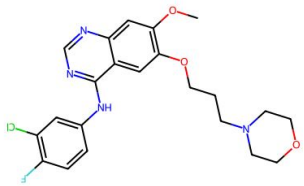
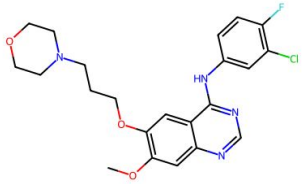
HS-BCOM-134-T2



MediSeeker: An AI-based Pipeline for Personalized Drug Discovery

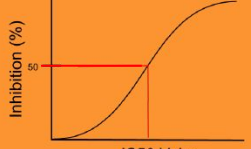
All images and graphics created by the student researchers

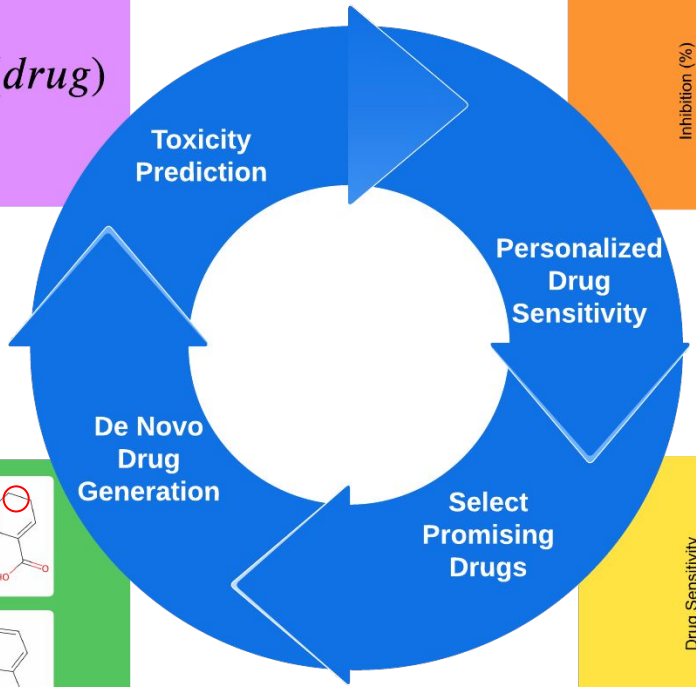


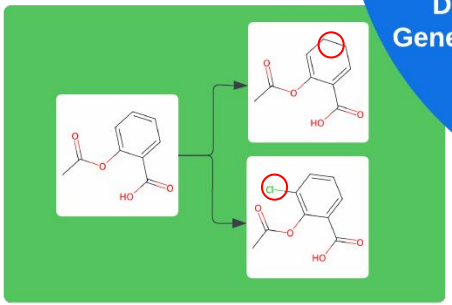
We certify that we have not used AI tools, like ChatGPT, or other AI tools to construct this document or responses to questions in this form. We certify the item is created by team members without any AI.

