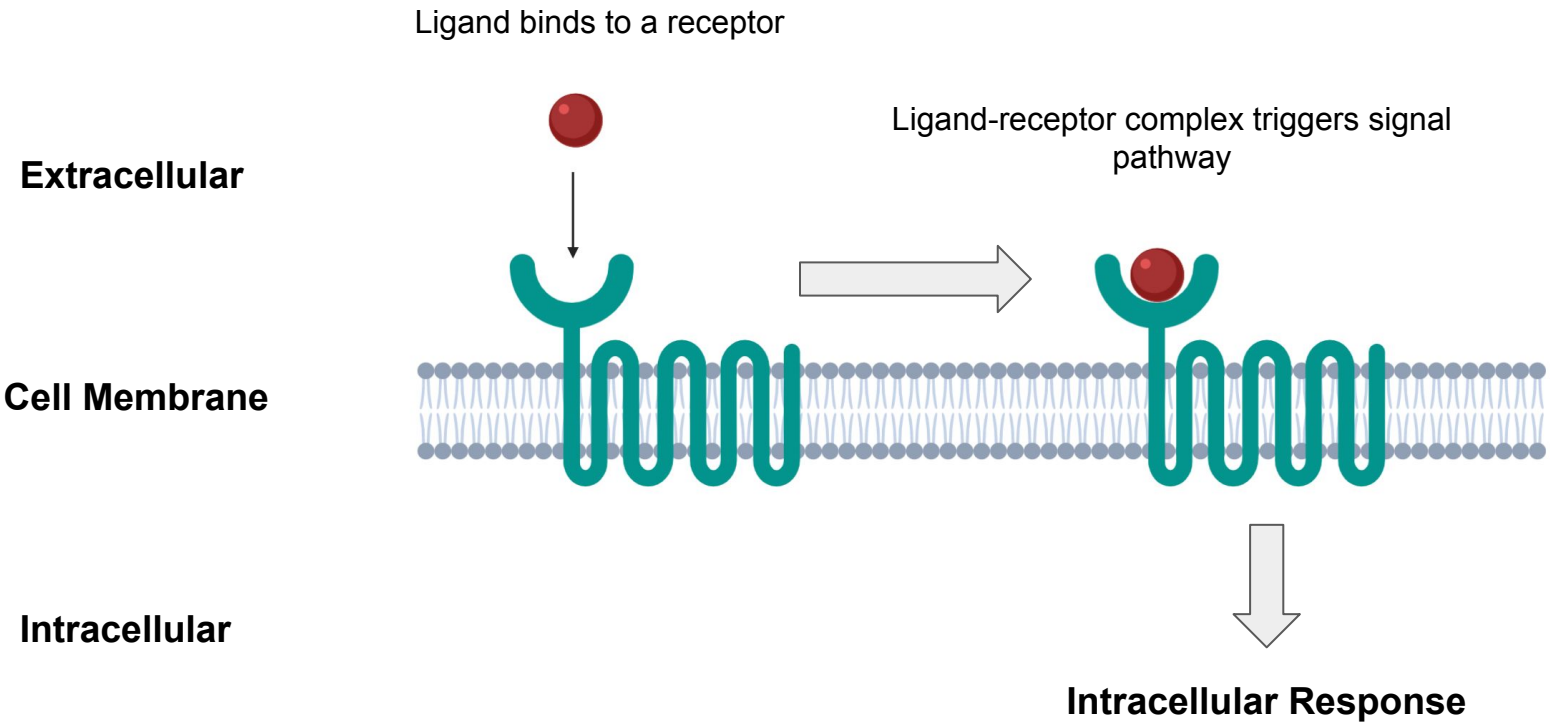
Predicting receptor activity from structural features of chemical ligands

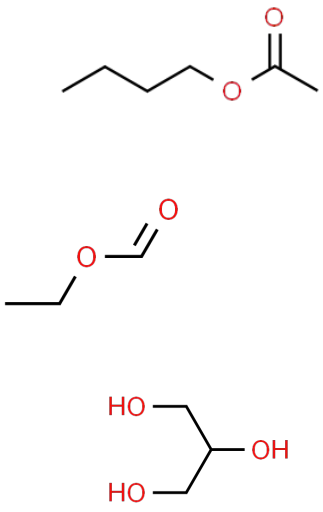
Jonathan Yin
Mentor: Dr. Hattie Chung, Regev Group, Broad Institute

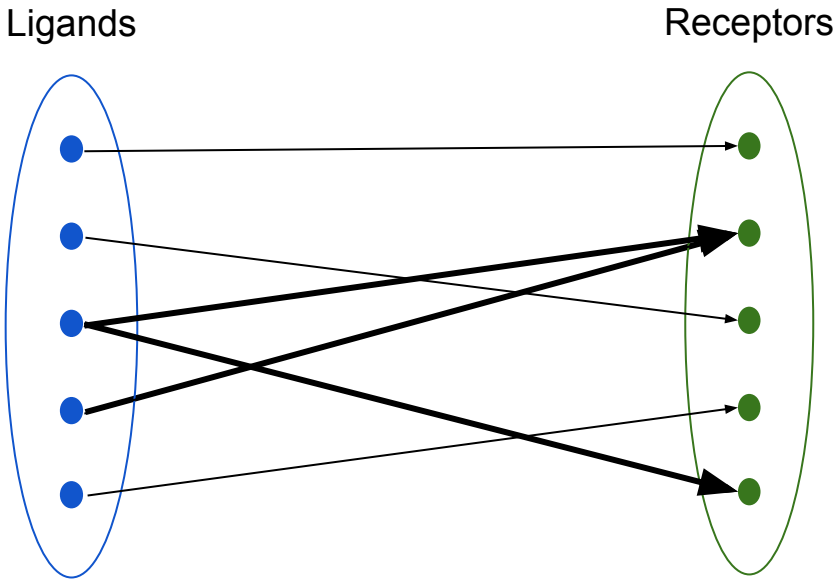
MIT PRIMES Conference, June 7th, 2020

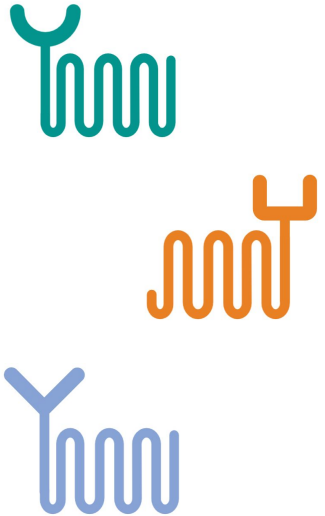
Receptors process signals from the environment



