Machine Learning and Drug Discovery

Amir Saffari, Director of Applied AI

@amirsaffari

amir.saffariazar@benevolent.ai

BenevolentAI

We are BenevolentAI

Our mission is to invent and apply AI and Machine Learning technology to accelerate scientific discovery that benefits society.

BenevolentAI Overview

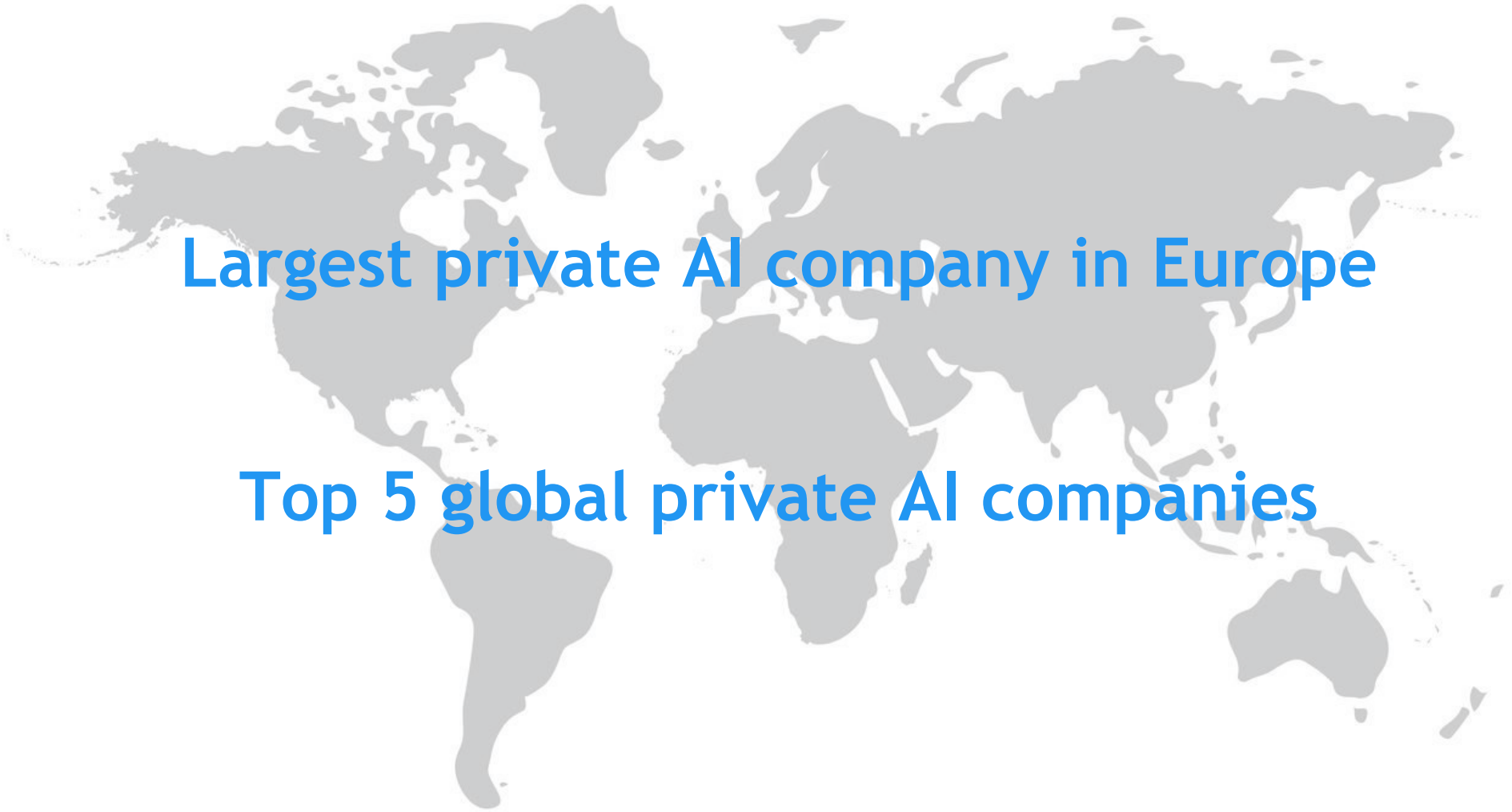
Founded in 2013

$BN company

80+ team of world class scientists and technologists (50+ doctorates/advanced degrees)

Technology enabling **previously impossible** tasks in bioscience

Rich patent portfolio of over 400 patents



Largest private AI company in Europe

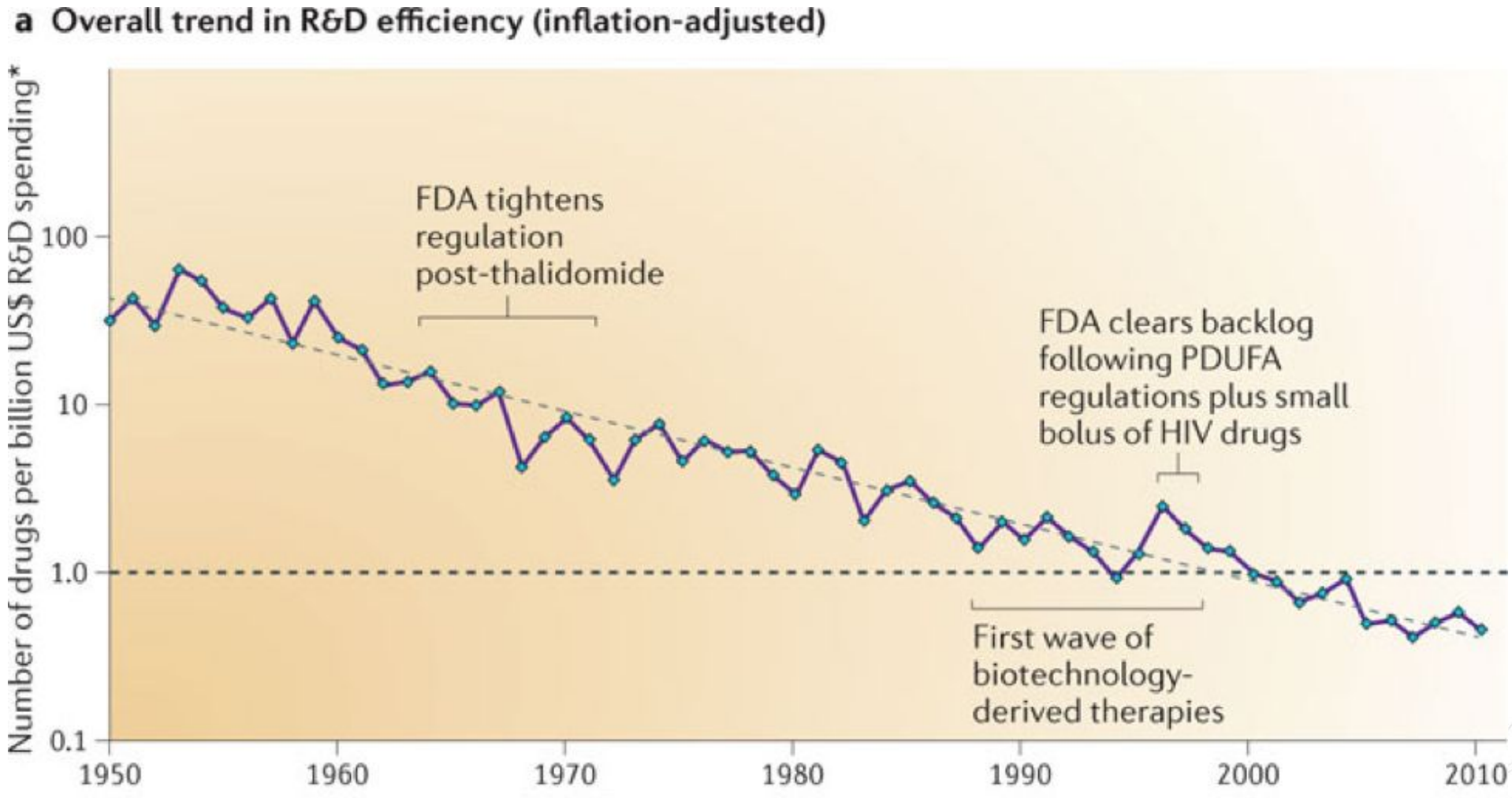
Top 5 global private AI companies

Focused on accelerating Drug Discovery as the first scientific domain

Drug Discovery is broken!