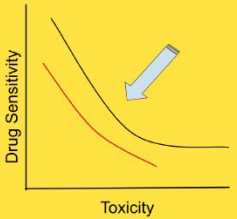
$$hERG_{tox} = f(drug)$$

$$IC_{50} = g(drug, gene\ expression)$$









Text

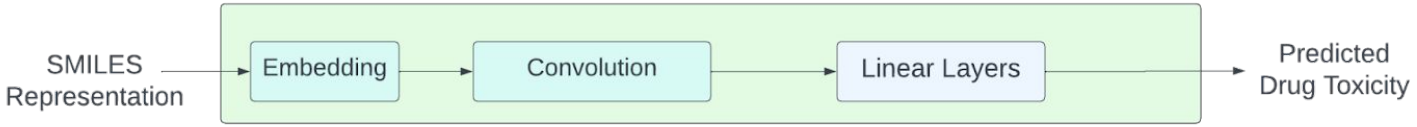
Procedure: Developing Mediseeker

1. Setup:
   a. Train a model to predict the toxicity, measured by the hERG inhibition using data from the hERG central dataset [9] (306,893 drugs).
   b. Train a model to predict the personalized sensitivity of a drug (IC50 value) using the data from the GDSC dataset [10] (92,703 pairs, 805 cancer cell lines, 137 drugs). The cell line gene expression data is obtained from the CCLE [11] dataset (17737 mRNA expression values per cell line).
   c. Train a model for drug generation using the ChEMBL dataset [12] (1,961,462 molecules) which consists of the SMILES representation of many drug like molecules.
   d. Use the GDSC1 dataset to obtain the SMILES format of drugs and the cell line gene expression data to start the iteration process
2. Iteration
   a. Using the trained models, predict the toxicity and the sensitivity for drug, cell line combinations
   b. Find the pareto points showing the best tradeoff between the toxicity and IC50 value for drugs on each cell line
   c. Feed the best performing drugs to the generator to generate multiple variants of each drug
   d. Repeat step 2 until more accurate results are obtained.

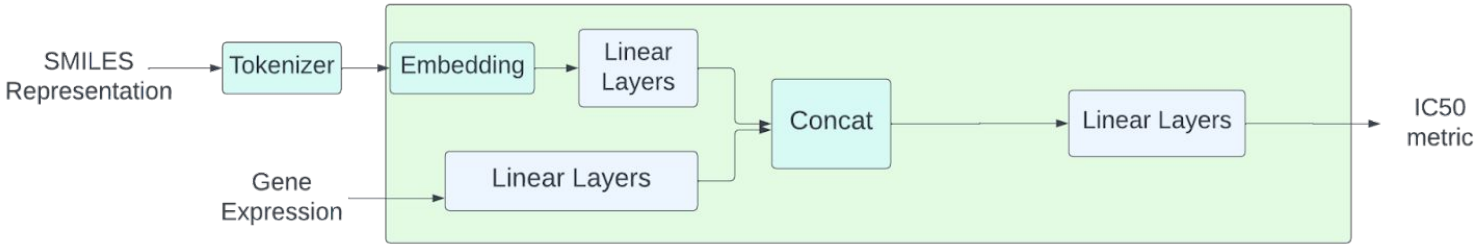
All data was obtained from the tdcommons database [13]