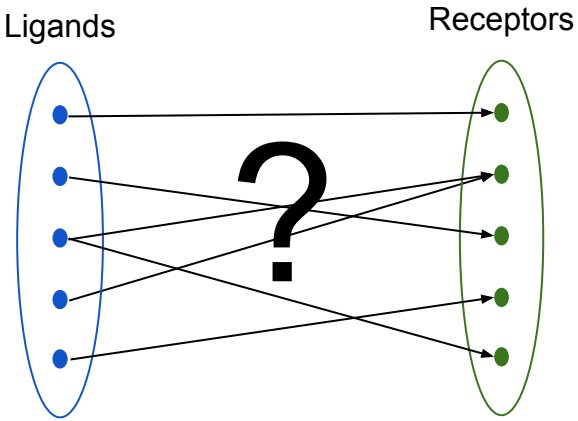
Understanding receptor-ligand interactions







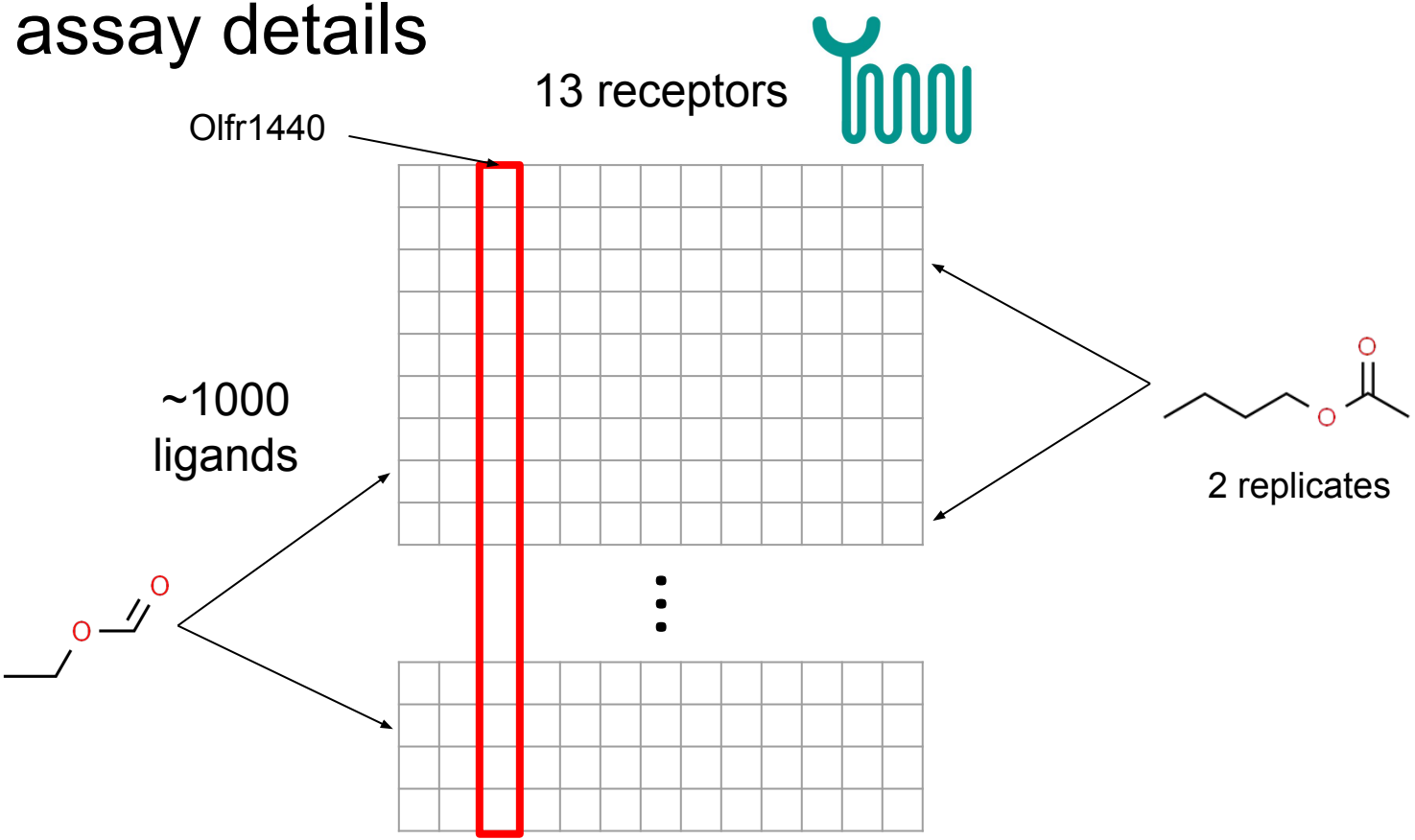
Motivation



Predicting receptor-ligand interactions

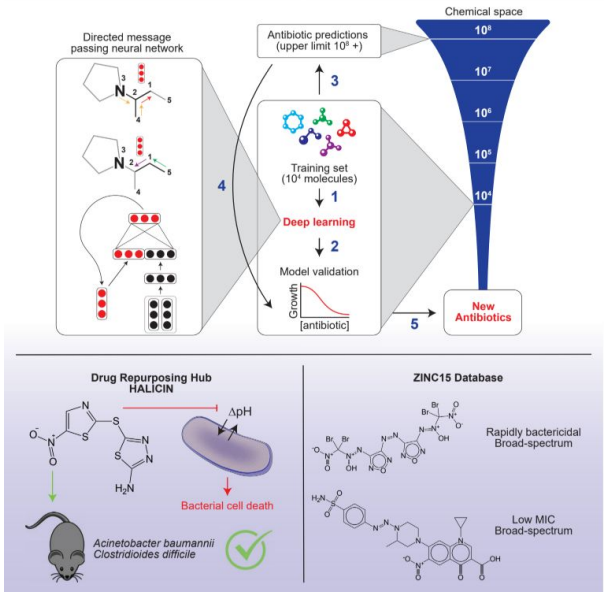
Ability to control intracellular behaviors

Pilot assay details



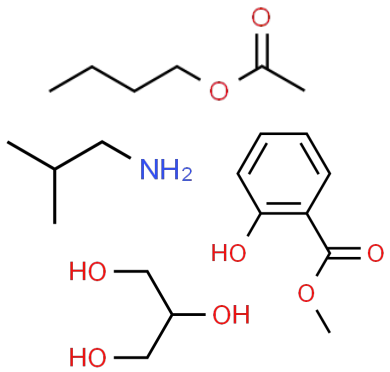
Data from Motohiko Kadoki, Ramnik Xavier lab (Broad)

