Latent representations of chemical ligands to predict combinatorial receptor-ligand interactions

Jonathan Yin
Mentors: Dr. Hattie Chung, Michael Truell, Regev Group, Broad Institute

MIT PRIMES Conference, October 20, 2019
Receptors process signals from the environment

Ligand binds to a receptor

Ligand-receptor complex triggers signal pathway

Extracellular

Cell Membrane

Intracellular

Intracellular Response
Understanding combinatorial receptor-ligand interactions
Similar chemicals activate similar sets of receptors

Esters

- Methanol
- Ethanol
- 1-propanol
- 1-butanol
- 1-pentanol
- 1-hexanol
- 1-octanal
- 1-pentanol
- 1-methylbutanol
- 1-methyl-2-buten-1-ol
- 1-penten-3-ol
- 1-octen-3-ol
- 2-hexanol
- 2-hexenol
- 3-hexenol
- 3-hexenol
- Glycerol
- 2,3-butanediol
- Methyl acetate
- Ethyl acetate
- Propyl acetate
- Butyl acetate
- Pentyl acetate
- Hexyl acetate
- Isobutyl acetate
- Isopentyl acetate
- E2-hexenyl acetate
- Methyl butyrate
- Ethyl butyrate
- Hexyl butyrate
- Ethyl 3-hydroxybutyrate
- Ethyl propionate
- Ethyl methanoate

Alcohols

- Ex: 1-Propanol

- Ex: Ethyl acetate

Motivation

Predicting ligand-receptor interactions

Ability to control intracellular behaviors
Challenge: representing inputs for predictive algorithms

Images

Text

Raw Text

Bag-of-words

vector

it 2
they 0
puppy 1
and 1
cat 0
aardvark 0
cute 1
extremely 1
...
...

it is a puppy and it is extremely cute
Low-level representation schemes: molecular fingerprints

Extracts molecular substructures and converts to bit vector

1024 or 2048 bits
Low-level representation schemes: SMILES

Simplified Molecular-Input Line-Entry System

A

\[
\text{HN}\text{N}\text{F} \quad \text{N} \quad \text{N} - \text{O} \\
\text{HN}\text{N} - \text{O} \quad \text{F} \quad \text{N} - \text{O}
\]

B

C

D

\[
\text{N1CN(CC1)C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O}
\]
Machine learning models typically require high quantities of data

> $10^6$ ligands

> $10^3$ receptors
In the absence of a large dataset, feature abstraction is necessary.
Recent trend: feature abstraction with variational autoencoders (deep neural network)

Low-level input (SMILES, fingerprints, etc)

Compressed low dimensional representation of input
Current 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

Cc1cn2c(CN(C)C(=O)c3ccc(F)cc3C)c(C)nc2s1
Cc1cc(F)c1ccc1C(=O)N(C)Cc1c(C)nc2cc2c(C)n12
Current models overemphasize molecular geometry

GrammarVAE latent space visualization

Kushner et al. 2017
Evaluating the latent space of current models

Molecular fingerprint still performs the best

Correlation between distance in latent space and receptor activity

Our Approach
Incorporating prior knowledge helps the model
Our model: two-tower approach

Functional group encoding

Tower 2

Tower 1

Extract functional groups from latent representation

Concatenated into encoder layers

Latent representation
Two-tower architecture
Our results

Two-tower model VAE
$r = -0.3449$

Morgan Fingerprint
$r = 0.4159$
Recap

Combinatorial Ligand-receptor binding

Feature abstraction with VAEs: Two-tower approach

Predicting receptor activities