Model Architecture Diagrams

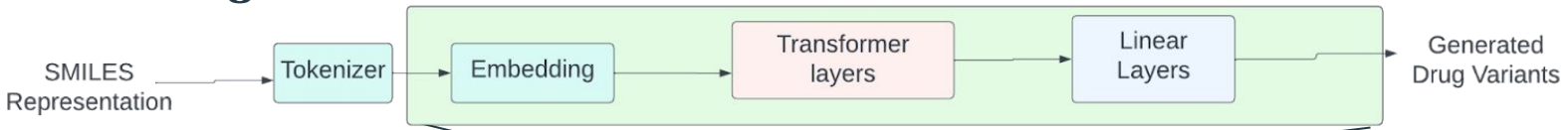
Toxicity Prediction Model



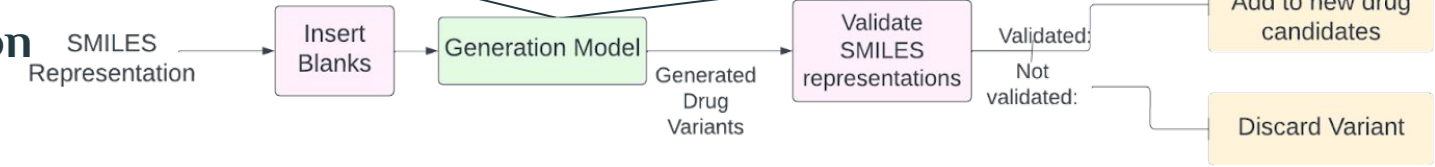
Drug Response Prediction Model



De Novo Drug Generation Model



Drug Generation via perturbation

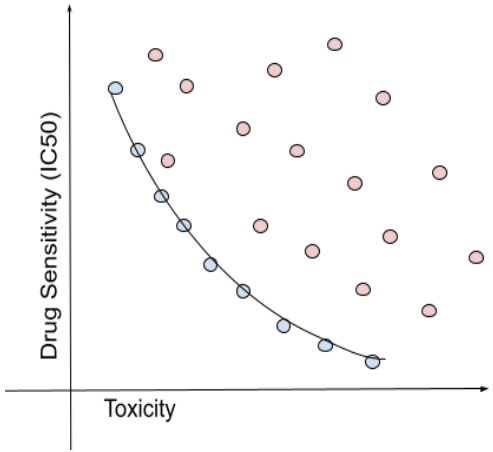


Drug Generation Explained: De Novo Drug Design

- Drug generation component is responsible for generating new drug variants that can be sent back through the pipeline, resulting in novel drug candidates.
- Our drug generation model works by adapting the architecture of the Nano GPT [14] model for SMILES string perturbation. We perform “infilling” on the SMILES strings, after inserting a few blanks at random, adapting the approach in [15].
- Example:

  C_COO_ => Model => C**X**COO**Y** (X and Y are placeholders)

- We use rdKit [16] as an initial pre-check to filter out drugs that are not valid.
- We determine Pareto points to explore the tradeoff between drug sensitivity and toxicity.
  - **Pareto Point**: (IC50, toxicity) pairs representing each drug on a cell line which have no drug with both better toxicity AND IC50 values.
- Each time new drugs are generated, they are added to existing variants to form a **pareto curve** where the best pareto points are selected.