Feasibility of deep learning on small biological datasets

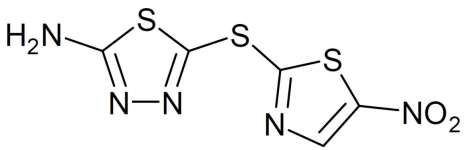


A Deep Learning Approach to Antibiotic Discovery

Stokes et al. 2020



Training dataset of ~2300 molecules



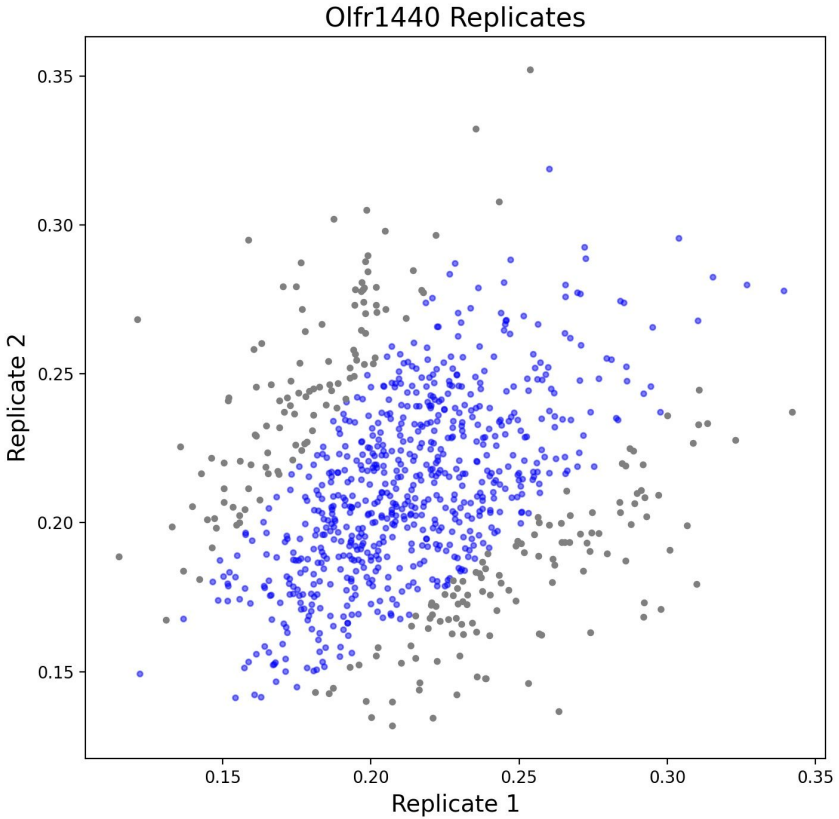
Halicin

Normalization and quality control

$$\textit{Normalized counts} = \frac{\text{Counts}_{\textit{Olfr i, well j}}}{\sum \text{Counts}_{\text{well j}}}$$

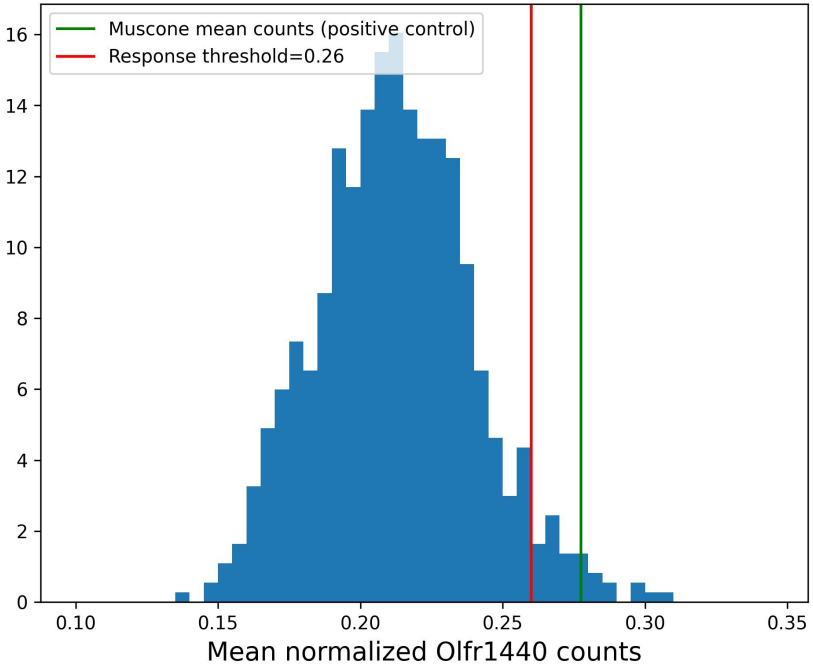
$$\textit{Quality score (QS)} = \frac{\mu_{\textit{replicates}}}{\sigma_{\textit{replicates}}}$$

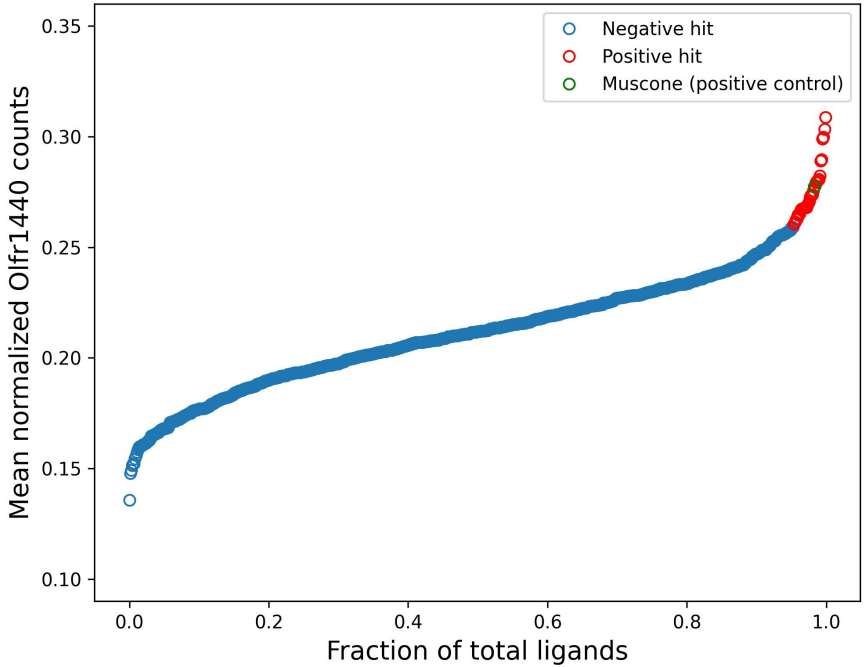
$$\textit{Quality threshold} = \textit{median}\ (QS_{\textit{control chemicals}})$$



