NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

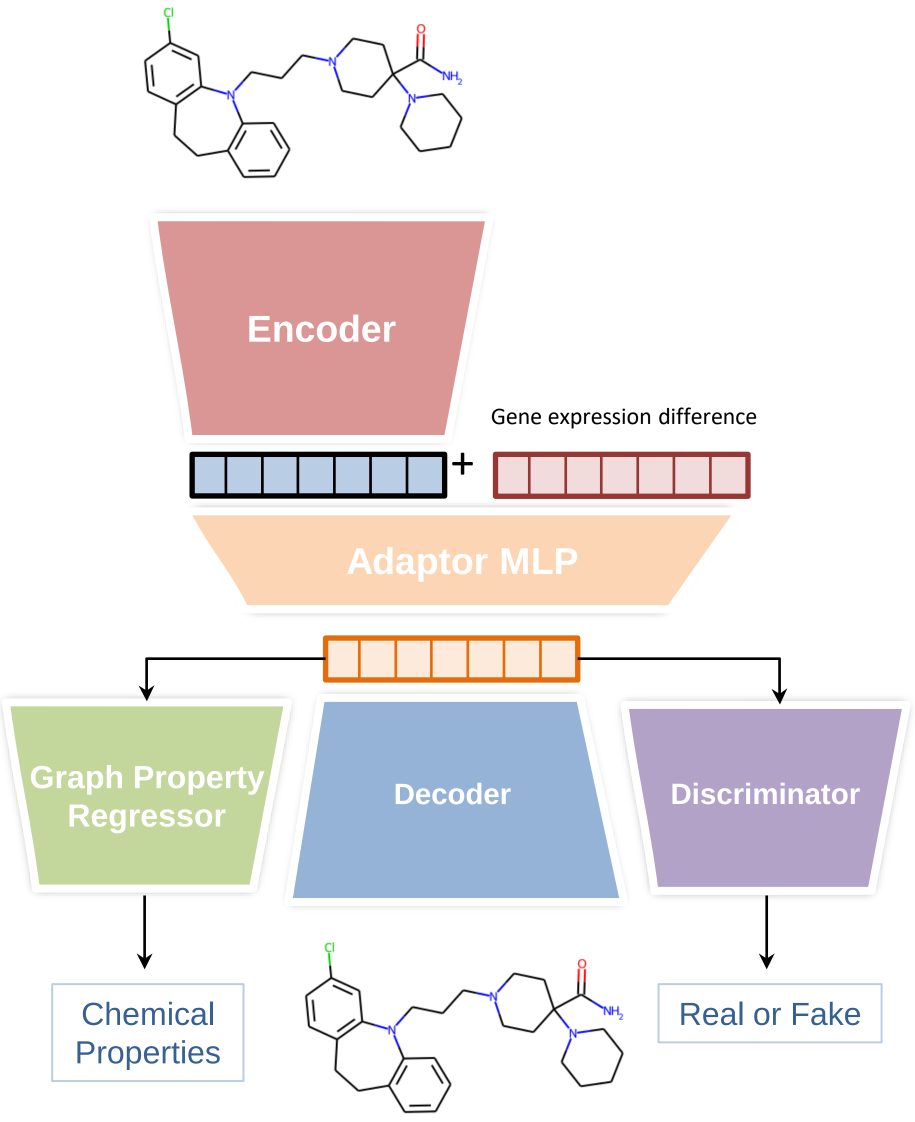
Drug Discovery with Adversarial Autoencoder Networks Conditioned on Gene Expression Profiles

