De Novo Drug Design as GPT Language Modeling

Gavin Ye
The drug discovery process is time consuming

Identify target protein

Traditionally costly and time consuming

Drug (candidates) designing

Entire process takes 10–15 years

Drug selection

Drug synthesis, testing, etc.

(DiMasi et al., 2016)
GPT = Generative ML that specializes in sequential data

Hello my… name is …
Q: 1+1=? A:2
Q: What's ML? A:ML is …
… …
Chat-GPT is… a large …

Input text with correct response

Training

GPT model

Perform certain task
Such as:
- Text generation
Molecules can be represented using sequences.
GPT specializes in sequential data

Input with correct response

<table>
<thead>
<tr>
<th>Hello my name is …</th>
<th>Training</th>
<th>Perform certain task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q: 1+1=?</td>
<td>A: 2</td>
<td>Such as:</td>
</tr>
<tr>
<td>Q: What's ML?</td>
<td>A: ML is …</td>
<td>- Text generation</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td></td>
</tr>
<tr>
<td>Chat-GPT is…</td>
<td>a large …</td>
<td></td>
</tr>
</tbody>
</table>
GPT has not been used for designing effective drug molecules in previous studies.

GPT specializes in sequential data, performing tasks such as text generation and molecule generation.

Input with correct response desired molecule:

<table>
<thead>
<tr>
<th>Input</th>
<th>GPT model</th>
<th>Perform certain task</th>
</tr>
</thead>
<tbody>
<tr>
<td>C12H6...</td>
<td>CH4...</td>
<td>Text generation</td>
</tr>
<tr>
<td>CO...</td>
<td>OOH</td>
<td>Molecule generation</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>CH3C...</td>
<td>OOH</td>
<td></td>
</tr>
</tbody>
</table>
Brief Recap of Problems: Non GPT models have low validity

Problems:

Sequential representations have been used for non-GPT models for different tasks
(Segler et al., 2018); (Abbasi et al., 2022); (Frey et al., 2022)

Low Validity

(Abbasi et al., 2022); (Yasonik, 2020);
(Popova et al., 2018)

Low Novelty or Efficacy

(Gao et al., 2020);

(Frey et al., 2022) 100% Validity
... Not applied to any specific task such as drug design.

This Study:
GPT applied to Drug Design by optimizing drug efficacy
My study: GPT applied to drug design

**Goal:**

Train GPT to generate drug-like molecules with high efficacy while maintaining validity towards treating a disease.
Methodology

Objectives

1. Evaluate Drug Efficacy
2. Design Drug-Like Molecules
3. Optimize Drug Efficacy

Steps

- Reward Modeling
  - Drug Efficacy Evaluation Model
- Supervised Finetuning
  - Drug Design GPT Model
- Proximal Policy Optimization (PPO)

Steps:

1. Evaluate Drug Efficacy
2. Design Drug-Like Molecules
3. Optimize Drug Efficacy

Methods:

- Reward Modeling
- Supervised Finetuning
- Proximal Policy Optimization (PPO)
For case study, BindingDB dataset with molecules and experimentally determined drug efficacy values are used.

- Training Data (bindingdb.org)
  - drug candidates
  - measured efficacy toward APP (in °p; Η>)

More Effective = \( pIC_{50} = \log_{10} \)

(Liu et al., 2007)

°p; Η> 7 ⇒ Highly Effective (Sydow et al., 2019)
My novel drug efficacy evaluation model design combines both sequential and chemical representations.
My novel drug efficacy evaluation model design combines both sequential and chemical representations.
My novel drug efficacy evaluation model design combines both sequential and chemical representations.
My novel drug efficacy evaluation model design combines both sequential and chemical representations.

Sequential Representation

Chemical Representation

Experimental pIC50 7.3

Predicted pIC50 7.2

1032 Molecules

BindingDB

Sequential Processing Neural Network Layers

Numerical Processing Neural Network Layers

Output
My novel drug efficacy evaluation model design combines both sequential and chemical representations.
My novel drug efficacy evaluation model design combines both sequential and chemical representations. Baseline: Abbasi et al. (2022), previous state-of-art efficacy evaluation model which uses LSTM. Same model structure is used for: SMILES, SELFIES, Mol2Vec.
Combining sequential representation with chemical descriptors improves accuracy.

