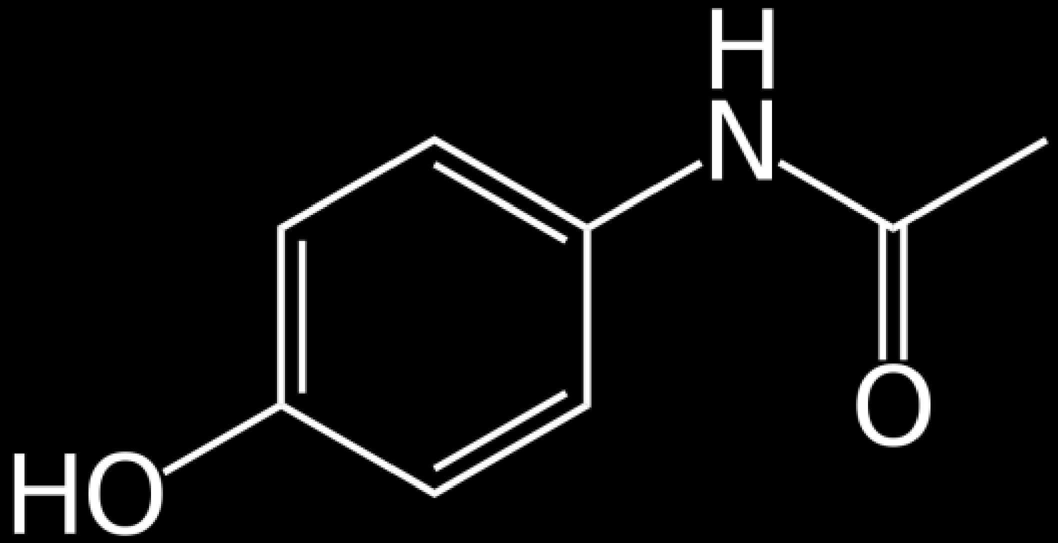


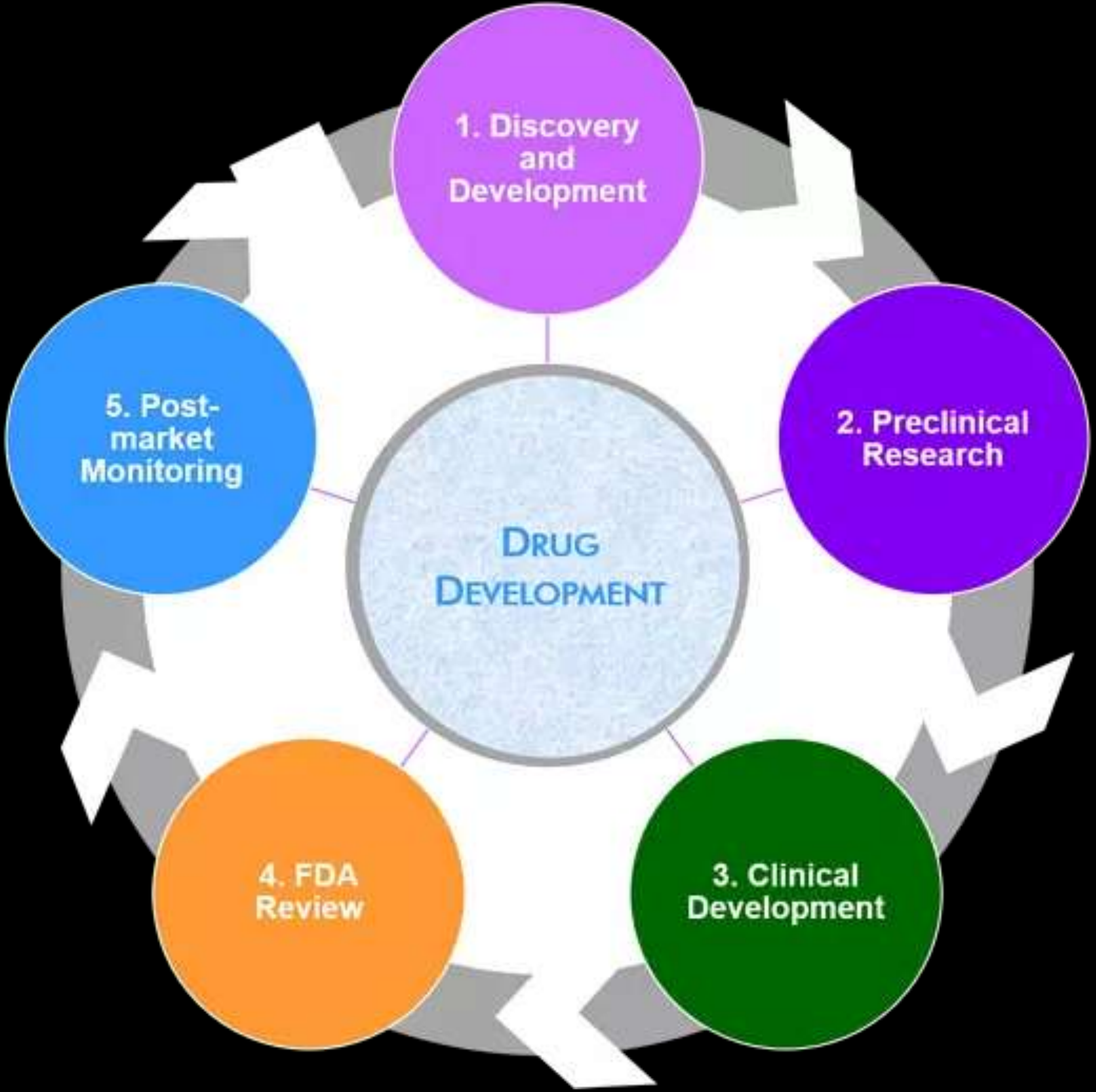
MACHINE LEARNING IN DRUG DEVELOPMENT

By Syeam Bin Abdullah

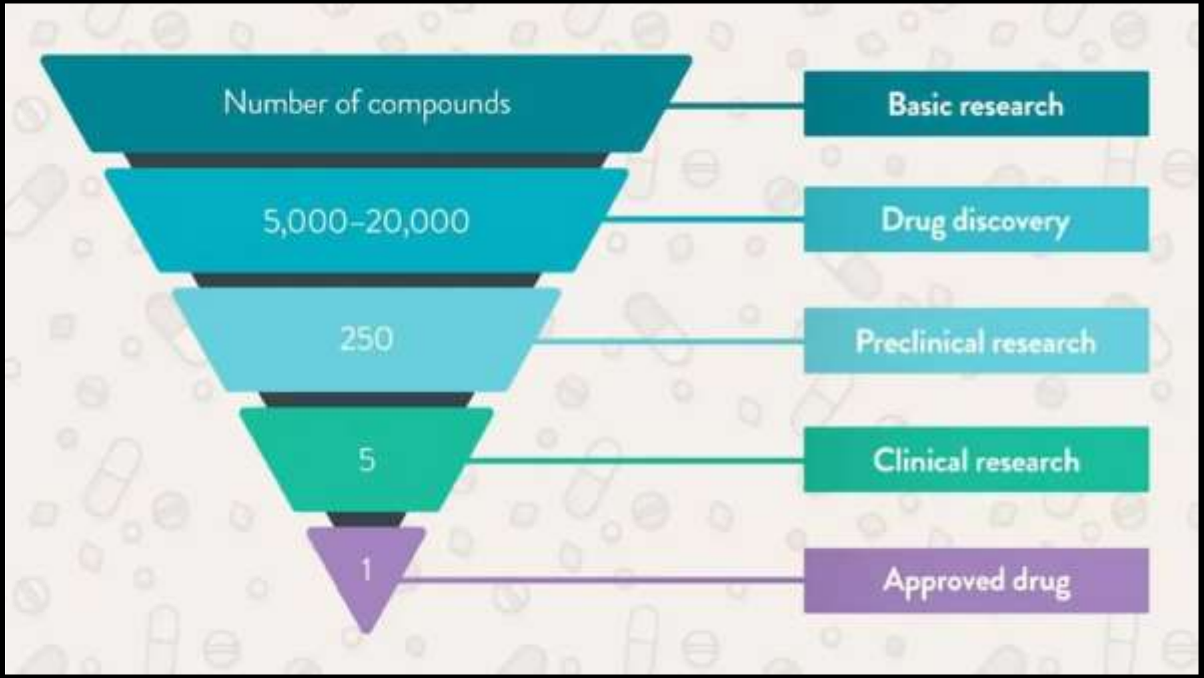
a1850041

INTRODUCTION

- Drugs are a common form of treatment for treating various illnesses
- Various in silico methods (machine learning in particular) are employed in all stages of drug development.
- Drug discovery and trials/testing are critical components of the process.