Student: Ong Hiok Hian

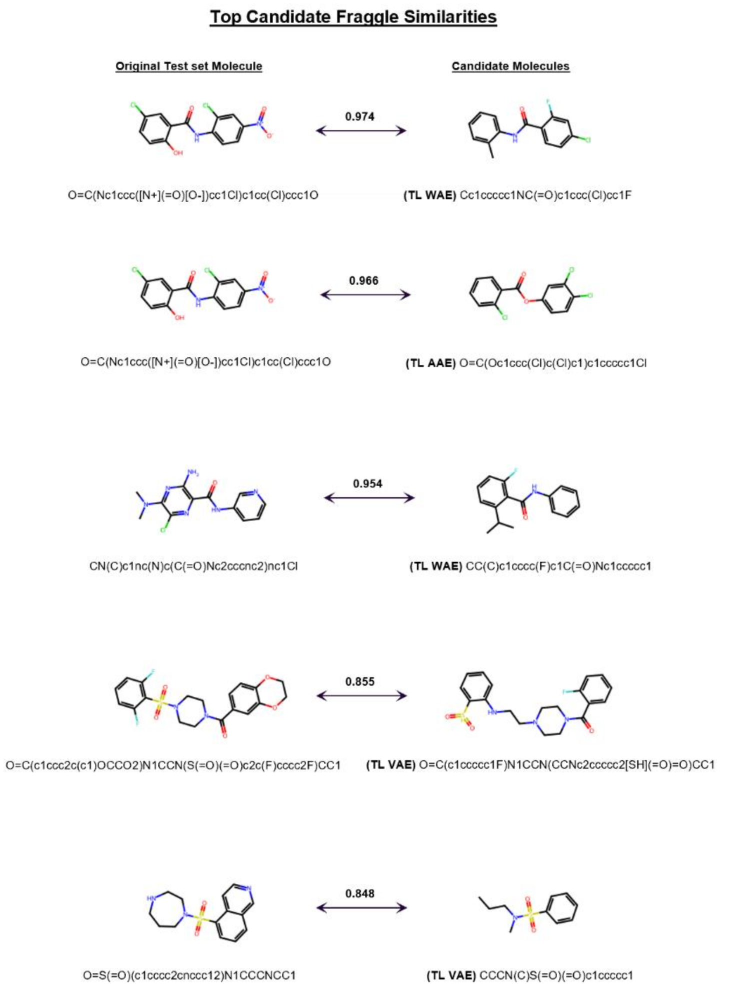
Supervisor: Professor Jagath Chandana Rajapakse
Mentor: Wang Conghao

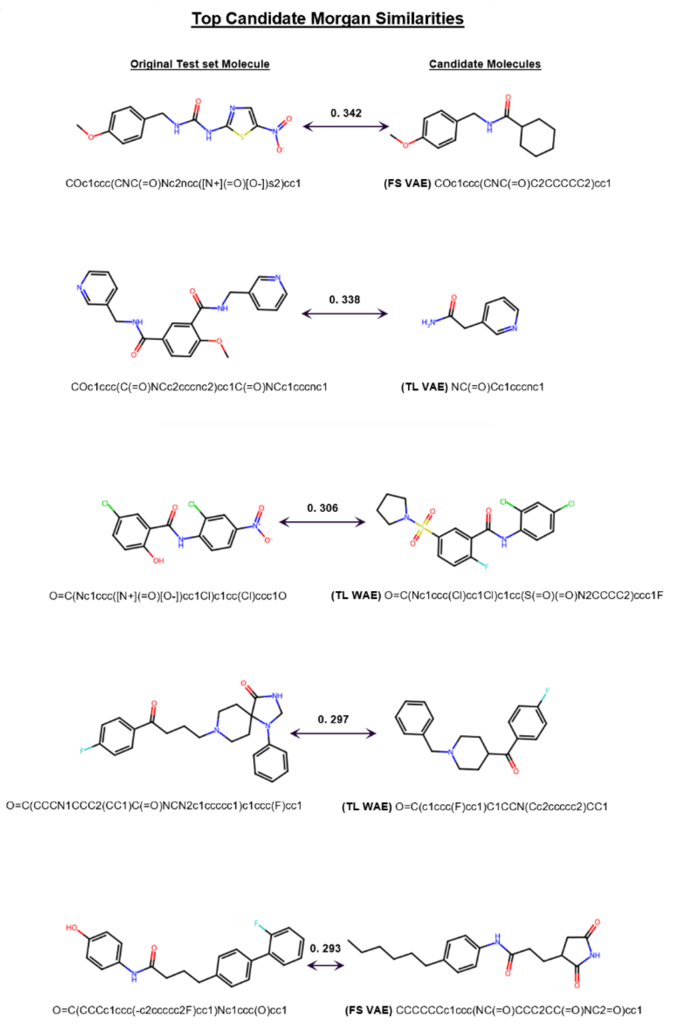
Project Objectives:

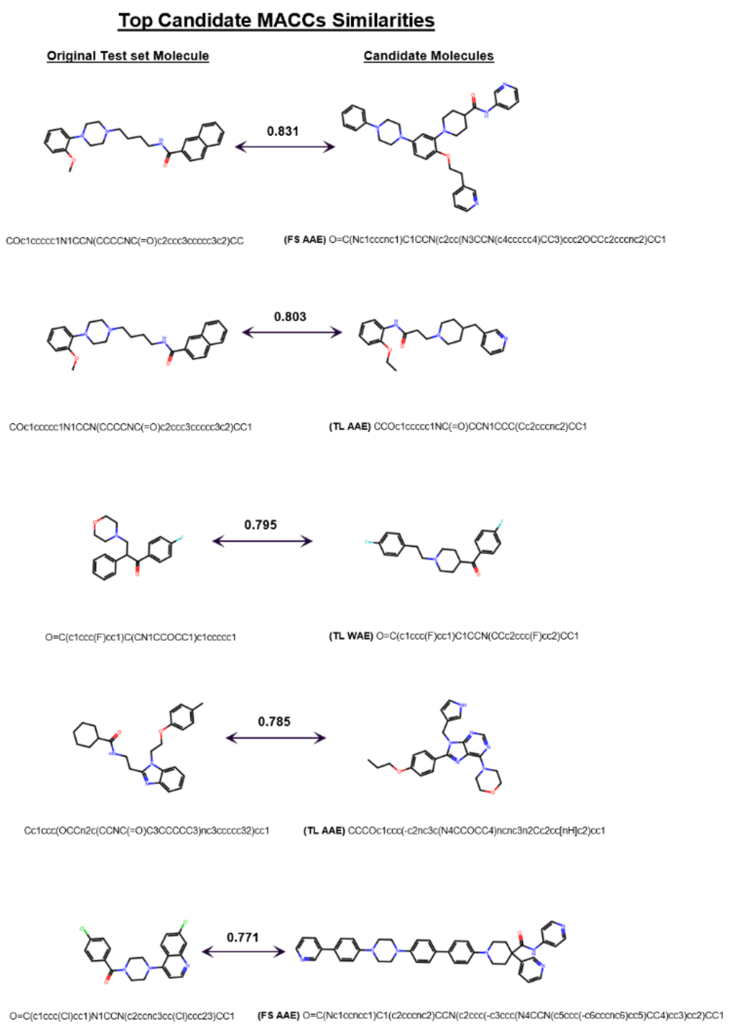
Conventional drug design and discovery is a time-consuming and computationally expensive task considering the comprehensive structure of the target protein and the necessity of generating chemically valid molecules with optimal properties. Recently, the advancement in computational power and machine learning approaches has accelerated this task by providing the more efficient algorithms for handling the available data. However, discovering the de novo molecular structure remains to be a challenge. Advancements in Conditional Graph Autoencoders show promise in empowering drug discovery by modelling the problem as the generation of molecular graphs. The aim of this project is to develop an Adversarial Graph Autoencoder model that designs the drug compound with the guidance of gene expression data, which has been demonstrated to have a strong impact on cancer progression in numerous literatures. The model is developed in two stages: (1) Large scale pretraining using Molecular Graphs as input; (2) Finetuning by conditioning on gene expression data to formulate desirable drugs using an adaptor module. In the end, the candidate molecules are evaluated against a suite of evaluation benchmarks from the Guacamol dataset as well as test set rediscovery, internal diversity, drug-likeness, ease of synthesisability.



Results:


Top Candidate Fraggle Similarities
Original Test set Molecule
Candidate Molecules
0.974
O=C(Nc1ccc([N+](=O)[O-])cc1Cl)c1cc(Cl)ccc1O
(TL WAE) Cc1ccccc1NC(=O)c1ccc(Cl)cc1F
0.966
O=C(Nc1ccc([N+](=O)[O-])cc1Cl)c1cc(Cl)ccc1O
(TL AAE) O=C(Oc1ccc(Cl)c(Cl)c1)c1ccccc1Cl
0.954
CN(C)c1nc(N)c(C(=O)Nc2cccnc2)nc1Cl
(TL WAE) CC(C)c1cccc(F)c1C(=O)Nc1ccccc1
0.855
O=C(c1ccc2c(c1)OCCO2)N1CCN(S(=O)(=O)c2c(F)cccc2F)CC1
(TL VAE) O=C(c1ccccc1F)N1CCN(CCNc2ccccc2[SH](=O)=O)CC1
0.848
O=S(=O)(c1cccc2cnccc12)N1CCCNCC1
(TL VAE) CCCN(C)S(=O)(=O)c1ccccc1