- On average it costs about $2.6bn to develop a new drug
- It requires 12-15 years of R&D from start to market
- 97% of drug programmes fail
- Only less that 40% of known diseases are currently treatable
- Novel discoveries are rare and is mainly done by academia research

Eroom's Law

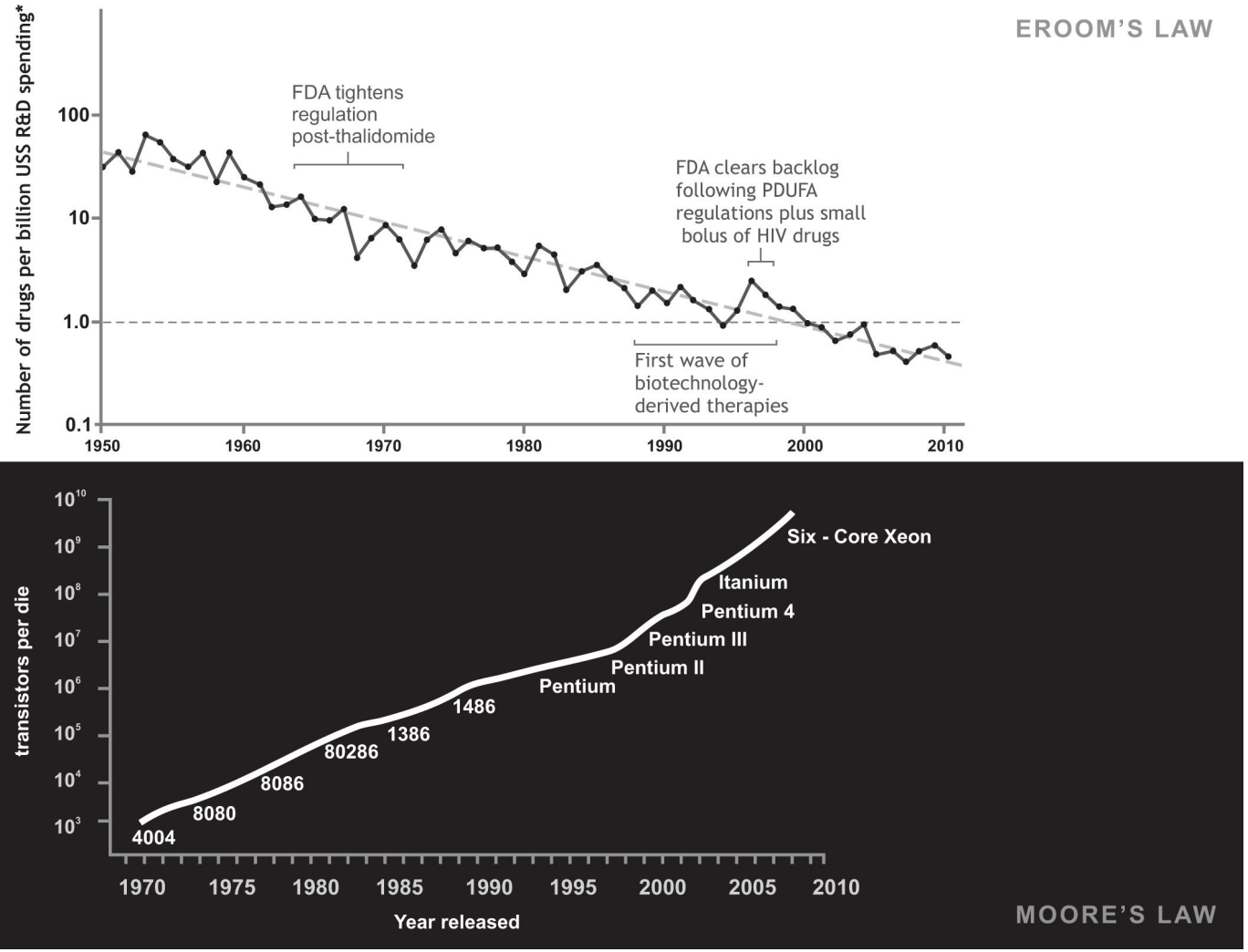


Diagnosing the decline in pharmaceutical R&D efficiency

Jack W. Scannell, Alex Blanckley, Helen Boldon & Brian Warrington

Nature Reviews Drug Discovery **11**, 191-200 (March 2012)

Eroom's Law and Moore's Law



https://buildingpharmabrands.com/2013/08/16/erooms-law-of-pharma-r-d/

Power of the AI technology - Knowledge & Reasoning

1 scientific paper is published every 30 seconds

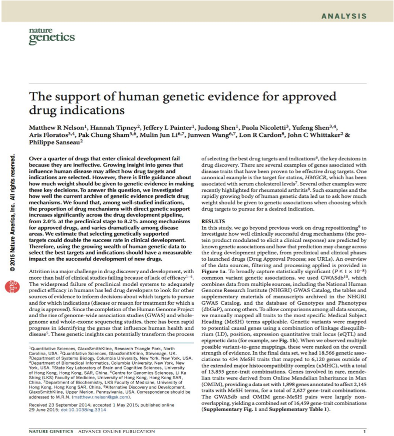
Individual scientists access a tiny fraction of available data in their lifetime - much less than the data generated in 1 day

All biological databases combined are less than 5% of available data

The BenevolentAI platform extracts facts and reasons from all relevant databases and literature, structured and unstructured

Proprietary Knowledge and Inference Models

Unstructured Data


ANALYSIS

nature genetics

The support of human genetic evidence for approved drug indications

Matthew R Nelson[1], Hannah Tipney[2], Jeffery L Painter[1], Judong Shen[1], Paola Nicoletti[3], Yufeng Shen[3,4], Aris Floratos[3,4], Pak Chung Sham[5,6], Mulin Jun Li[6,7], Junwen Wang[6,7], Lon R Cardon[8], John C Whittaker[2] & Philippe Sanseau[2]

Over a quarter of drugs that enter clinical development fail because they are ineffective. Growing insight into genes that influence human disease may affect how drug targets and indications are selected. However, there is little guidance about how much weight should be given to genetic evidence in making these key decisions. To answer this question, we investigated how well the current archive of genetic evidence predicts drug mechanisms. We found that, among well-studied indications, the proportion of drug mechanisms with direct genetic support increases significantly across the drug development pipeline, from 2.0% at the preclinical stage to 8.2% among mechanisms for approved drugs, and varies dramatically among disease areas. We estimate that selecting genetically supported targets could double the success rate in clinical development. Therefore, using the growing wealth of human genetic data to select the best targets and indications should have a measurable impact on the successful development of new drugs.

Attrition is a major challenge in drug discovery and development, with more than half of clinical studies failing because of lack of efficacy[1–4]. The widespread failure of preclinical model systems to adequately predict efficacy in humans has led drug developers to look for other sources of evidence to inform decisions about which targets to pursue and for which indications (disease or reason for treatment for which a drug is approved). Since the completion of the Human Genome Project and the rise of genome-wide association studies (GWAS) and whole-genome and whole-exome sequencing studies, there has been rapid progress in identifying the genes that influence human health and disease[5]. These genetic insights can potentially transform the process of selecting the best drug targets and indications[6], the key decisions in drug discovery. There are several examples of genes associated with disease traits that have been proven to be effective drug targets. One canonical example is the target for statins, *HMGCR*, which has been associated with serum cholesterol levels[7]. Several other examples were recently highlighted for rheumatoid arthritis[8]. Such examples and the rapidly growing body of human genetic data led us to ask how much weight should be given to genetic associations when choosing which drug targets to pursue for a desired indication.

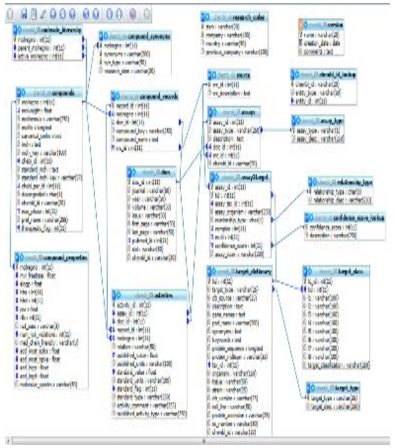
[1]Quantitative Sciences, GlaxoSmithKline, Research Triangle Park, North Carolina, USA. [2]Quantitative Sciences, GlaxoSmithKline, Stevenage, UK. [3]Department of Systems Biology, Columbia University, New York, New York, USA. [4]Department of Biomedical Informatics, Columbia University, New York, New York, USA. [5]State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong, SAR, China. [6]Centre for Genomics Sciences, Li Ka Shing (LKS) Faculty of Medicine, University of Hong Kong, Hong Kong SAR, China. [7]Department of Biochemistry, LKS Faculty of Medicine, University of Hong Kong, Hong Kong SAR, China. [8]Alternative Discovery and Development, GlaxoSmithKline, Upper Merion, Pennsylvania, USA. Correspondence should be addressed to M.R.N. (matthew.r.nelson@gsk.com).

Received 23 September 2014; accepted 1 May 2015; published online 29 June 2015; doi:10.1038/ng.3314