CHALLENGES



- Drug discovery is extremely tedious – chemical space is vast. Companies start with thousands, and eventually narrow down to 1 drug.
- Various trial phases before market approval – need to go through various hoops.
- Late-stage clinical trials take many years and millions of dollars to conduct

LITERATURE REVIEW – CURRENT SOLUTIONS

- Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules – **Drug Discovery**
  - *https://pubs.acs.org/doi/full/10.1021/acscentsci.7b00572*

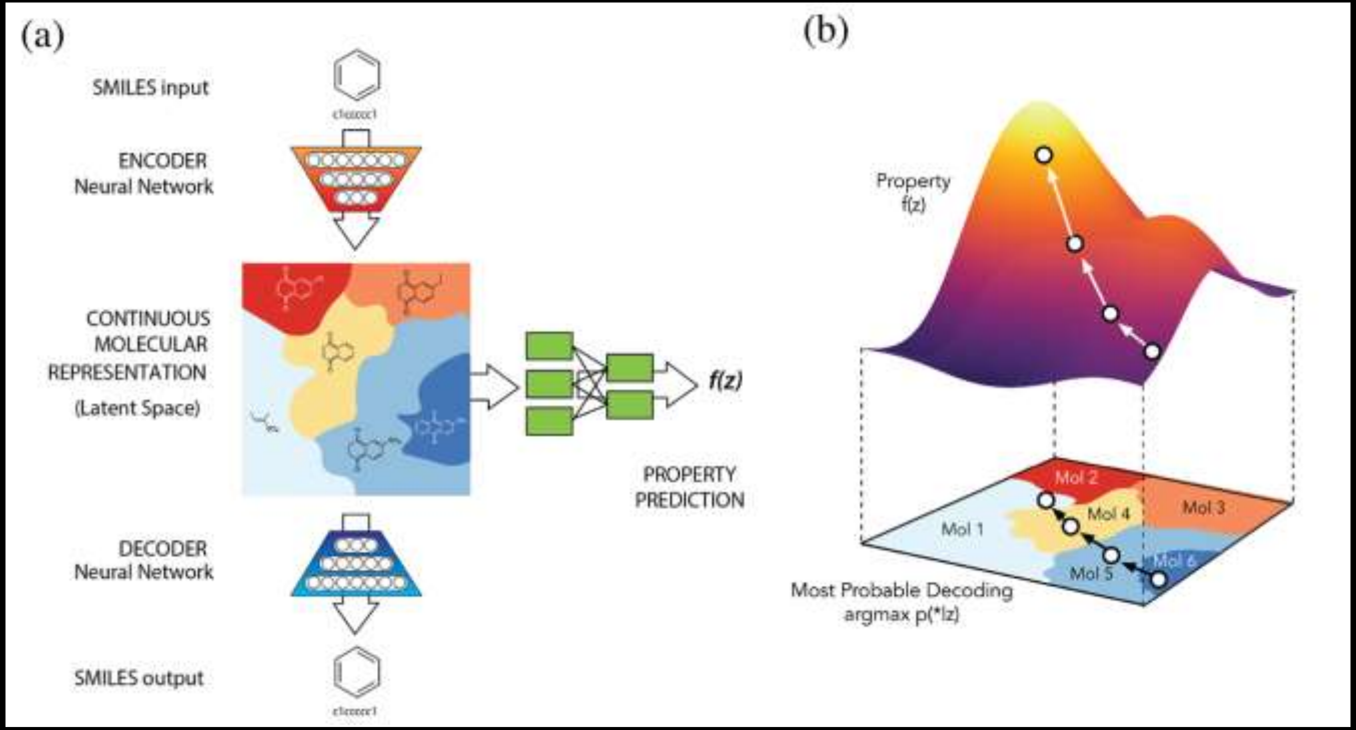
ž druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico – **Drug Discovery**

