**Molecular Generation Using Variational Auto-Encoders**

Daniel Flockhart

Artificial Intelligence and Computer Science BSc

Contents

[Abstract 2](#_Toc143964662)

[Project Overview - Introduction 3](#_Toc143964663)

[Related Work 4](#_Toc143964664)

[Datasets 5](#_Toc143964665)

[Zinc 250k 5](#_Toc143964666)

[Conditions 5](#_Toc143964667)

[Pre-Processing 6](#_Toc143964668)

[Vectors 6](#_Toc143964669)

[Conditions 6](#_Toc143964670)

[Target Images 6](#_Toc143964671)

[Methods 7](#_Toc143964672)

[ChemBERTa 7](#_Toc143964673)

[Variational Auto Encoder 7](#_Toc143964674)

[QSAR 7](#_Toc143964675)

[Training 8](#_Toc143964676)

[Computation/Hardware 8](#_Toc143964677)

[Hyper Parameters 8](#_Toc143964678)

[Analysis 8](#_Toc143964679)

[Genetic Information 8](#_Toc143964680)

[Generation 9](#_Toc143964681)

[Input Synthesis 10](#_Toc143964682)

[Inputs 10](#_Toc143964683)

[Outputs 10](#_Toc143964684)

[Analysis 11](#_Toc143964685)

[Simulation 12](#_Toc143964686)

[Ranking 13](#_Toc143964687)

[Personalised Molecule Generation 14](#_Toc143964688)

[Genetic Tailoring 14](#_Toc143964689)

[Synthesis Routes 15](#_Toc143964690)

[Experimentation 16](#_Toc143964691)

[Results 17](#_Toc143964692)

[Discussion 18](#_Toc143964693)

[Future Work 19](#_Toc143964694)

[Conclusion 20](#_Toc143964695)

[Acknowledgements 21](#_Toc143964696)

[References 22](#_Toc143964697)

[Academic Papers 22](#_Toc143964698)

[Videos 22](#_Toc143964699)

[Appendices 23](#_Toc143964700)

# Abstract

# Project Overview - Introduction

A screenshot of a computer screen

Description automatically generated

# Related Work

# Datasets

## Zinc 250k

## Conditions

# Pre-Processing

## Vectors

## Conditions

## Target Images

# Methods

## ChemBERTa

## Variational Auto Encoder

## QSAR

# Training

## Computation/Hardware

## Hyper Parameters

## Analysis

# Genetic Information

Ranking with genetics

A colorful lines with text

Description automatically generated with medium confidence

Compress Genetic Information

A diagram of a different color

Description automatically generated with medium confidence

# Generation

# Input Synthesis

## Inputs

The inputs to this section is the illness and any other required information

## Outputs

The goal of this component is to generate a chosen starting molecule for the VAE and generate a set of conditions.

# Analysis

# Simulation

# Ranking

Heuristic

* Performance estimation based on simulation score + genetic ranking

# Personalised Molecule Generation

## Genetic Tailoring

I propose a new technique for personalized healthcare.

It consists of two separate (variational) Auto-Encoders and one decoder

Input Synthesis

* Generates a starting molecule and a condition

Encoders

* One compresses a target users genetic information into a latent space – We will not use the decoder of this
* One compresses a starting molecules representation into a latent space

Decoder

* Takes the Compressed genetic Information
* Take the starting molecule latent space
* Takes the user specified condition

# Synthesis Routes

# Experimentation

# Results

# Discussion

# Future Work

# Conclusion

# Acknowledgements

# References

## Academic Papers

<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0286-7> -

Journal of Informatics - Molecule Generative Model based on conditional variational autoencoder for de novo molecular design.

<https://arxiv.org/pdf/2209.01712.pdf>

Arxiv - ChemBERTa-2: Towards Chemical Foundation Models

<https://www.nature.com/articles/s41597-022-01142-7>

Nature - Organic Materials Repurposing, a dataset for theoretical predictions of new applications for existing compounds.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570547/>

National Library of Medicine - Disease-Drug Database for Pharmacogenomic-Based Prescribing

<https://www.kaggle.com/datasets/jithinanievarghese/drugs-related-to-common-treatments>

Kaggle – Drugs Related to Medical Conditions

<https://www.nature.com/articles/s42256-022-00557-6>

Nature - Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework

## Videos

<https://www.youtube.com/watch?v=4mygq7Brtu8>

Microsoft YouTube Channel - Research Talk: AI for drug discovery

# Appendices