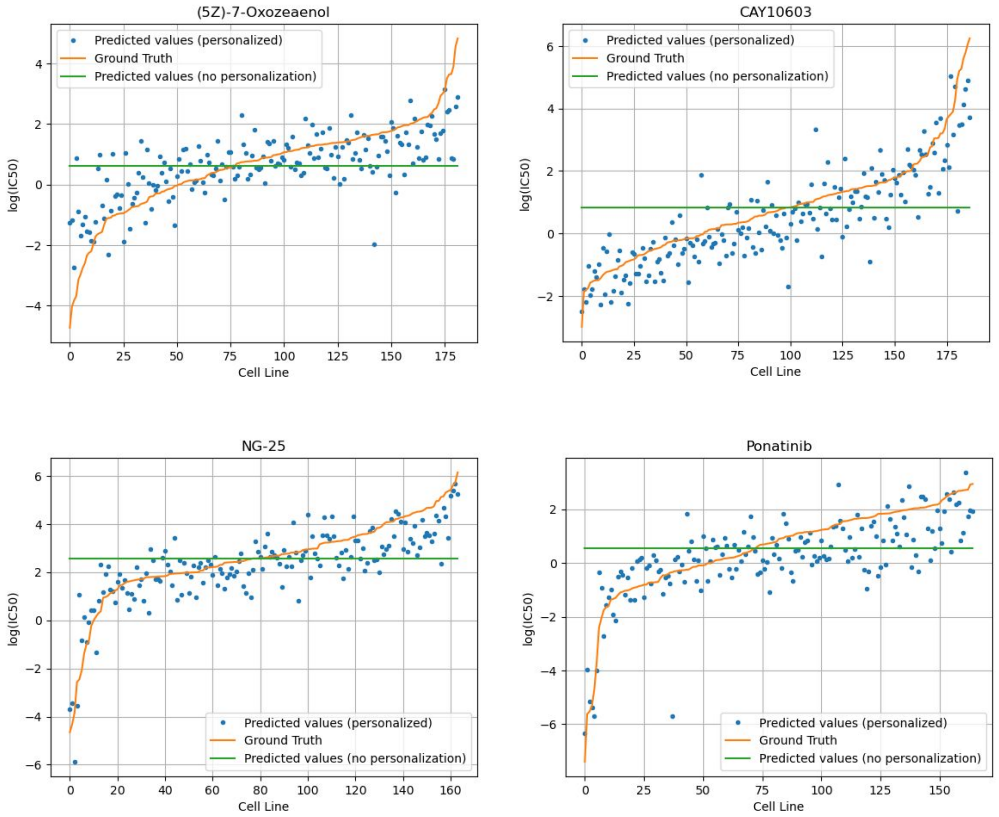


Pareto Curve

Drug Response Prediction

RESULTS

Visualization of performance for four drugs



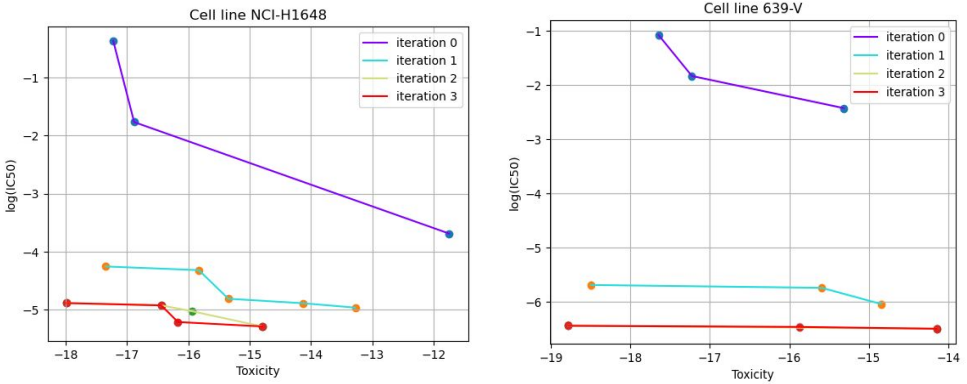
DISCUSSION

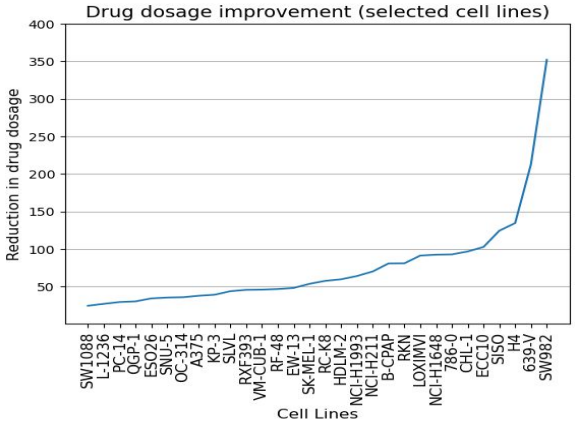
Comparison of true IC50 values and the predicted values generated by the model - shows difference between a *personalized* algorithm for drug sensitivity and a general algorithm.

- The graphs are generated with 4 randomly sampled drugs from the test data.
- The x axis are the respective cell lines and the y axis are the IC50 of each drug on all of the cell lines
- Predicted values match the shape of the ground truth IC50 curve, showing that our trained model can successfully predict drug sensitivity, with a root mean squared error (rms) of 1.1.
- **Importance of personalized drug treatment:** The trend of both the orange curve and blue scatter plot indicate the effect of drugs on different cell lines → shows the importance of not prescribing one drug for all patients (green line).

Generation Pipeline

RESULTS





DISCUSSION

The first two plots show the pareto curves for two cell lines, 639-V and NCI-H1468 over 4 iterations of the Mediseeker pipeline (consisting of drug generation, drug response prediction, and toxicity prediction).