Top Candidate Morgan Similarities
Original Test set Molecule
Candidate Molecules
0. 342
COc1ccc(CNC(=O)Nc2ncc([N+](=O)[O-])s2)cc1
(FS VAE) COc1ccc(CNC(=O)C2CCCCC2)cc1
0. 338
COc1ccc(C(=O)NCc2cccnc2)cc1C(=O)NCc1cccnc1
(TL VAE) NC(=O)Cc1cccnc1
0. 306
O=C(Nc1ccc([N+](=O)[O-])cc1Cl)c1cc(Cl)ccc1O
(TL WAE) O=C(Nc1ccc(Cl)cc1Cl)c1cc(S(=O)(=O)N2CCCC2)ccc1F
0. 297
O=C(CCCN1CCC2(CC1)C(=O)NCN2c1ccccc1)c1ccc(F)cc1
(TL WAE) O=C(c1ccc(F)cc1)C1CCN(Cc2ccccc2)CC1
0. 293
O=C(CCCc1ccc(-c2ccccc2F)cc1)Nc1ccc(O)cc1
(FS VAE) CCCCCCc1ccc(NC(=O)CCC2CC(=O)NC2=O)cc1

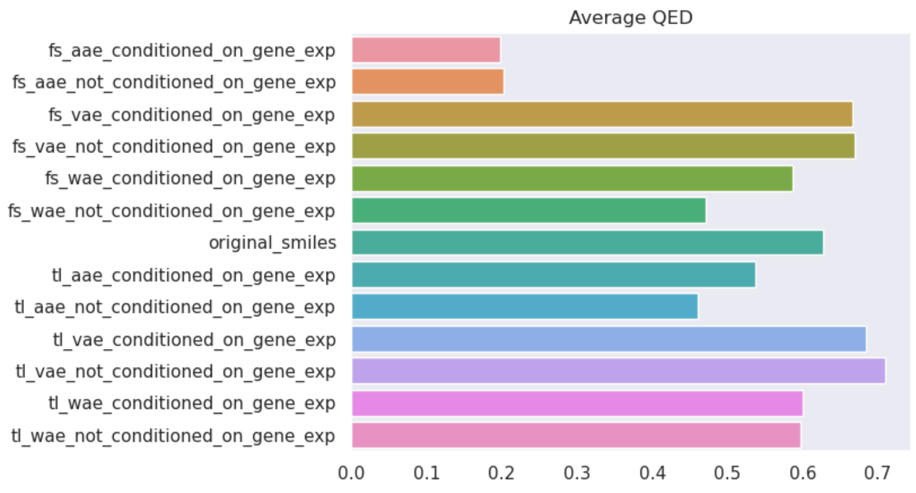

Top Candidate MACCs Similarities
Original Test set Molecule
Candidate Molecules
0.831
COc1ccccc1N1CCN(CCCCNC(=O)c2ccc3ccccc3c2)CC
(FS AAE) O=C(Nc1cccnc1)C1CCN(c2cc(N3CCN(c4ccccc4)CC3)ccc2OCCc2cccnc2)CC1
0.803
COc1ccccc1N1CCN(CCCCNC(=O)c2ccc3ccccc3c2)CC1
(TL AAE) CCOc1ccccc1NC(=O)CCN1CCC(Cc2cccnc2)CC1
0.795
O=C(c1ccc(F)cc1)C(CN1CCOCC1)c1ccccc1
(TL WAE) O=C(c1ccc(F)cc1)C1CCN(CCc2ccc(F)cc2)CC1
0.785
Cc1ccc(OCCn2c(CCNC(=O)C3CCCCC3)nc3ccccc32)cc1
(TL AAE) CCCOc1ccc(-c2nc3c(N4CCOCC4)ncnc3n2Cc2cc[nH]c2)cc1
0.771
O=C(c1ccc(Cl)cc1)N1CCN(c2ccnc3cc(Cl)ccc23)CC1
(FS AAE) O=C(Nc1ccncc1)C1(c2cccnc2)CCN(c2ccc(-c3ccc(N4CCN(c5ccc(-c6cccnc6)cc5)CC4)cc3)cc2)CC1

