reconstructed from their noised versions.

MODEL EVALUATION

Currently, the evaluation methods use PCA of generated embeddings against the ground-truth embeddings, in a 2-dimensional plane. We then take the output 512 features and take a sample protein (1), and then we plot a scatterplot of the 512 features against each other, to see if the generated embedding features resemble the ground-truth molecular embeddings.

To further evaluate our model, we have created a clever, yet simple, look-up table decoding scheme. The look-up table decoding scheme is a vital component for translating embeddings back into meaningful biological and chemical sequences.

- Protein Embedding to Sequence Mapping:
    - A dictionary (protein_embedding_sequence_dict) is created to map original dataset protein embeddings to their respective protein sequences. It is important to note that this mapping is the ground truth embeddings and sequences. This is achieved by iterating through the dataset, where each unique protein embedding is stored as a key, and its corresponding protein sequence as the value.
- Molecule Embedding to SMILES Mapping:
    - Similarly, another dictionary (embedding_smiles_dict) is constructed to map ground truth molecule embeddings to their corresponding ground truth SMILES strings. This look-up table ensures that each unique molecule embedding is directly associated with its SMILES representation.
- Nearest Neighbor Search
    - A function 'find_closest_embedding' is defined to find the closest matching SMILES string for a given generated molecule embedding. This function iterates through the embedding_smiles_dict, calculating the Euclidean distance between the generated embedding and each embedding in the dictionary. The SMILES string corresponding to the embedding with the minimum distance is then selected as the closest match.
- Evaluation and Visualization
    - Distance Calculation and Data Aggregation:
        - For each generated embedding, the closest SMILES string is identified (via Euclidean Distance calculation), and its distance is recorded. These details, along with the ground truth SMILES, are collated into a DataFrame for analysis.
    - Tanimoto Similarity Measurement:
        - Using RDKit, Tanimoto similarity scores are calculated between the original and decoded SMILES strings. This metric provides a quantitative assessment of how closely the decoded SMILES match the original ones.
    - Visualization:
        - The results are visualized using scatter and box plots, showcasing the distribution of Tanimoto similarities, which helps in understanding the decoding accuracy and consistency across the dataset.

There are some shortcomings of this look-up table decoding scheme which lends way to the LSTM decoding scheme. The table scheme relies on existing embeddings in the look-up table, which restricts its ability to generalize to unseen data. It can only decode embeddings that closely resemble those in the training set, limiting its applicability to novel protein or molecule embeddings. The look-up tables provide a static mapping and do not account for the dynamic nature of biological systems. They cannot capture the nuances of changes in protein or molecular structures under different conditions. In addition, If certain proteins or molecules are underrepresented in the training data, the decoder's ability to accurately reconstruct these entities is compromised.

In the future, we aim to create a custom LSTM decoder. The LSTM decoder in the context of translating molecule embeddings to SMILES strings functions as a sequence generator, utilizing the inherent properties of LSTM networks to deal with sequential data. In this process, the molecule embedding, a fixed-size representation capturing the molecular structure, acts as the initial state or context for the LSTM. The LSTM then generates a sequence of tokens, one at a time, each representing a character in the SMILES string.

At each step of the sequence generation, the LSTM receives an input (which could be the molecule embedding or a function of the previously generated tokens) and updates its internal state. This state encapsulation is key, as it allows the LSTM to maintain a memory of what has been generated so far, enabling it to make informed decisions about subsequent characters in the sequence. The output at each step is a probability distribution over the possible characters in the SMILES vocabulary, from which the most likely character is selected and appended to the growing SMILES string.

This approach, however, can still be further optimized. The use of repeated embeddings as input to each LSTM step, while straightforward, might not be the most effective way to leverage the LSTM's potential. Alternatives could involve dynamically updating the input at each step based on the previously generated token or incorporating attention mechanisms to focus on different aspects of the molecule embedding at different stages of the sequence generation. These refinements could potentially lead to more accurate and contextually relevant SMILES generation, making the most of the LSTM's capabilities in handling sequential and contextual information.

After generating the SMILES, we have an existing scheme (from the look-up table) to generate structures from the SMILES, using the RDKit package and then calculating the Tanimoto similarity of the generated structures against the ground-truth embeddings.

Furthermore, the further, more complex evaluation of generated molecules will be multifaceted, using the following criteria:

- Docking Score: Leveraging SMINA and various in-silico benchmarking tools (AutoDock, rDock, etc) for binding affinity calculations, where higher scores indicate better docking and potential for high binding affinity.
- Uniqueness: Assessment of the model's ability to generate diverse molecules for different proteins.
- LogP: Ensuring the lipophilicity and solubility of the molecules are within the desired range.
- QED: Quantitative Estimate of Drug-likeness will be used to measure the drug-likeness of the molecules.
- SA Score: Synthetic accessibility will be gauged to ensure practicality in drug synthesis.
- NP-likeness: The natural product likeness score will be used to assess the potential of the molecules to be derived from or inspired by natural products.

To validate our model's performance, we will benchmark against contemporary models, and also do a virtual screen using DrugCLIP, which employs a contrastive learning framework for virtual screening, enabling fast screening over large-scale chemical libraries without relying on explicit binding-affinity scores. This comparison will allow us to demonstrate our model's relative strengths and efficacy in drug discovery tasks.

CHALLENGES AND ALTERNATIVE APPROACHES

In the realm of employing diffusion models for generating molecule embeddings and their subsequent translation into SMILES strings, several challenges arise. One key challenge is capturing the complexity of molecular structures within embeddings accurately. Molecules are inherently complex, and their embeddings must encapsulate various chemical properties and structural nuances. Additionally, the diffusion process involves a delicate balance between adding and removing noise. Ensuring that this process preserves the essential molecular information while also allowing for meaningful transformations is a non-trivial task. Generalization is another significant hurdle, as the model must be able to handle a wide variety of molecular structures, including those not present in the training data. This is particularly challenging in the context of drug discovery, where novel compounds are continuously being synthesized.

To address these challenges, alternative approaches can be considered. For instance, enhancing the molecule embedding process with more sophisticated models or additional contextual information might lead to more robust and informative embeddings. In the diffusion process, advanced techniques like adaptive noise scheduling or incorporating attention mechanisms could offer more control and precision. Finally, for the decoding phase, going beyond standard LSTM decoders to models that can dynamically adapt to the input embeddings at each step, or employing more advanced sequence generation techniques like transformer models, could improve accuracy and generalization.

OUTCOMES AND SUCCESS

- **Accuracy of SMILES Reconstruction:** One of the primary indicators of success is the accuracy of the reconstructed SMILES strings. This can be quantitatively measured using similarity metrics like the Tanimoto coefficient, comparing the generated SMILES strings against the ground truth. High similarity scores indicate that the model accurately captures and reproduces the molecular structures.
- **Chemical Validity:** The chemical validity of generated SMILES strings is crucial. The model's success can be partially measured by the proportion of generated SMILES that correspond to chemically valid structures. Tools like RDKit can be used to check the validity of these structures.
- **Novelty of Molecules:** In drug discovery, the ability to generate novel molecules that are not present in the training set is highly desirable. Novelty can be assessed by comparing the generated molecules against a database of known compounds.
- **Benchmarking Against Existing Methods:** Comparing the model's performance against existing state-of-the-art methods in molecular generation can offer a perspective on its relative effectiveness and areas for improvement.