RESULTS

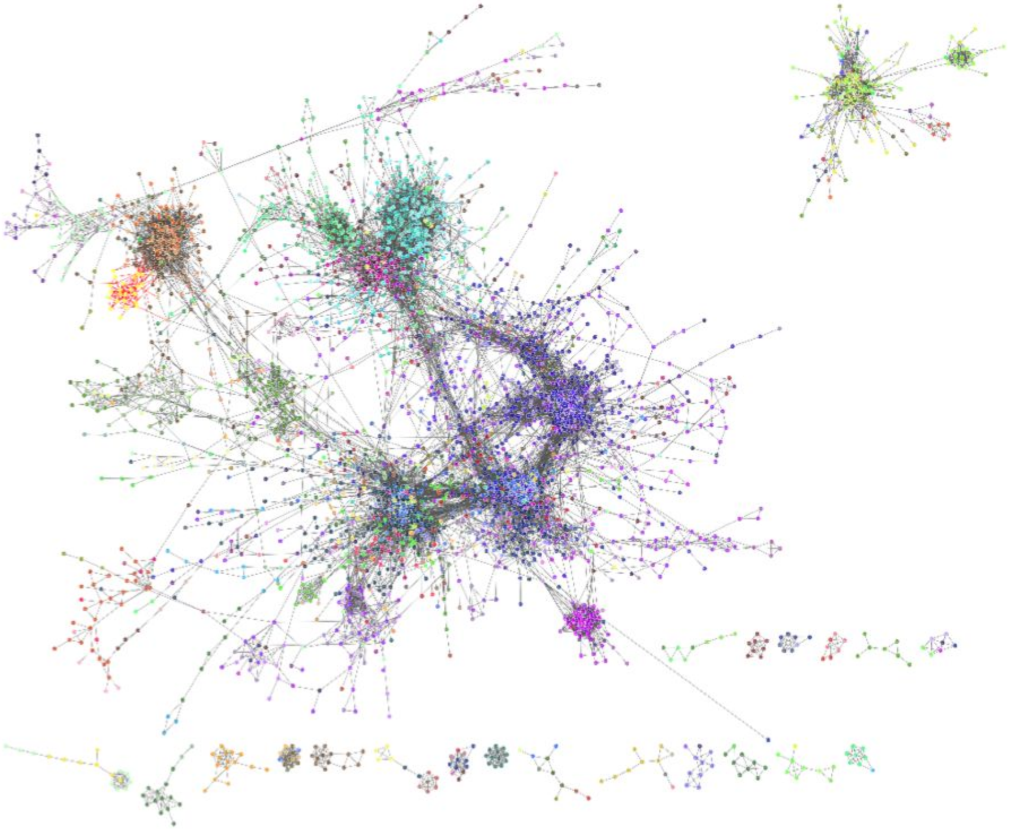
In this study, we go beyond previous work on drug repositioning[9] to investigate how well clinically successful drug mechanisms (the protein product modulated to elicit a clinical response) are predicted by known genetic associations and how that prediction may change across the drug development pipeline, from preclinical and clinical phases to launched drugs (Drug Approval Process; see URLs). An overview of the data sources, filtering and processing applied is provided in **Figure 1a**. To broadly capture statistically significant ($P \leq 1 \times 10^{-8}$) common variant genetic associations, we used GWASdb[10], which combines data from multiple sources, including the National Human Genome Research Institute (NHGRI) GWAS Catalog, the tables and supplementary materials of manuscripts archived in the NHGRI GWAS Catalog, and the database of Genotypes and Phenotypes (dbGaP), among others. To allow comparisons among all data sources, we manually mapped all traits to the most specific Medical Subject Heading (MeSH) terms applicable. Genetic variants were mapped to potential causal genes using a combination of linkage disequilibrium (LD), position, expression quantitative trait locus (eQTL) and epigenetic data (for example, see **Fig. 1b**). When we observed multiple possible variant-to-gene mappings, these were ranked on the overall strength of evidence. In the final data set, we had 18,566 genetic associations to 434 MeSH traits that mapped to 6,120 genes outside of the extended major histocompatibility complex (xMHC), with a total of 13,855 gene-trait combinations. Genes involved in rare, mendelian traits were derived from Online Mendelian Inheritance in Man (OMIM), providing a data set with 1,898 genes annotated to affect 2,145 traits with MeSH terms, for a total of 2,627 gene-trait combinations. The GWASdb and OMIM gene-MeSH pairs were largely non-overlapping, yielding a combined set of 16,459 gene-trait combinations (**Supplementary Fig. 1** and **Supplementary Table 1**).

Structured Data

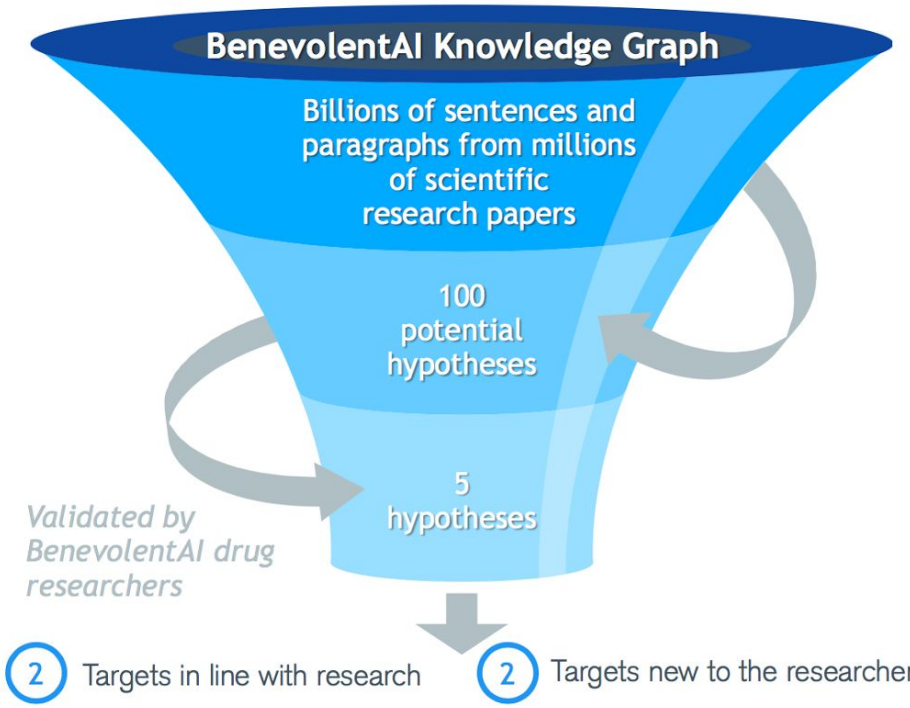


Large Scale NLP, Machine Learning, Deep Learning, and Neural Reasoning Technology

Rich proprietary knowledge



BenevolentAI finds potential treatments for ALS



Validated by SITraN, expert researchers on ALS

"I was impressed by the initial hypotheses that were presented at our first meeting together. One of the ideas was very much a central part of our ongoing research and immediately validated the technology. The second idea was quite a novel one for treatments for ALS. Without the augmentation and insight that the technology allows us, these are possibilities we may have otherwise overlooked."

Dr. Mead, Sheffield Institute of Translational Neuroscience

Medicine and computing

The shoulders of gAInts

The Economist

Artificial intelligence may help unpick the complexity of biology

"BenevolentAI has already validated on pathway for drug discovery and opened up a surprising new one. What their engine can do is look across vast swaths of information to pick novel ideas repurpose…BenevolentAI has given me two ideas for drugs for ALS, one is bang in the middle of what I and my teams are doing already…AI is generating good ideas. The other is complicated and not obvious, but mechanistically interesting."

Dr Richard Mead.. Lecturer in Neuroscience, University of Sheffield

"Benevolent AI is not a one-off. More and more people and firms believe that AI is well placed to help unpick biology and advance human health…It can in weeks elucidate salient links and offer new ideas that would take lifetimes of human endeavor to come up with. I can also weigh up the evidence for its hypotheses in an evenhanded manner."

The Economist

FINANCIAL TIMES

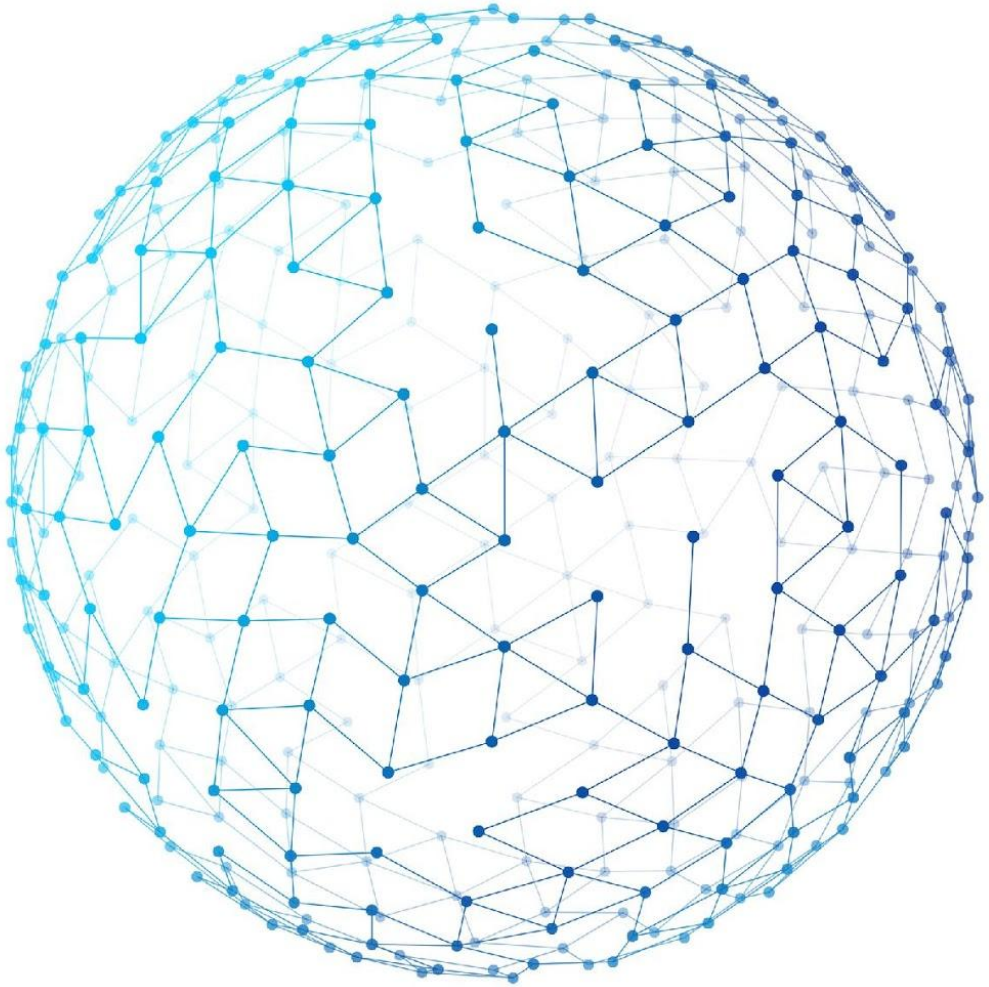
Bold claims for AI are hard to compute for economist