Categorizing responses as binary

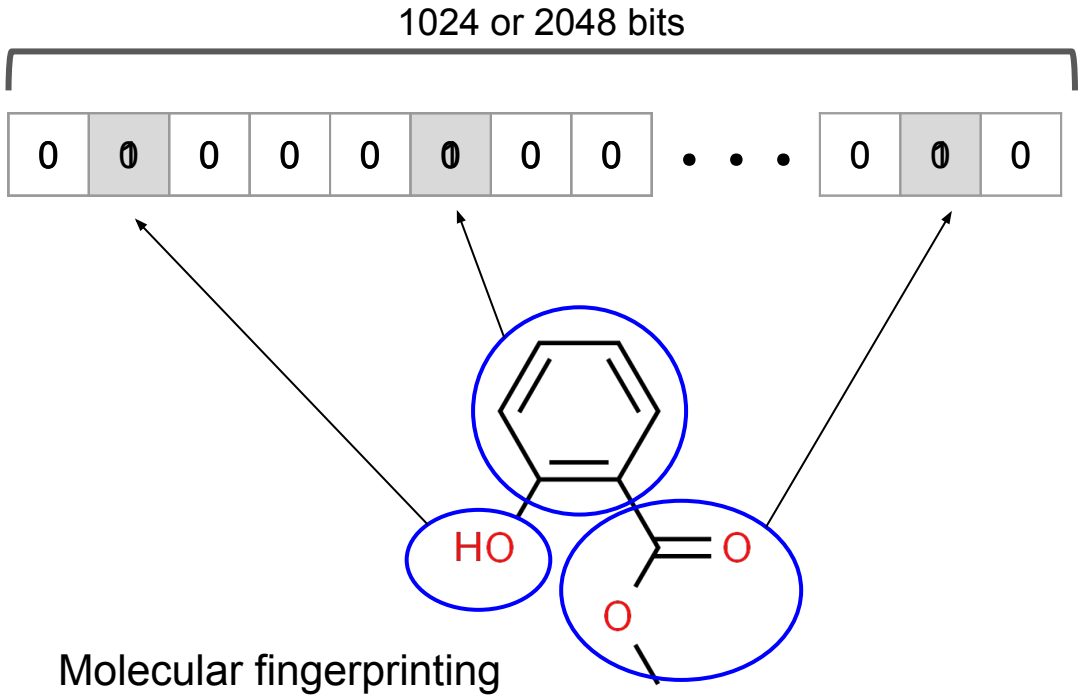
Bacterial Olfr1440 receptor activity (Stoka et al. 2020)



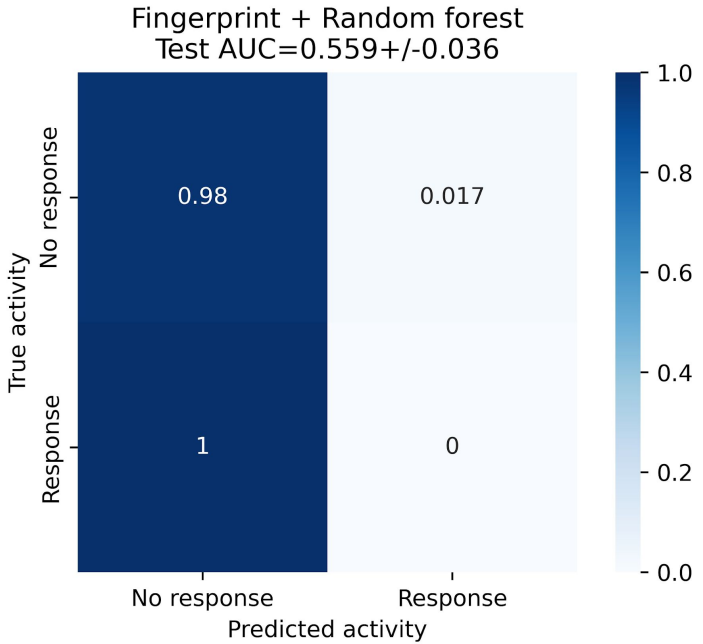


Predicting receptor activity from representations of chemical ligands

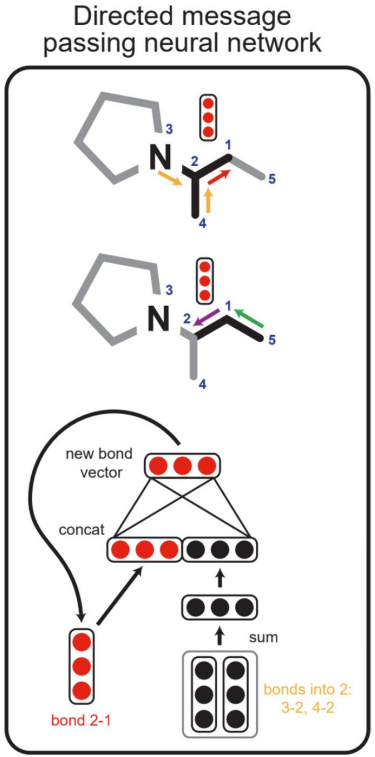
Classic machine learning models do not perform well on molecular fingerprints