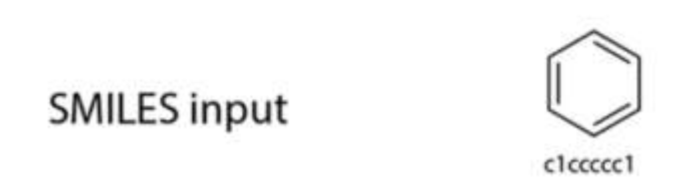
  - *https://pubs.acs.org/doi/full/10.1021/acs.molpharmaceut.7b00346#*

- Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes – **Clinical Outcome Prediction**
  - *https://www.biorxiv.org/content/10.1101/095653v2*

#1 – AUTOMATIC CHEMICAL DESIGN USING A DATA-DRIVEN CONTINUOUS REPRESENTATION OF MOLECULES

- Uses a variational autoencoder (VAE) architecture.
- Takes existing compound representations, encodes it into latent space and that encoding can be decoded with a neural net. **(a)**
- The latent space can be used as an optimisation landscape. The desired point in space is fed to decoder to output a molecular structure. **(b)**



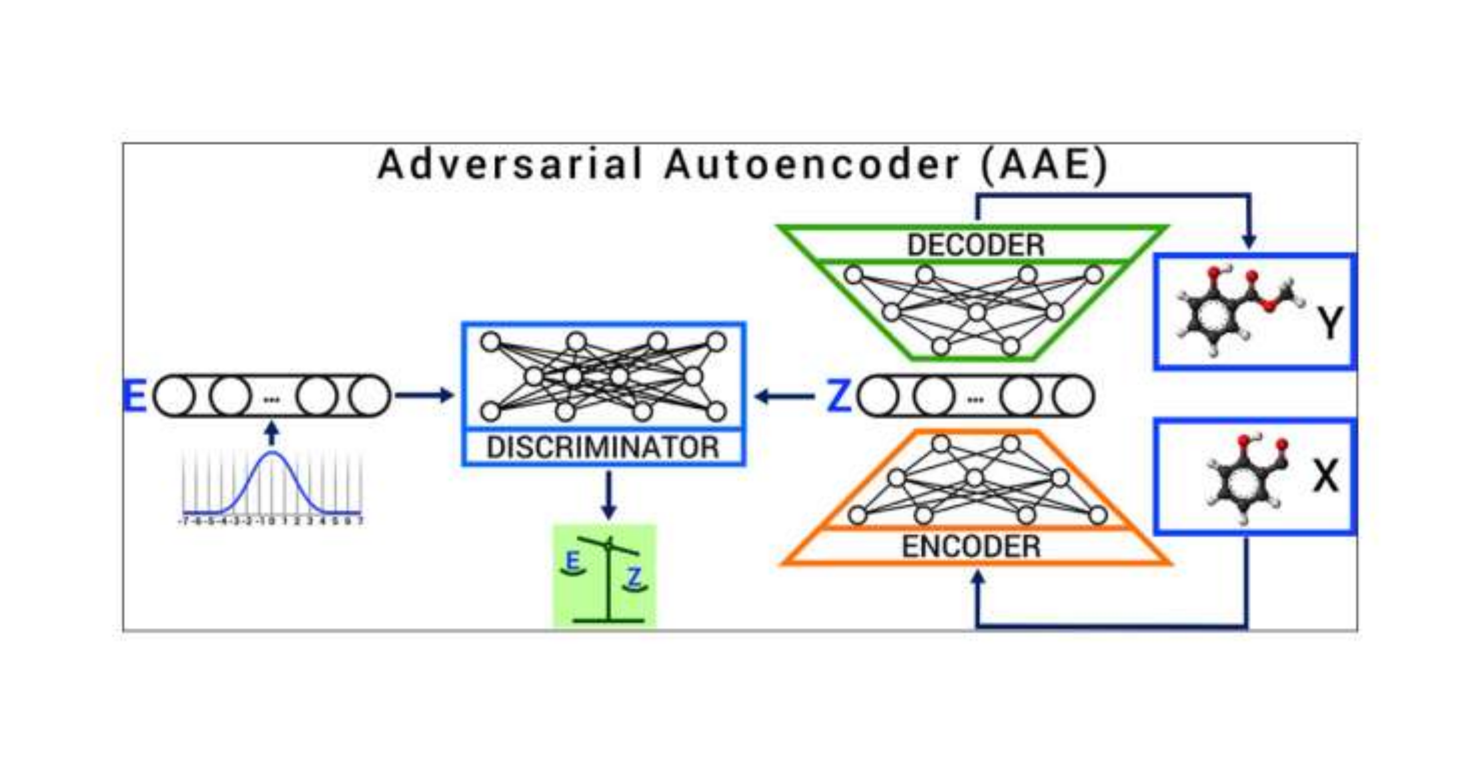


CRITIQUES

- Requires large pre-existing collection of compounds – not too much of a problem.
- Chosen input representation format (SMILES) does not capture fully the complexities of compounds – Don’t consider pharmacokinetics.

#2 – DRUGAN: AN ADVANCED GENERATIVE ADVERSARIAL AUTOENCODER MODEL FOR DE NOVO GENERATION OF NEW MOLECULES WITH DESIRED MOLECULAR PROPERTIES IN SILICO

- Very similar to #1 – uses a different loss function due to the added GAN architecture
- Essentially has another model which 'judges' the authenticity of molecules the auto-encoder generates – the discriminator
- The discriminator can be used as a proxy for the validity of novel compound arrangements