"AI-enabled technologies could double the economic growth rates of many advance countries by 2035"

The Financial Times

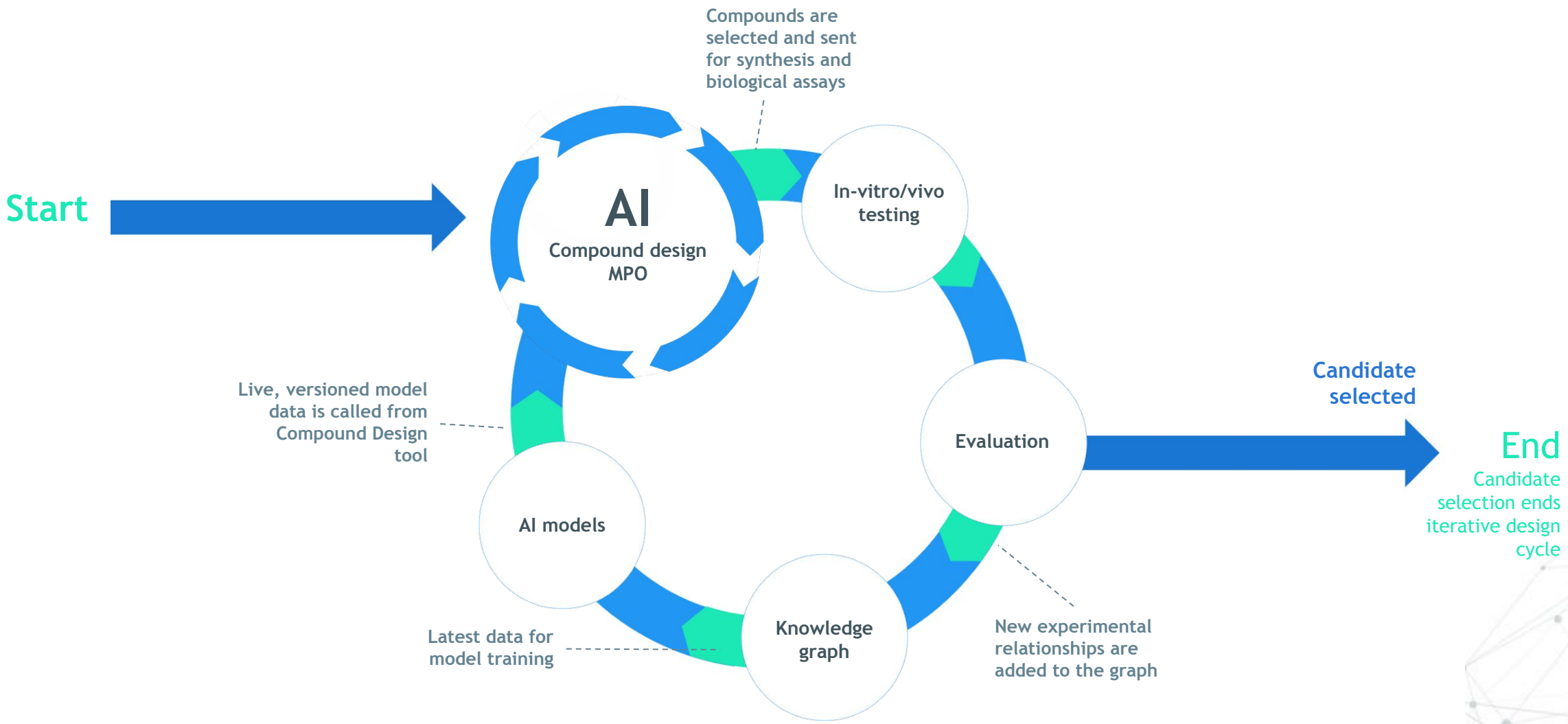
"It [BenevolentAI] can also personalise solutions for individuals according to their genetic make up, we are really excited about it, the potential is incredible"