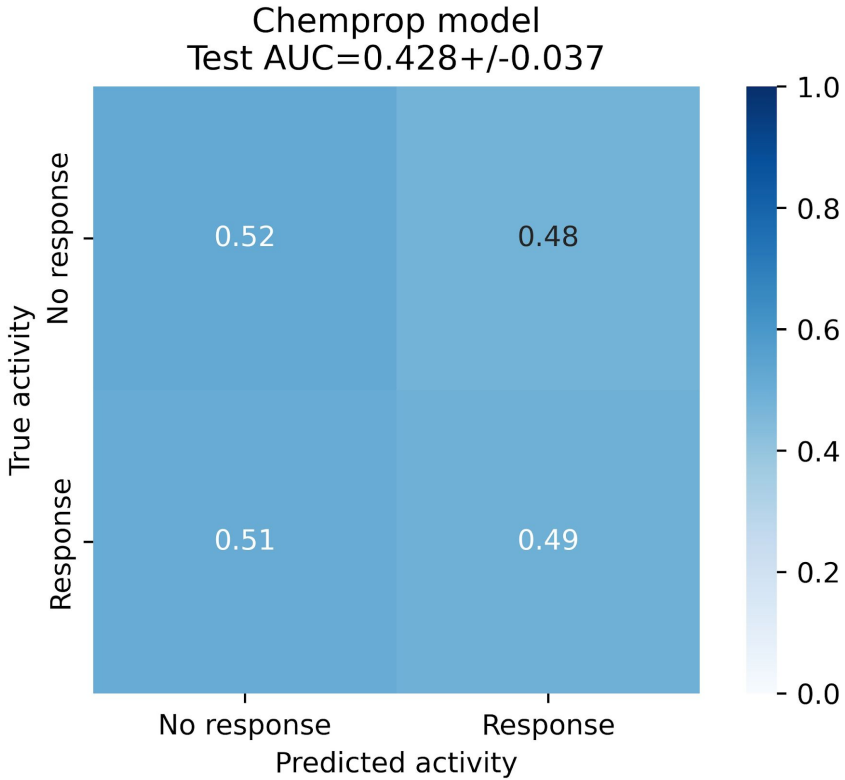
Molecular fingerprinting



Graph-based message passing neural network

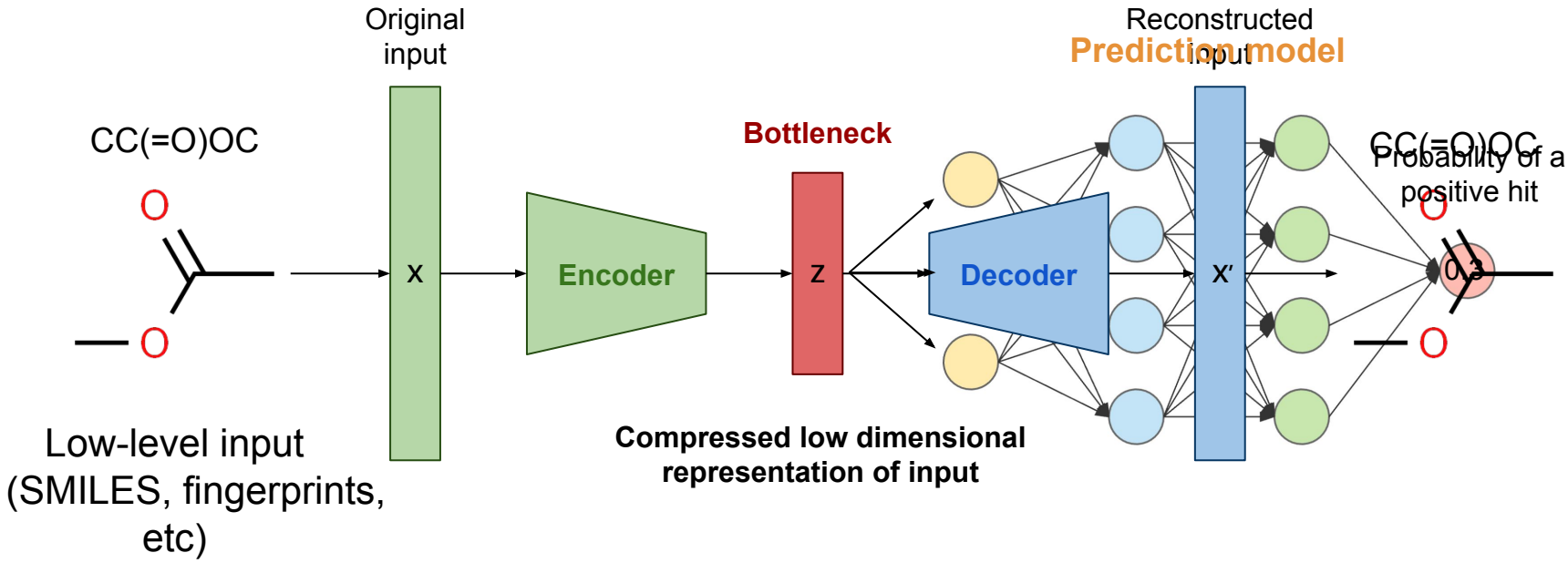


Stokes et al. 2020



Given a small dataset, reducing task complexity is essential

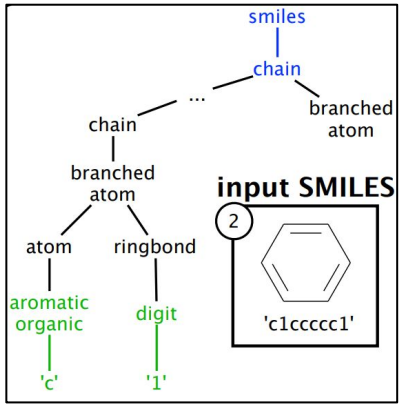
Feature abstraction with variational autoencoders (deep neural network)



Recent models

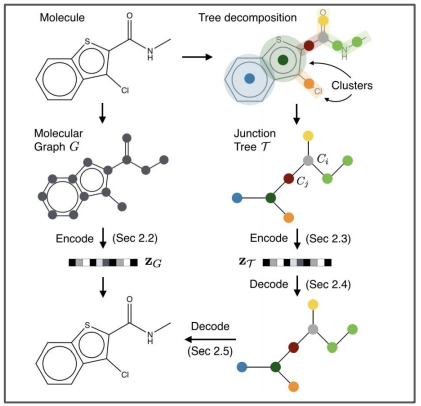
Grammar Variational Autoencoder

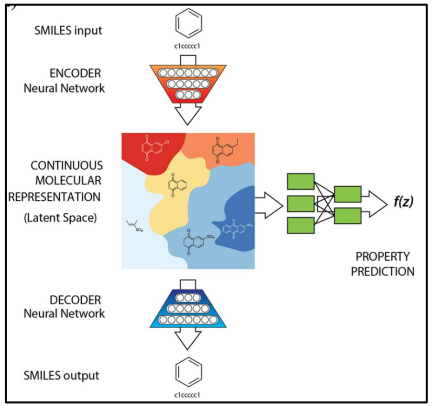
Kushner et al. 2017



Junction Tree Variational Autoencoder for Molecular Graph Generation

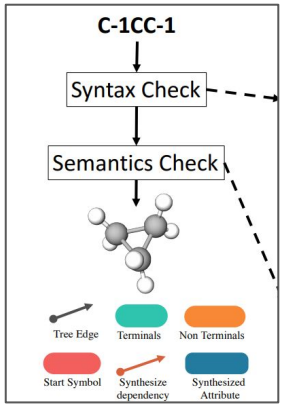
Jin et al. 2018





Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Gómez-Bombarelli et al. 2016