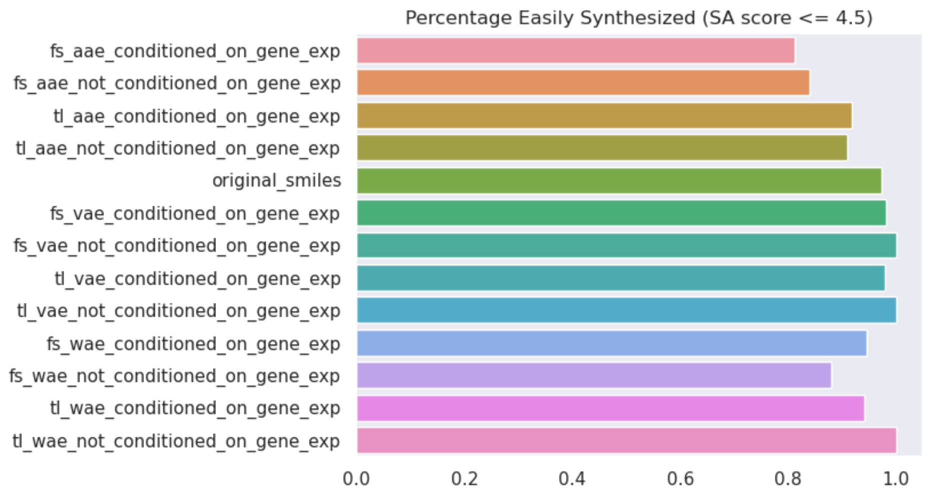
From our 2 stage evaluation we gather that within our autoregressive autoencoders:

Our Adversarial autoencoder (AAE) and Wasserstein autoencoder (WAE) outperform the Variational Autoencoder (VAE) on the Guacamol benchmark (bottom left table) and exhibit greater ability to extrapolate beyond small constrained datasets like L1000 (see FS WAE; bottom right table) and model the physicochemical and biological properties of chemical datasets without large scale pretraining.

Additionally, while VAE seems to favour more simplistic molecular substructures and AAE trades off complexity to better model the chemical properties of the dataset, WAE bridges this gap, generating molecules that are less complex and yet more drug like and easily synthesized than the candidate molecules from AAE.

Finally all 6 models, both trained from scratch and those that underwent large scale pretraining outperformed baseline similarity search, implying that these methods are better than simple similarity search for the closest molecule in terms of substructure similarity.





	Validity	Uniqueness	Novelty	KL Divergence	Frechet Chemnet Distance	Average Guacamol Score	Percentage connected	Percentage easily synthesized (SA score <= 4.5)
Best VAE	1.0	0.955	0.968	0.471	0.0874	0.696	0.85	0.97
Best AAE	1.0	0.999	0.997	0.824	0.331	0.830	0.94	0.77
Best WAE	1.0	0.994	0.993	0.917	0.354	**0.851**	0.95	0.76

*Model performance from pretraining on Guacamol dataset (without conditioning on gene expression data)

	Validity	Uniqueness	Novelty	KL Divergence	Frechet Chemnet Distance	Average Guacamol Score	Percentage connected	Internal diversity
FS AAE	1.0	0.670	0.999	0.342	0.000744	0.602	1.00	0.791
TL AAE	1.0	0.991	0.997	0.851	0.140	**0.796**	0.99	0.878
FS VAE	1.0	0.2939	0.558	0.571	0.00422	0.485	0.87	0.854
TL VAE	1.0	0.7613	0.945	0.238	0.02101	0.593	1.00	0.892
FS WAE	1.0	0.9609	0.998	0.838	0.122	**0.784**	0.88	0.871
TL WAE	1.0	0.8970	0.991	0.642	0.0509	0.716	0.96	0.880

*Model performance on Guacamol benchmark after finetuning on L1000 dataset (with conditioning on gene expression data); FS = models trained from scratch on only limited L1000 dataset; TL = transfer learning from pretraining on Guacamol then finetuning on L1000 dataset