Dr Laura Ferraiuolo, PhD. Lecturer in Translational Neurology

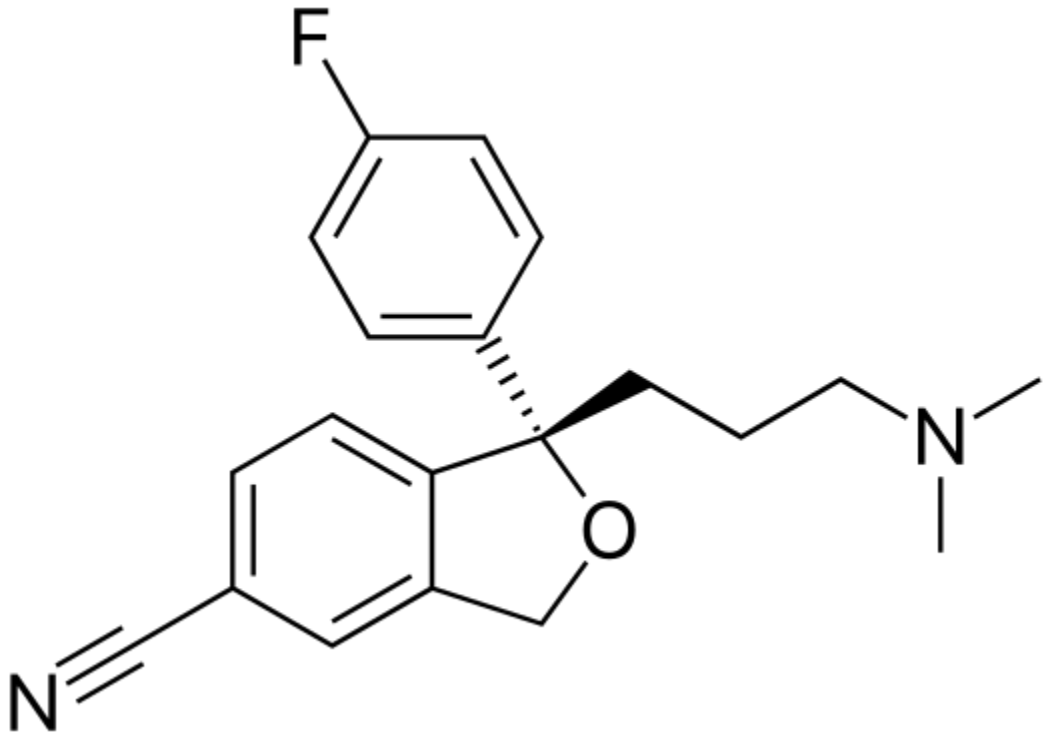


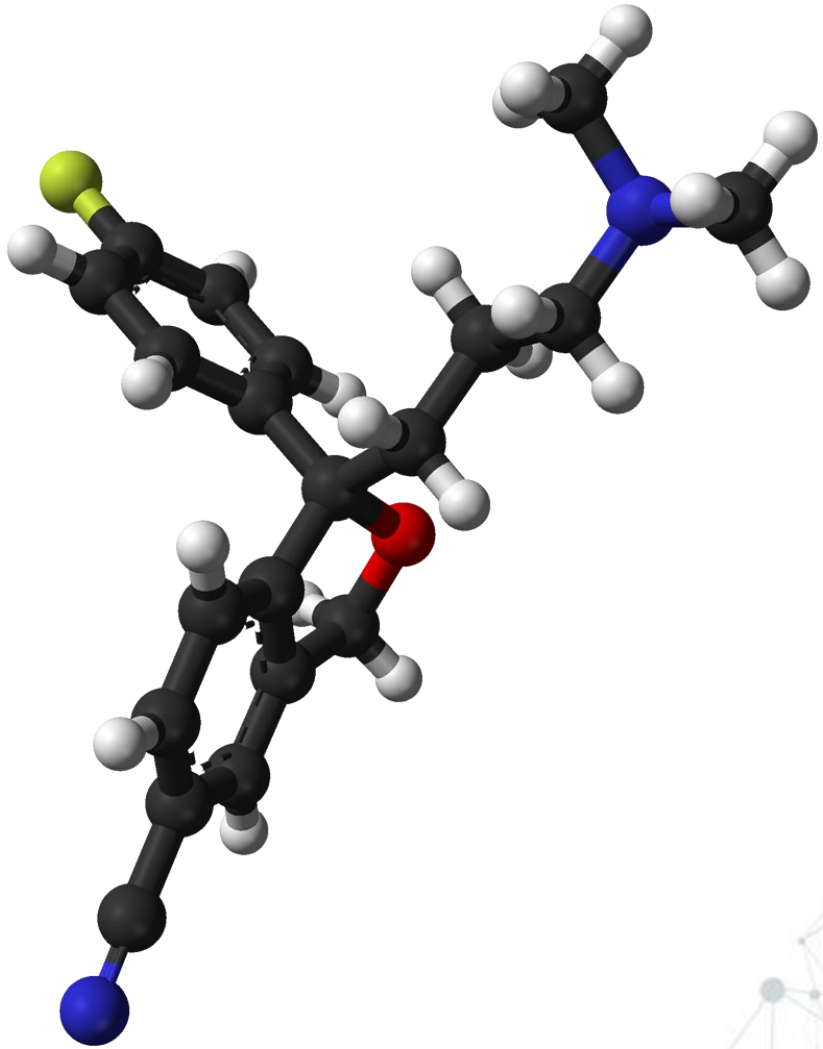
AI-Driven Compound Design

AI-Driven Compound Design Overview



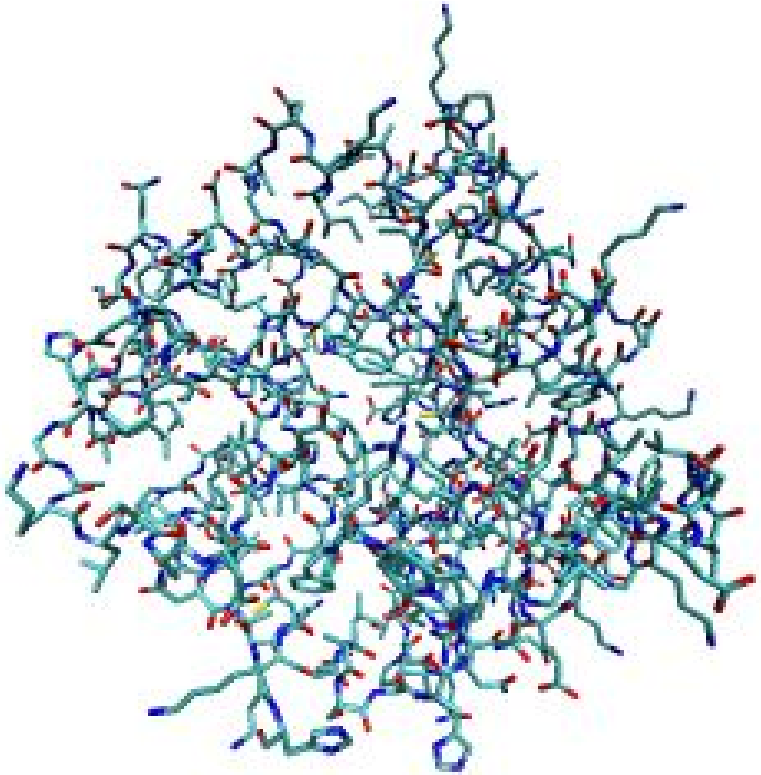
Basics: Drugs are molecules

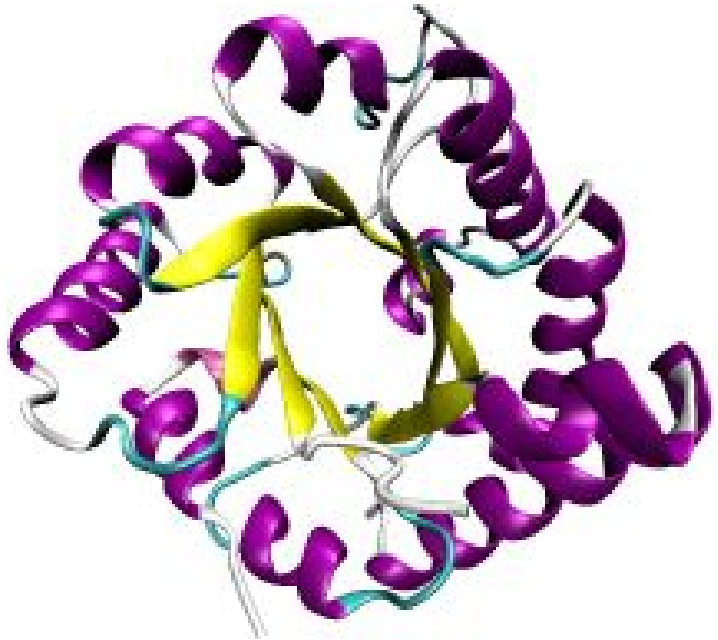


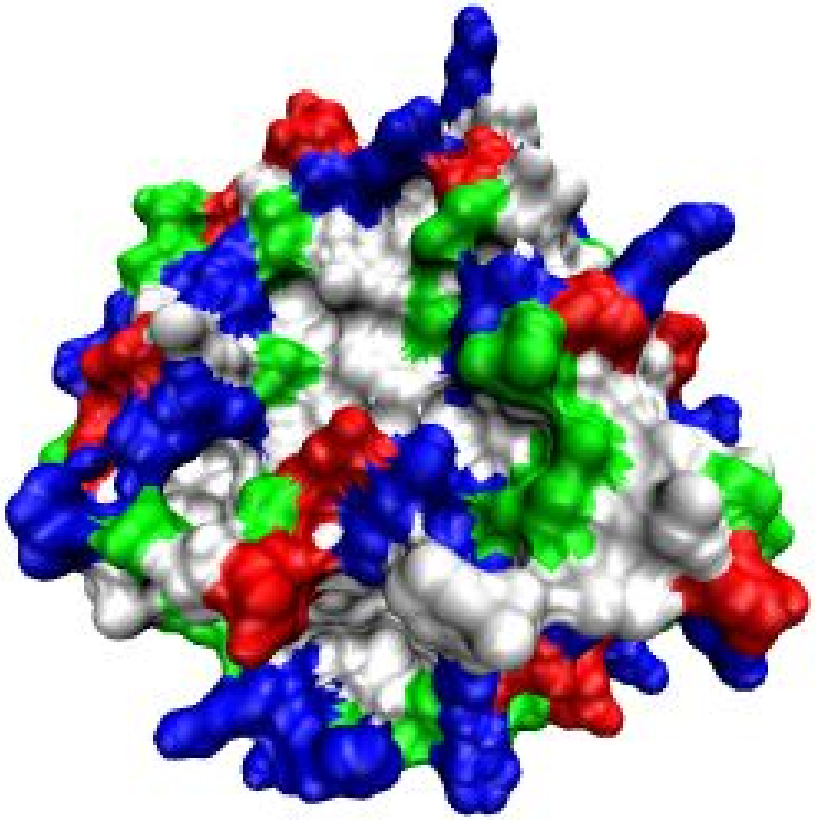


https://en.wikipedia.org/wiki/Escitalopram

Basics: Proteins are biomolecules

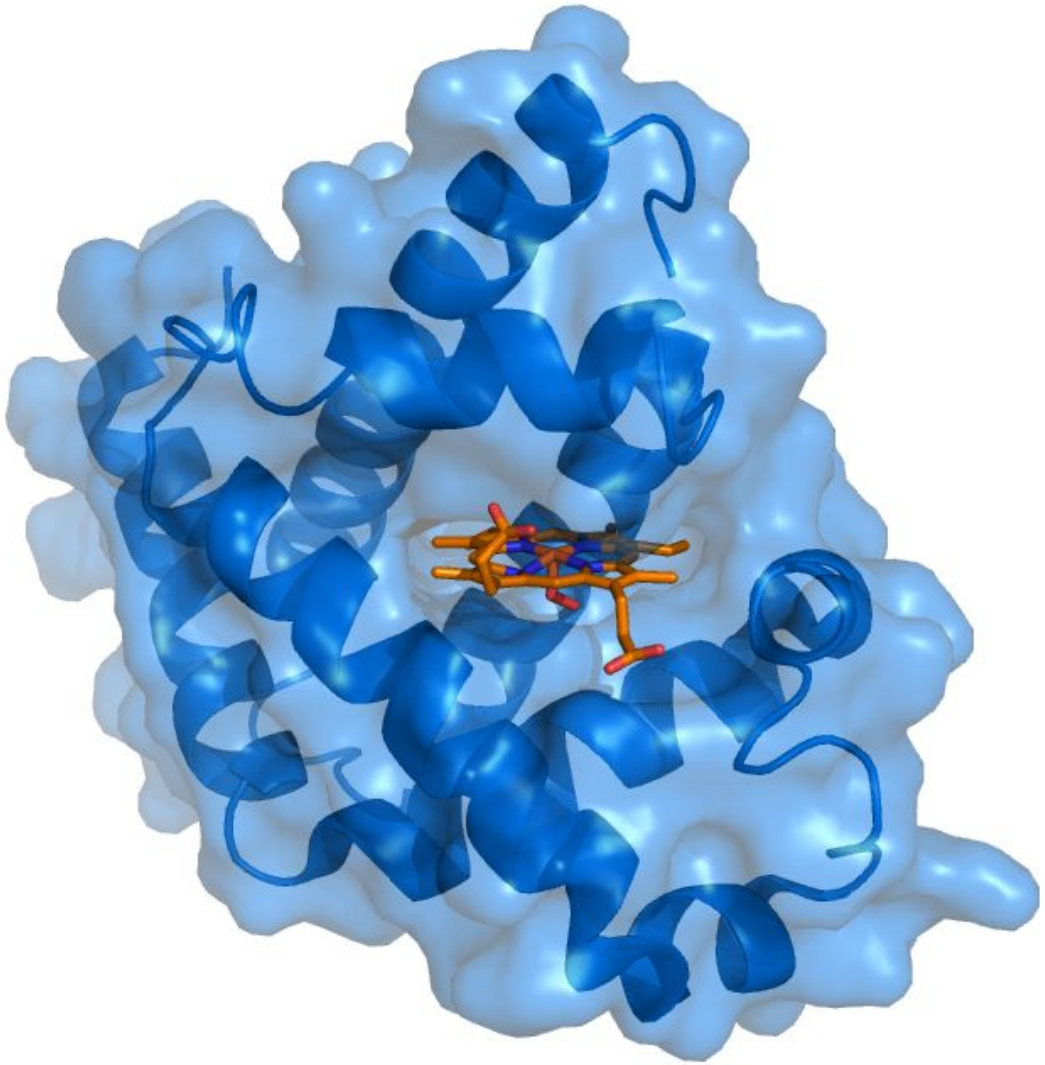






https://en.wikipedia.org/wiki/Protein

Basics: Ligand-Protein Binding

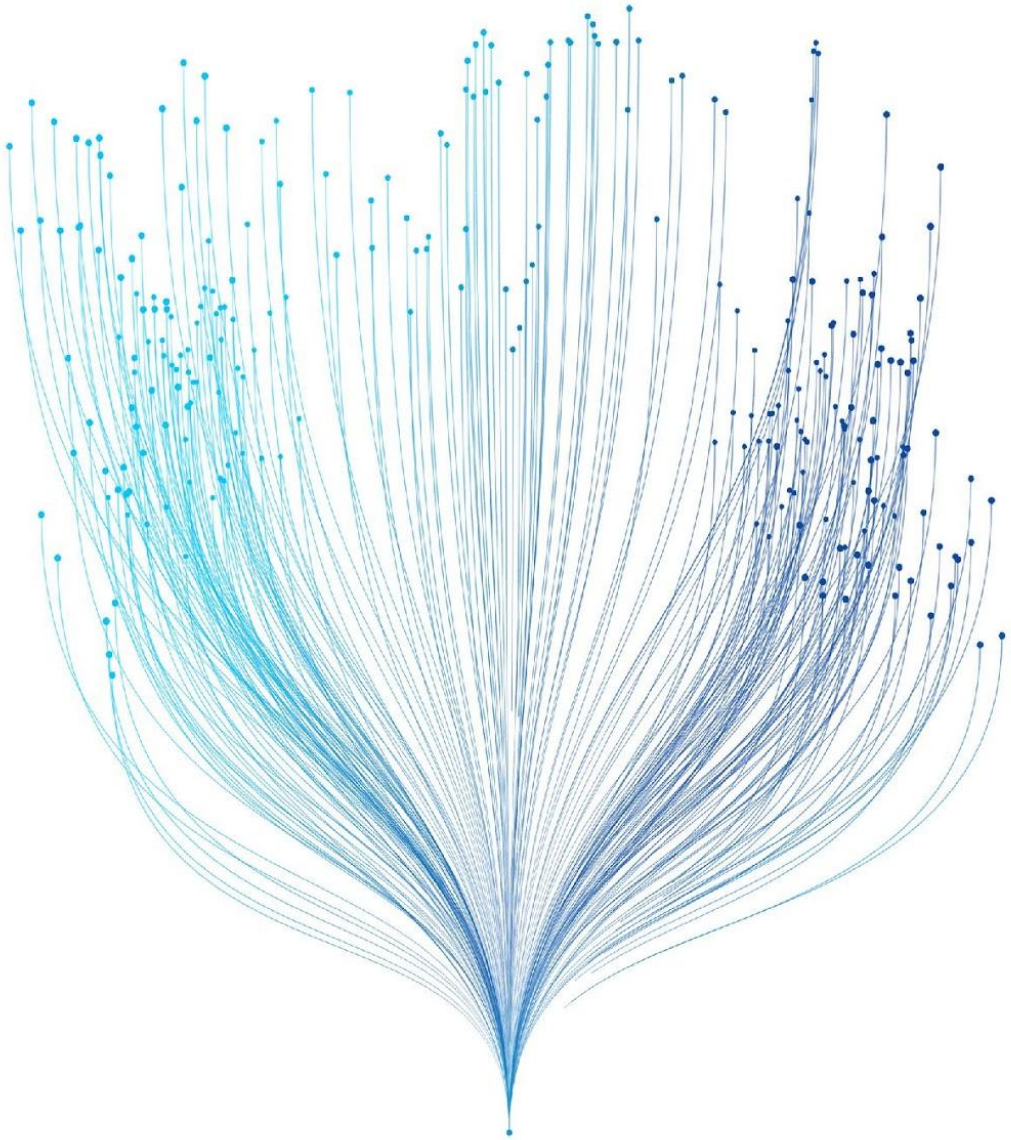


https://en.wikipedia.org/wiki/Ligand_(biochemistry)

ML Tasks for Drug Design

- Property prediction
  - Given a compound, predict biochemical properties
  - Given a compound and a protein, predict their binding affinity

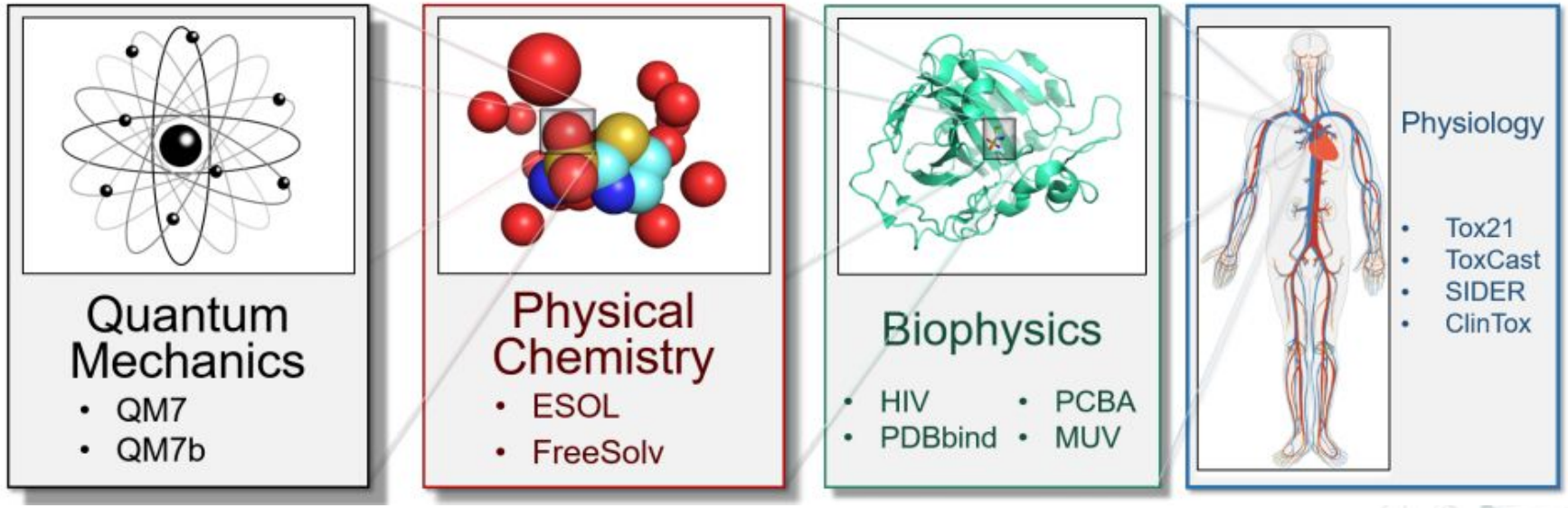
- Generative chemistry