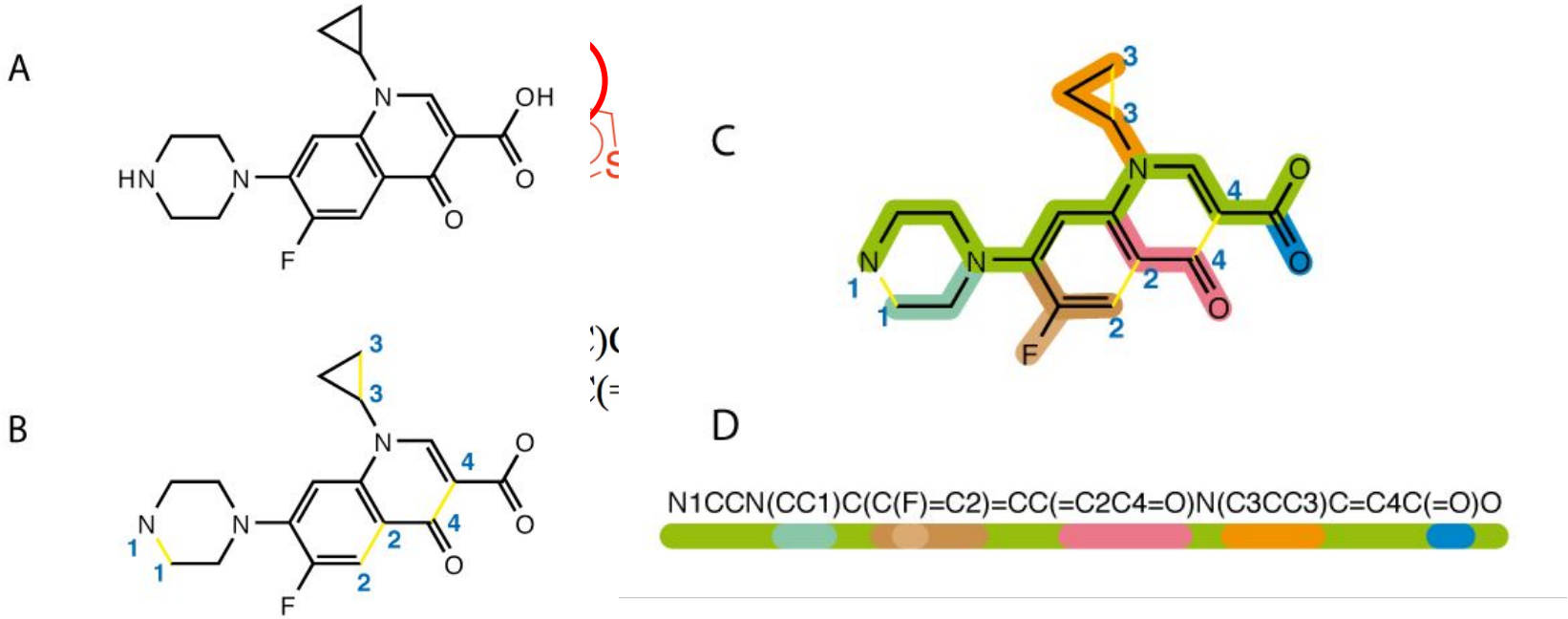


Syntax-Directed Variational Autoencoder for Structured Data

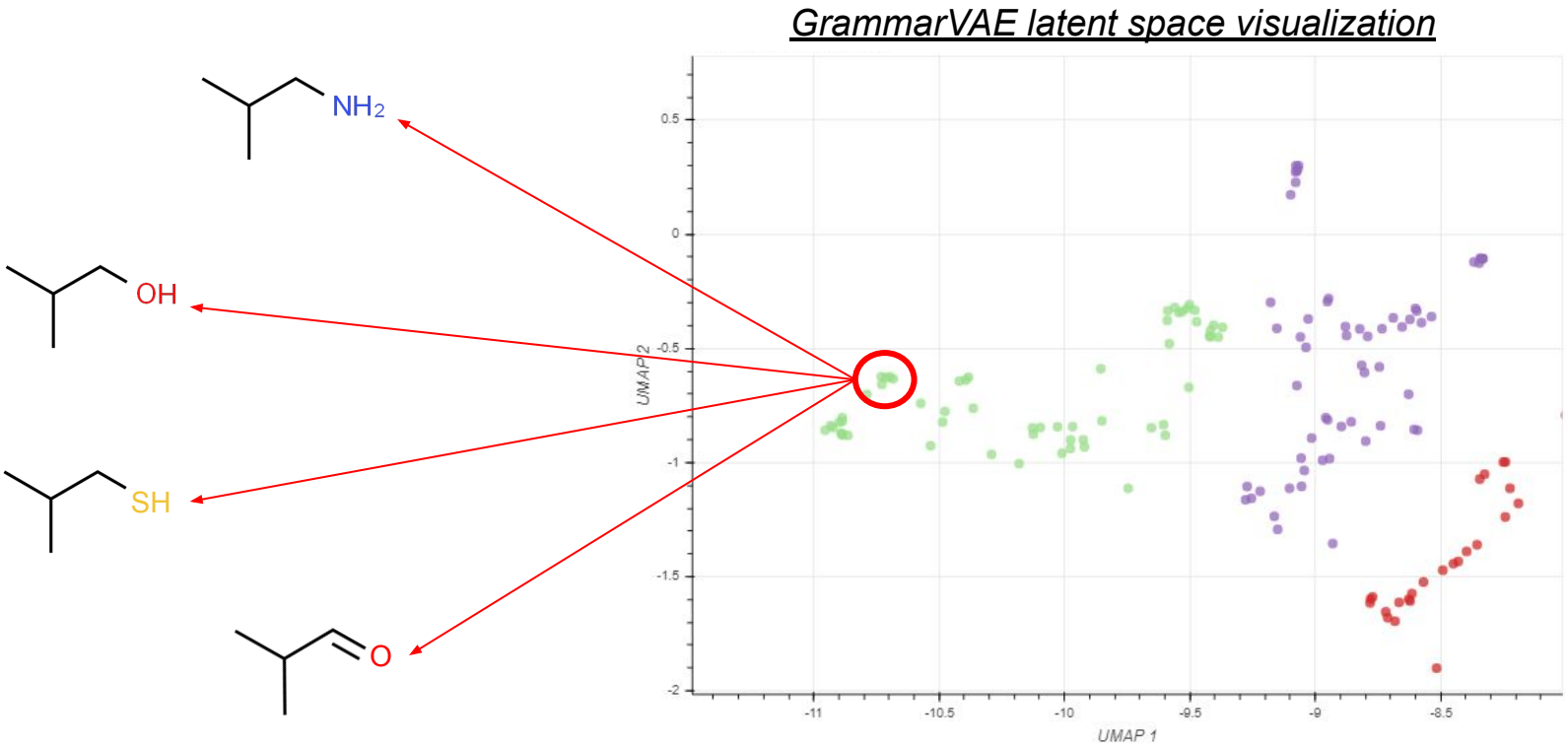
Dai et al. 2018