- In both plots,the log(IC50) value is reduced by 5, which means that we have a drug candidate that works at **150 times** lesser concentration!($e^5 \approx 150$)
- Note that this achieved at the same or better level of drug cardiotoxicity

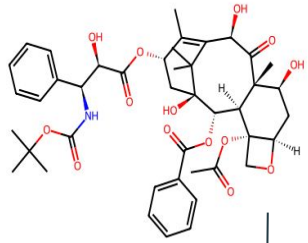
In the bottom plot, we illustrate the dramatic reduction in the drug dosage for selected cell-lines to achieve the same sensitivity. We obtain this plot by computing the IC50 improvement for the top left pareto points. The newer drugs are less toxic than the original ones and are much more effective.

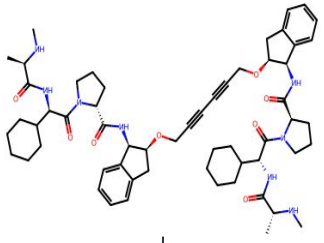
Results: Drug Generation

RESULTS

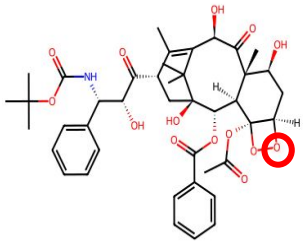
= differences in drug structure

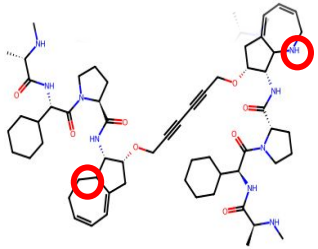
Original





Variant





DISCUSSION

The diagrams on the left were generated from SMILES representations output by the model.

- The top figures show the original drug's molecular structure for the drugs with the lowest toxicity in the GDSC dataset for the two cell-lines: NCI-H1648 and 639-V
- Both drugs go through four iterations of our drug generation pipeline, resulting in the variants shown at the bottom, which now are predicted to have much lower IC50s.
- Since our method introduces small perturbations at each step, the final drug candidates are different from the original approved treatments in just a few select positions.

Conclusions and Next Steps

Conclusions

- Designed a novel pipeline that was able to generate effective drugs on individual cell lines. These drug variants were predicted to be as effective as the original drugs, but at a 10 to 100x lower dosage. The newer drug variants are also less toxic than the original drugs.
- Designed a generative model to assist with de novo anti-cancer drug design, creating drugs with minute variances than the original.
- Predicted the toxicity of drugs based on chemical representation.
- Used the SMILES representation of various drugs and individual gene expression from cell lines to predict drug effectiveness, as measured by log(IC50). Our drug sensitivity model had a rms prediction error of +/- 1.1.

Limitations

- Our pipeline is highly dependent on the ability of models to accurately predict toxicity and IC50 for novel drugs. The toxicity prediction model show a large rms error of about 10%

Impacts of our approach:

- Can accelerate effective drug development.
- Can also lead to targeted, personalized therapy for patients based on the gene expression of their tumors.
- Can be easily extended to find new antibiotics to combat antibiotic resistance.

Future Work/Extensions

- Build highly accurate models for predicting toxicity and IC50.
- Incorporate different measurements of effectiveness besides toxicity, such as chemical absorption, distribution, metabolism, and excretion
- Build improved drug generation models that generate more structured variants.

References

[1] https://www.cancer.gov/about-cancer/treatment/research/drug-combo-resistance
[2] https://pubmed.ncbi.nlm.nih.gov/31801986/
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9963982/
[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9957434/
[5] https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2020.00733/full
[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9798519/
[7] https://www.frontiersin.org/articles/10.3389/fmed.2023.1086097/full
[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9773863/
[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3232635/
[10] https://www.cancerrxgene.org/
[11] https://sites.broadinstitute.org/ccle/
[12] https://www.ebi.ac.uk/chembl/
[13] https://tdcommons.ai/
[14] https://github.com/karpathy/nanoGPT
[15] https://aclanthology.org/2020.acl-main.225/
[16] https://www.rdkit.org/

Note: All images created by student authors