**A: Mean Squared Error (MSE) by Evaluation Model**

- Abbasi et al: 0.89
- Random Forest (Non LSTM): 7.57
- Mol2Vec Embedding: 0.45
- SMILES Representation: 0.5
- SELFIES Representation: 0.38

**Previous Studies**

**This Study**

**Performance Increases**

**B: R2 Score by Evaluation Model**

- Abbasi et al: 0.55
- Mol2Vec Embedding: 0.79
- SMILES Representation: 0.76
- SELFIES Representation: 0.81

**Previous Studies**

**This Study**

**Performance Increases**

**Effect of Dataset Size on Performance?**
My efficacy evaluation model outperforms baseline model even with less data.

Experiment repeated with different dataset sizes.

Performance Increases

Baseline Model MSE = 0.8931

Outperforms Baseline
Methodology

Objectives

1. Evaluate Drug Efficacy
2. Design Drug-Like Molecules
3. Optimize Drug Efficacy

Steps

Reward Modeling

Supervised Finetuning

Proximal Policy Optimization (PPO)

Drug Efficacy Evaluation Model

Drug Design GPT Model
Methodology

Objectives

1. Evaluate Drug Efficacy
2. Design Drug-Like Molecules
3. Optimize Drug Efficacy

Steps

1. Reward Modeling
2. Supervised Finetuning
3. Proximal Policy Optimization (PPO)

Drug Efficacy Evaluation Model

Drug Design GPT Model
Supervised finetuning training uses the same dataset for designing drug-like molecules.

Dataset from BindingDB

Training

My drug design model

Design molecules with similar properties

Images (225×225), n.d.
Generated molecules exhibit similar properties as ones from the dataset using Supervised finetuning.
Objectives

1. Evaluate Drug Efficacy
2. Design Drug-Like Molecules
3. Optimize Drug Efficacy

Steps

Reward Modeling

Supervised Finetuning

Proximal Policy Optimization (PPO)
Methodology

Objectives

1. Evaluate Drug Efficacy
2. Design Drug -Like Molecules
3. Optimize Drug Efficacy

Steps

- Reward Modeling
- Supervised Finetuning
- Proximal Policy Optimization (PPO)

Drug Efficacy Evaluation Model
Drug Design GPT Model

(first time used for drug design)
**PPO Workflow**

Drug Efficacy Evaluation Model

Predicted Reward (in pIC50)

Drug Design Model

Designed Molecule (with chemical descriptors calculated)

Update parameters to maximize pIC50 reward

*(Images (225 ×225), n.d.)*
Result: PPO effectively optimized efficacy of molecules for drug design for the first time.

100% Validity (used RDKit for validation)

99.2% pIC50 > 7
Limitations & Next Steps

Drug Design Model

GPT

Predicted Reward (in pIC50)

Drug Efficacy

Evaluation Model

Designed Molecule (with chemical descriptors calculated)
Future study can optimize multiple properties of the drug design model using similar methodology.
Future study can optimize multiple properties of the drug design model using similar methodology.

Drug Design Model

Use human chemists to provide feedback (a.k.a RLHF with PPO) (Ouyang et al., 2022)
Brief Recap of Problems:
Non GPT models have low validity

Problems:

- Traditional Drug Discovery
  - Costly and time consuming
  - (Abbasi et al., 2022) Low Validity
  - (Yasonik, 2020) Low Validity

- (Frey et al., 2022) HIGH Validity
  - ...Not applied to any specific task

This Study
- (High Efficacy + Validity)
STEP 1: Novel evaluation model outperforms traditional efficacy evaluation models.

STEP 3: GPT model successfully designed molecules with high efficacy (for the 1st time).

Dataset
1032 Molecules

Drug Design Model

Evaluate Molecule

Drug Evaluation Model

Design Molecule

Drug Design Model

GPT

Evaluation model outperforms traditional evaluation models.
Training procedure is generalizable

Molecule with high efficacy targeting APP
Training procedure is generalizable

Any Protein Targeted Molecules → GPT → Molecule with high efficacy targeting any drug target

Drug Design Model
Significance: My drug design model can speed up drug discovery

Traditionally costly and time consuming

Identify target protein

Drug (candidates) Designing

Drug selection

Drug Synthesis, Testing, etc.
Significance: My drug design model can speed up drug discovery

Identify target protein → GPT → Drug Design Model → Drug selection → Speeding up drug design by providing high quality (effective) molecules quickly.

Drug Synthesis, Testing, etc.


Images (225×225). (n.d.). Retrieved January 23, 2024, from https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQWKthUMueq8uSWPmmCknHu0gYf0AVN32ZvH5k2crmB2X1aAd


GPT For Drug Design

For the first time

More accurate & data efficient

Designed molecules similar to dataset ones.

99.2% pIC50 >7