  - Given a set of goals, generate a series of new molecule (maybe from a starting one) that best optimise those goals



Property Prediction

Molecular Property Prediction: tasks and datasets

- MoleculeNet
  - Aims to be ImageNet of molecular ML
  - A mid-size dataset of about 600K molecules and less than 800 tasks
  - Binary classification or regression
  - Under active development
- BenevolentAI
  - In reality we care about more than 1800 tasks



MoleculeNet: A Benchmark for Molecular Machine Learning
Wu et al, arXiv 2017, https://arxiv.org/abs/1703.00564

Dataset Challenges

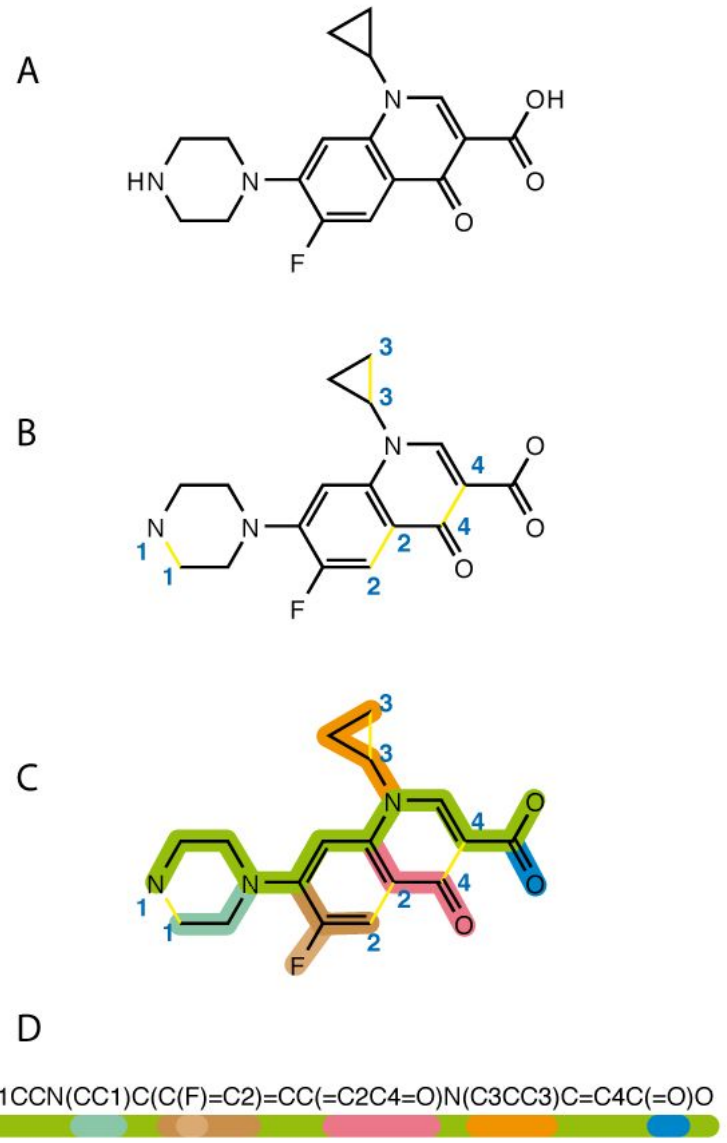
- Lack of labelled data
  - Small in scale compared to other data rich domains (e.g. vision)
- Unbalanced data
  - 1:50 pos to neg ratio
- Biased datasets
  - Similar molecules with similar properties
  - Makes it difficult to rely on pure random k-fold CV
  - Better to use scaffold
- Massive unlabelled data
  - GDB-13 ~ 1B
  - GDB-17 ~ 166B