Issues with existing SMILES-based models

Simplified Molecular-Input Line-Entry System



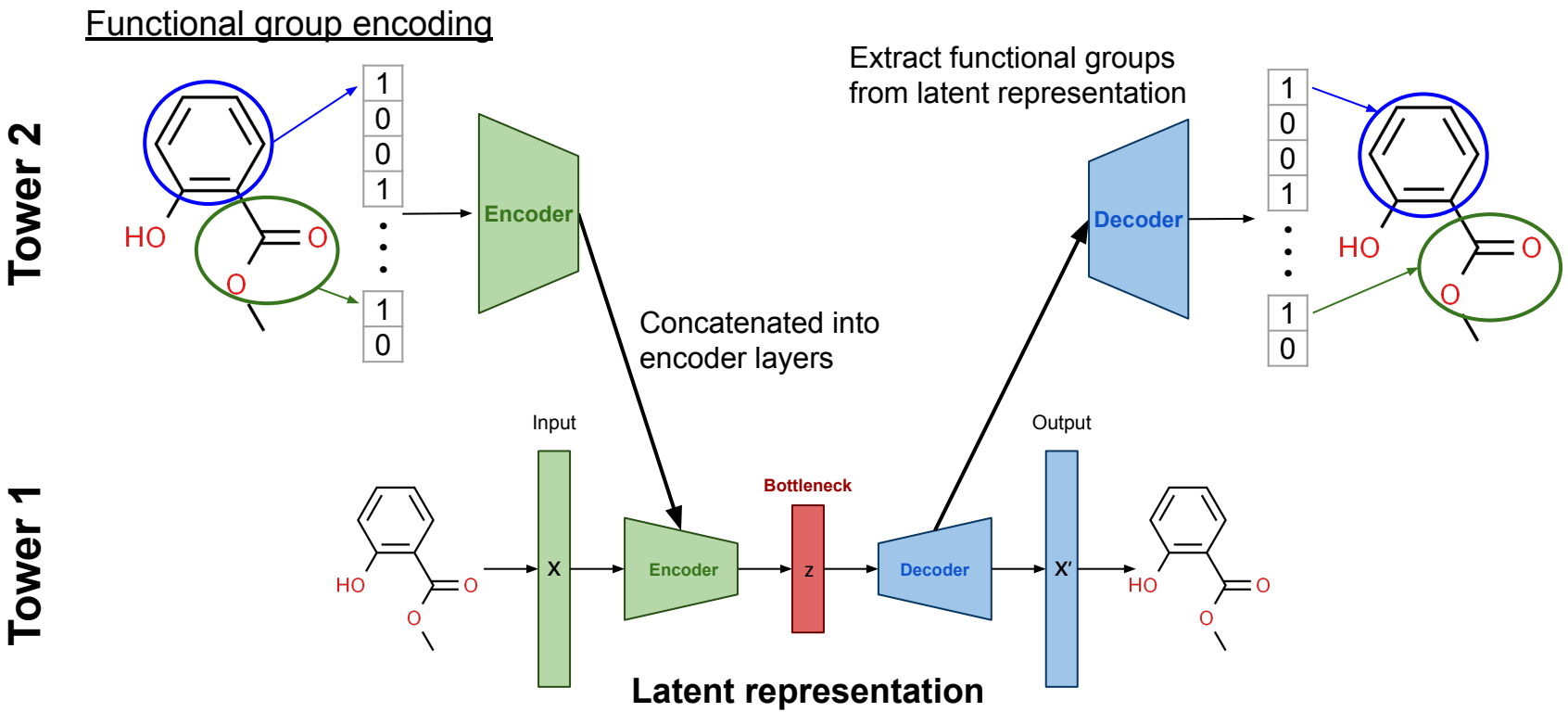
Some models overemphasize molecular geometry