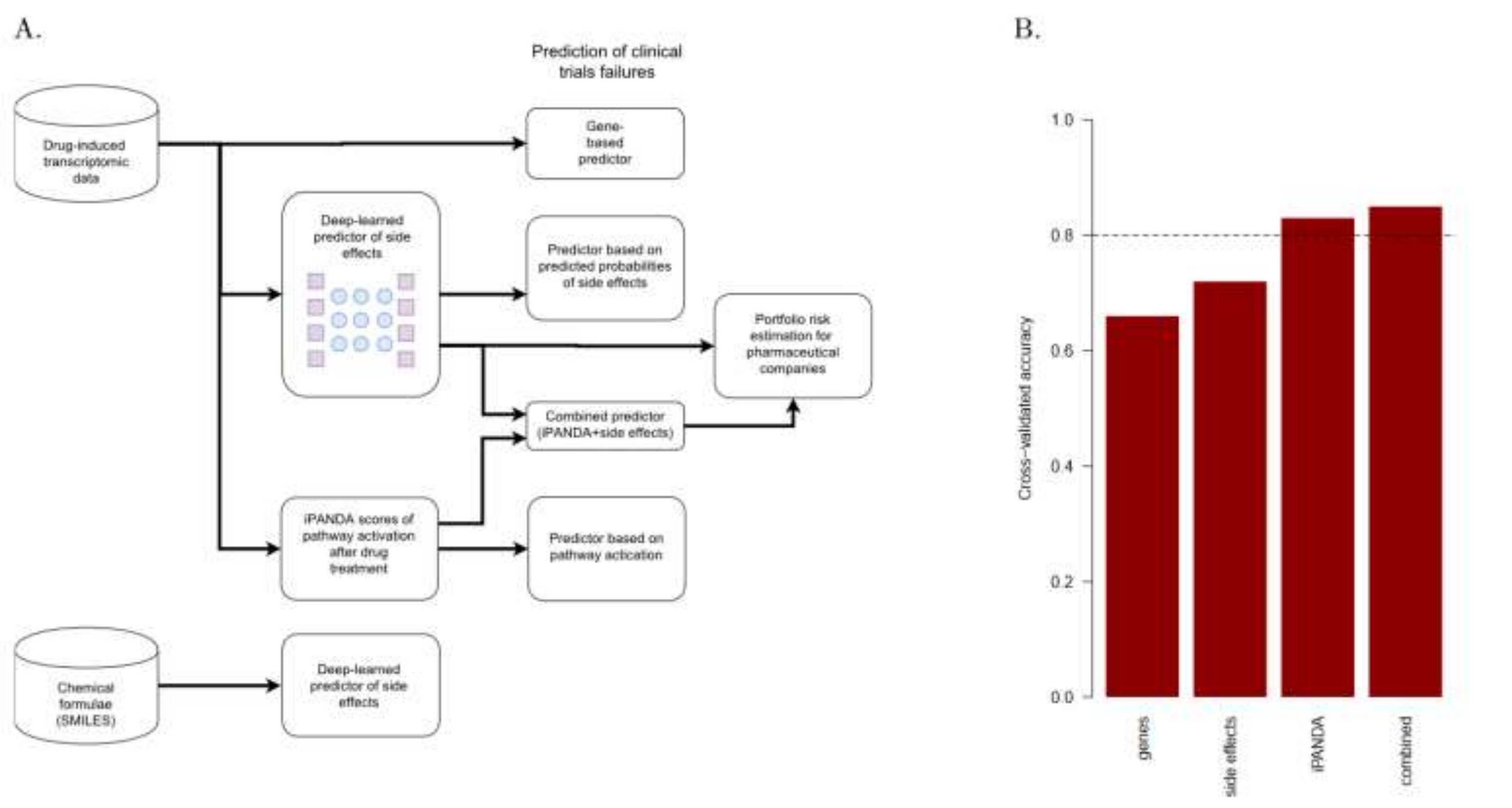


CRITIQUES

- Although a clever implementation of a GAN, it doesn't really address the limitations of #1.
- This is a great example of the ’More model vs. More data’ problem in machine learning.
- Should be possible to get a better representation of molecules instead of SMILES to learn a richer latent space instead of adding more complexity to the ML model.

#3 – INTEGRATED DEEP LEARNED TRANSCRIPTOMIC AND STRUCTURE-BASED PREDICTOR OF CLINICAL TRIALS OUTCOMES

- Used neural nets to predict side effects. Took in ‘pathway activation scores’ as input.
- Random forest for predicting outcome of clinical trial based on changes in drug-induced gene expression.


A.
Prediction of clinical trials failures
Drug-induced transcriptomic data
Gene-based predictor
Deep-learned predictor of side effects
Predictor based on predicted probabilities of side effects
Portfolio risk estimation for pharmaceutical companies
Combined predictor (iPANDA+side effects)
iPANDA scores of pathway activation after drug treatment
Predictor based on pathway actication
Chemical formulae (SMILES)
Deep-learned predictor of side effects
B.
Cross−validated accuracy
1.0
0.8
0.6
0.4
0.2
0.0
genes
side effects
iPANDA
combined