Category	Dataset	Description	Tasks	Compounds
Quantum Mechanics	QM7	atomization energy	1	7165
	QM7b	electronic properties	14	7211
Physical Chemistry	ESOL	solubility	1	1128
	FreeSolv	solvation energy	1	643
Biophysics	PCBA	bioactivity	128	439863
	MUV	bioactivity	17	93127
	PDBbind	binding affinity	1	11908
	HIV	bioactivity	1	41913
Physiology	Tox21	toxicity	12	8014
	ToxCast	toxicity	617	8615
	SIDER	side effect	27	1427
	ClinTox	clinical toxicity	2	1491

MoleculeNet: A Benchmark for Molecular Machine Learning
Wu et al, 2017, https://arxiv.org/abs/1703.00564

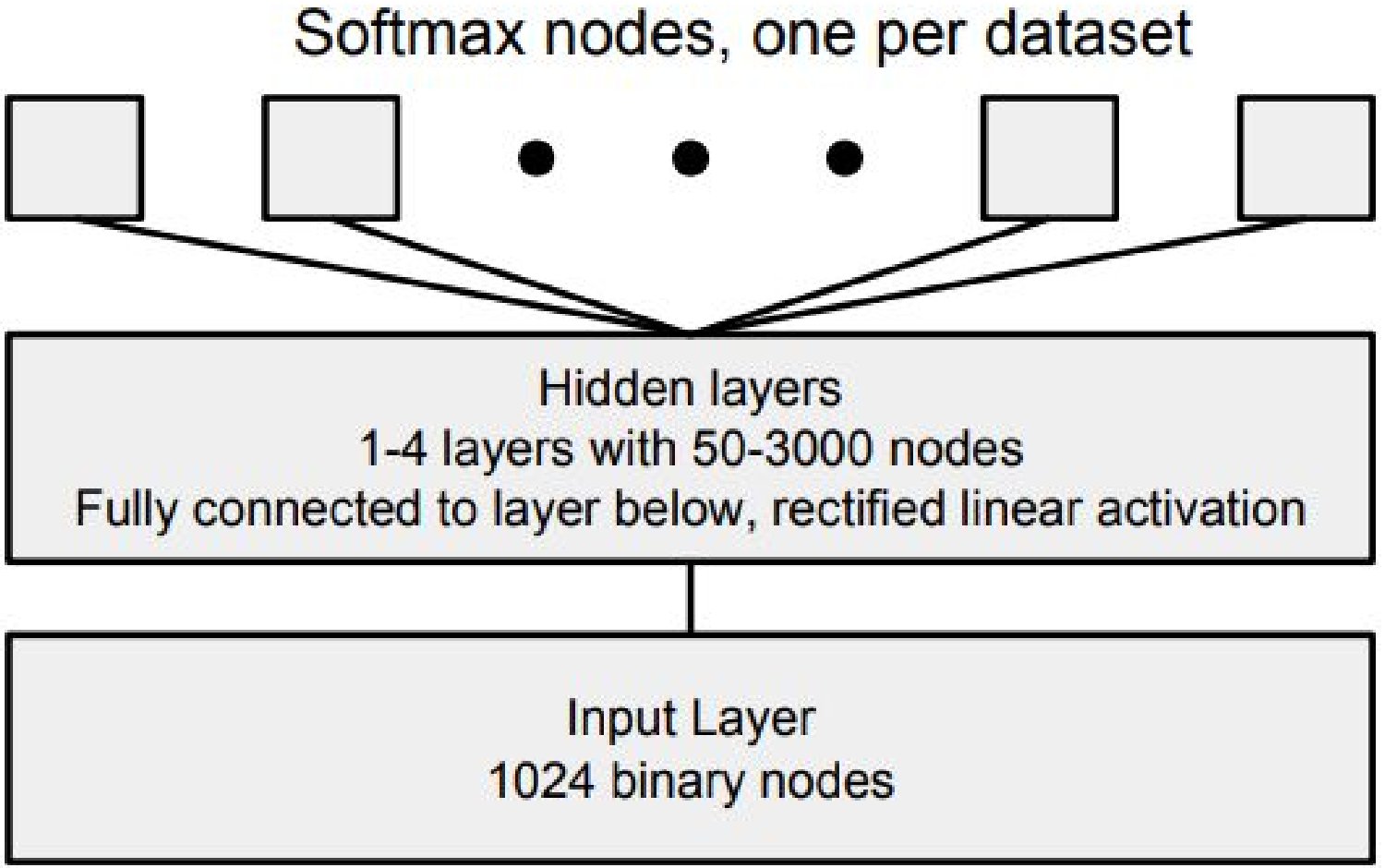
Structured Input Representations

- Molecular graph
  - Atoms as nodes and bonds as vertices
- SMILES
  - Simplified molecular-input line-entry system
  - String representation
  - SMILES is a string obtained by printing the symbol nodes encountered in a depth-first tree traversal of a chemical graph



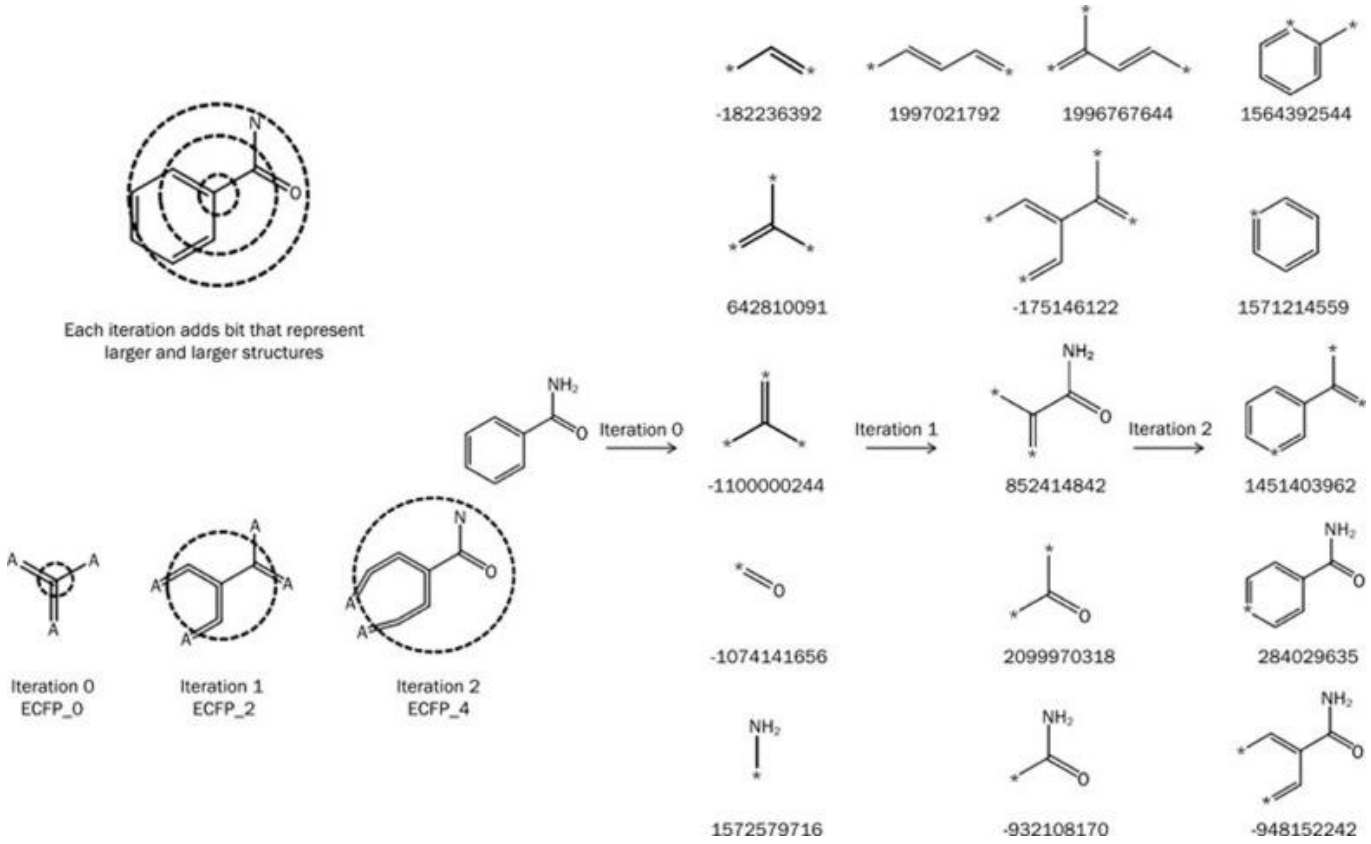
https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