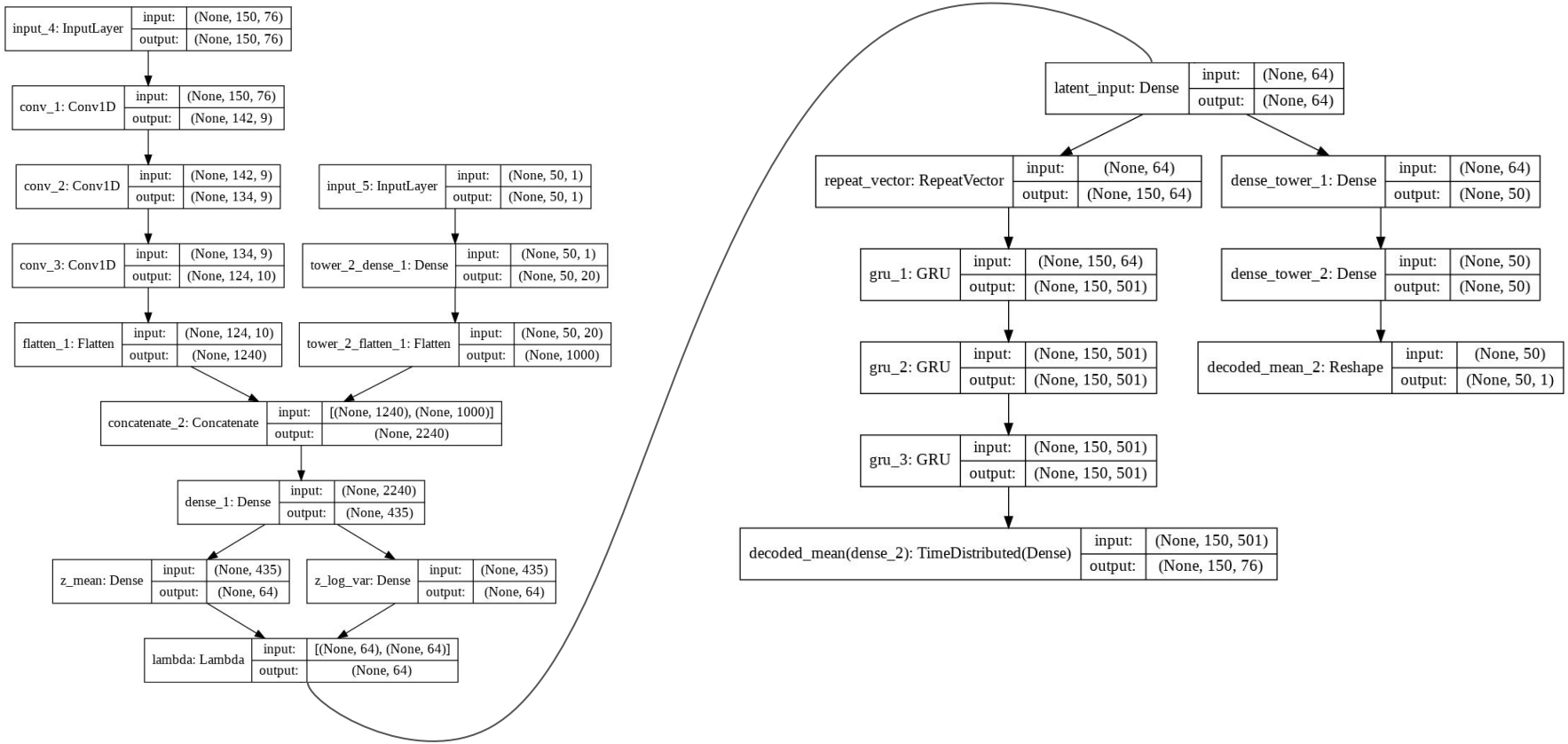
Kushner et al. 2017

Our Approach

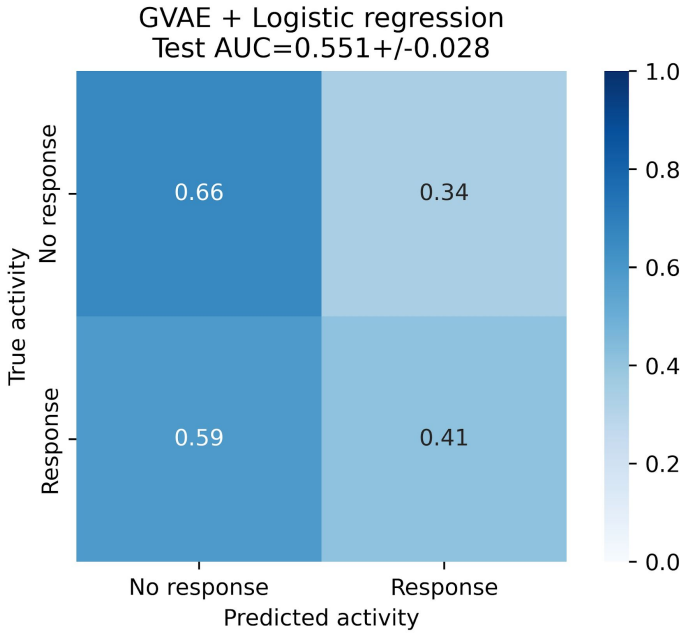
Our model: two-tower approach

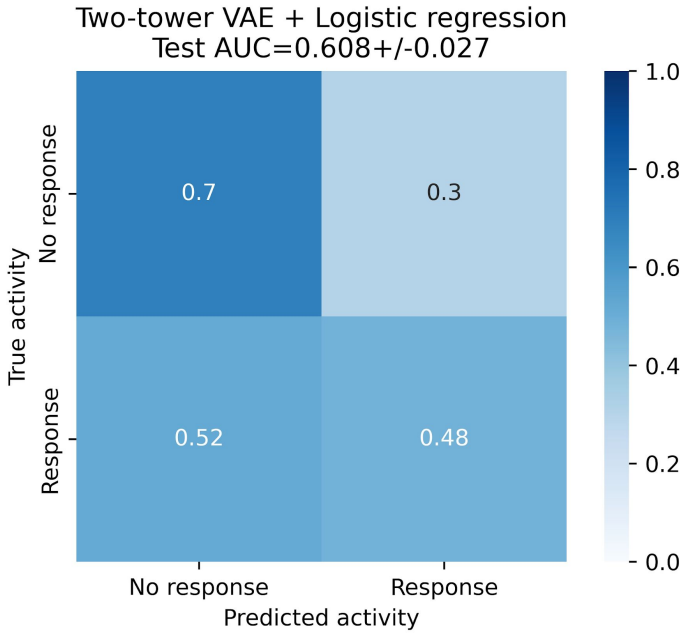


Two-tower architecture

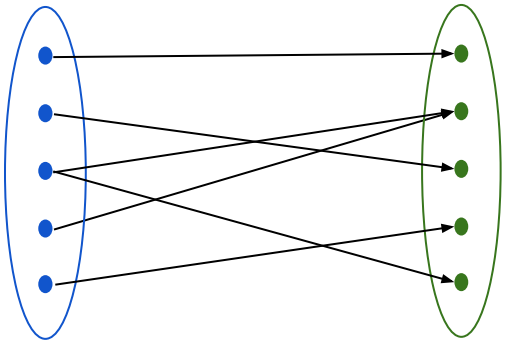


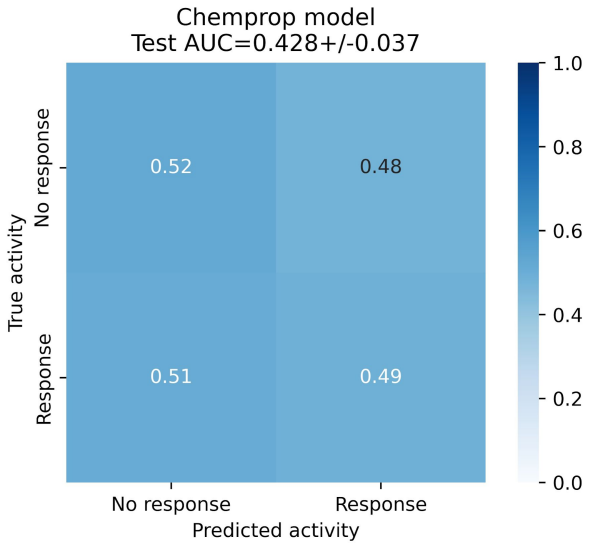
Our results

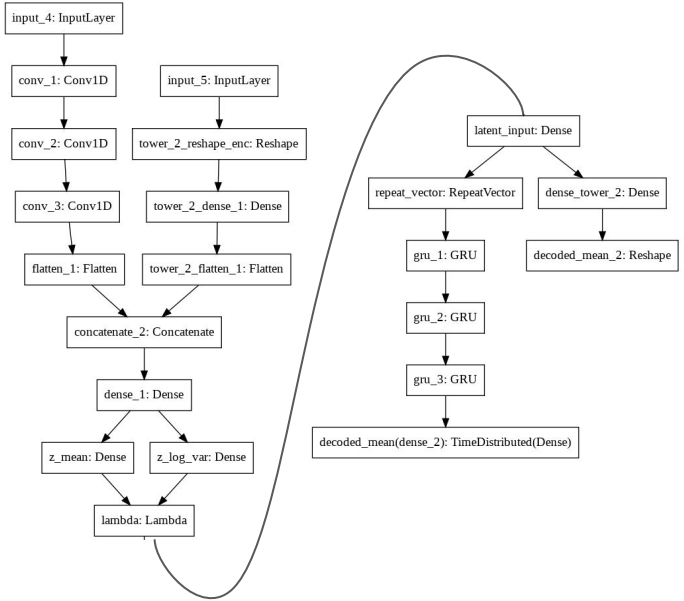

GVAE + Logistic regression
Test AUC=0.551+/-0.028
True activity
No response
Response
0.66
0.34
0.59
0.41
No response
Response
Predicted activity
1.0
0.8
0.6
0.4
0.2
0.0


Two-tower VAE + Logistic regression
Test AUC=0.608+/-0.027
True activity
No response
Response
0.7
0.3
0.52
0.48
No response
Response
Predicted activity
1.0
0.8
0.6
0.4
0.2
0.0

Recap







Importance of receptor-ligand binding

Difficulties of model training in data-limited settings

Predicting receptor Activities with VAEs