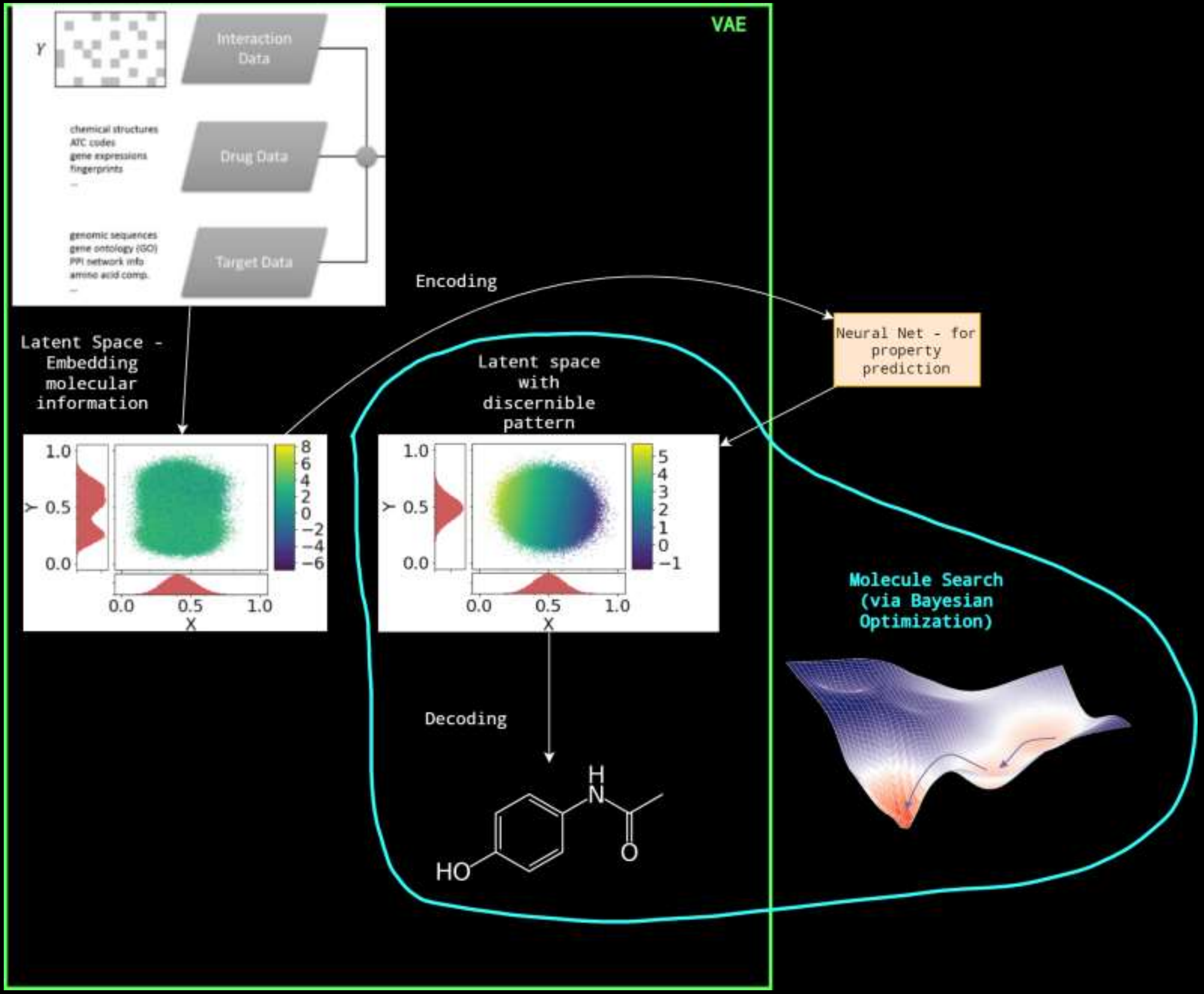
CRITIQUES

- Gene expression change data is gathered from inserting drugs into cell cultures -> This type data is not always available for compounds.
- Also, clinical trial outcome prediction was poor when using gene expressions. **(see (B) - genes)**. Genes alone do not fully capture the complexity of the cell signalling.
- Each side effect requires a totally separate neural network -> this is inefficient.

BETTER DRUG REPRESENTATIONS ARE REQUIRED

PROPOSED SOLUTION – BETTER REPRESENTATION LEARNING

- Input: Randomized canonical SMILES (drug representation) + Interaction data (cell signalling pathway activations) + target data (enzymes, proteins, etc.) – all encoded in matrices.
- Uses VAE for encoding into and out of latent space.
- VAE + NN for joint property prediction (i.e., logP, QED, how well a molecule targets a certain enzyme).
- Run optimization algo. on latent space, run the selected point in the space through the decoder to get a candidate molecule.



VAE

Encoding

Neural Net - for property prediction

Latent Space - Embedding molecular information

Latent space with discernible pattern

Decoding

Molecule Search (via Bayesian Optimization)

BENEFITS AND LIMITATIONS OF SOLUTION

- Richer chemical latent space for drug discovery – considers more complicated properties of compounds, not just topological like SMILES.
- Synthetic Data Generation opportunity - A version of the latent space can be updated for the public to use, as it gets better and better it can be utilized to generate more reliable synthetic data, helping alleviate the lack of clinical data available for more novel treatment targets.
- More data-points required for input – a little more resource-intensive than state-of-the-art

REFERENCES

- Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules
  - *https://pubs.acs.org/doi/full/10.1021/acscentsci.7b00572*
- druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico
  - *https://pubs.acs.org/doi/full/10.1021/acs.molpharmaceut.7b00346#*
- Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes
  - *https://www.biorxiv.org/content/10.1101/095653v2*
- Applications of machine learning in drug discovery and development
  - *https://www.nature.com/articles/s41573-019-0024-5*