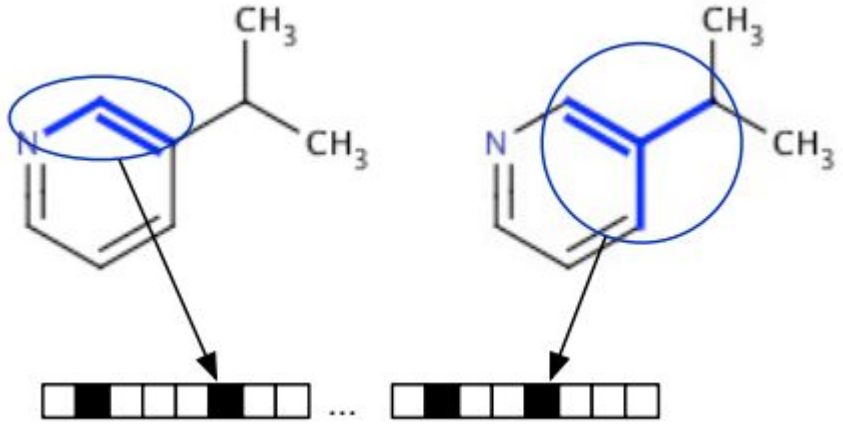
Large Scale Multitask Learning



Massively Multitask Networks for Drug Discovery
Ramsundar et al, 2015, https://arxiv.org/abs/1502.02072

ECFP-4 Fingerprint Features





Novel Bayesian classification models for predicting compounds blocking hERG potassium channels
Liu et al, 2014, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4125710/

Fingerprints in the RDKit
Landrum, 2012,
http://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf

Large Scale Multitask Learning

Table 2. Median 5-fold-average AUCs for various models. For each model, the sign test in the last column estimates the fraction of datasets (excluding the DUD-E group, for reasons discussed in the text) for which that model is superior to the PMTNN (bottom row). We use the Wilson score interval to derive a 95% confidence interval for this fraction. Non-neural network methods were trained using scikit-learn (Pedregosa et al., 2011) implementations and basic hyperparameter optimization. We also include results for a hypothetical "best" single-task model (Max{LR, RF, STNN, PSTNN}) to provide a stronger baseline. Details for our cross-validation and training procedures are given in the Appendix.

Model	PCBA ($n = 128$)	MUV ($n = 17$)	Tox21 ($n = 12$)	Sign Test CI
Logistic Regression (LR)	.801	.752	.738	[.04, .13]
Random Forest (RF)	.800	.774	.790	[.06, .16]
Single-Task Neural Net (STNN)	.795	.732	.714	[.04, .12]
Pyramidal (2000, 100) STNN (PSTNN)	.809	.745	.740	[.06, .16]
Max{LR, RF, STNN, PSTNN}	.824	.781	.790	[.12, .24]
1-Hidden (1200) Layer Multitask Neural Net (MTNN)	.842	.797	.785	[.08, .18]
Pyramidal (2000, 100) Multitask Neural Net (PMTNN)	**.873**	**.841**	**.818**	

Massively Multitask Networks for Drug Discovery
Ramsundar et al, 2015, https://arxiv.org/abs/1502.02072

Large Scale Multitask Learning

- Addition of more tasks and data helps with generalisation of models
- Shared learnt representation seems to be informative
- As expected, additional tasks act as regularisation
  - However, for some tasks, single-task model performance have been observed to be better than multitask
  - This is mainly due to nature of some of the tasks that require bespoke feature learning layers

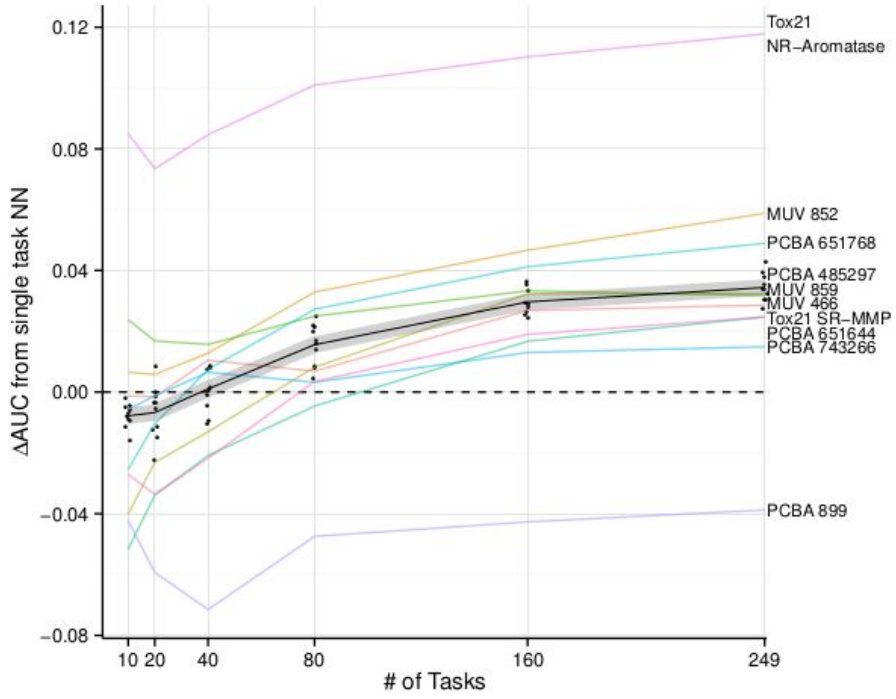
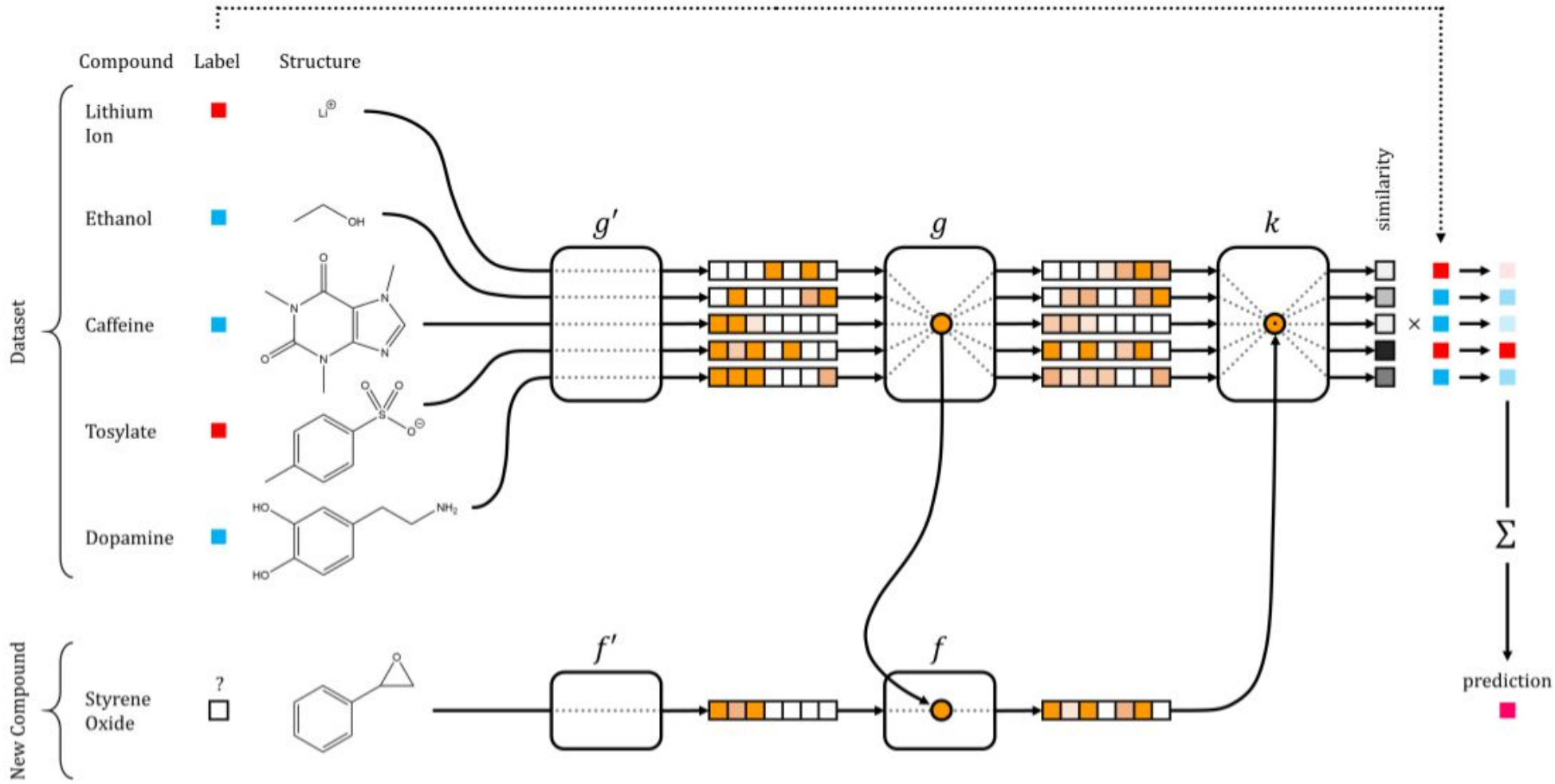


Figure 3. Held-in growth curves. The y-axis shows the change in AUC compared to a single-task neural network with the same architecture (PSTNN). Each colored curve is the multitask improvement for a given held-in dataset. Black dots represent means across the 10 held-in datasets for each experimental run, where additional tasks were randomly selected. The shaded curve is the mean across the 100 combinations of datasets and experimental runs.

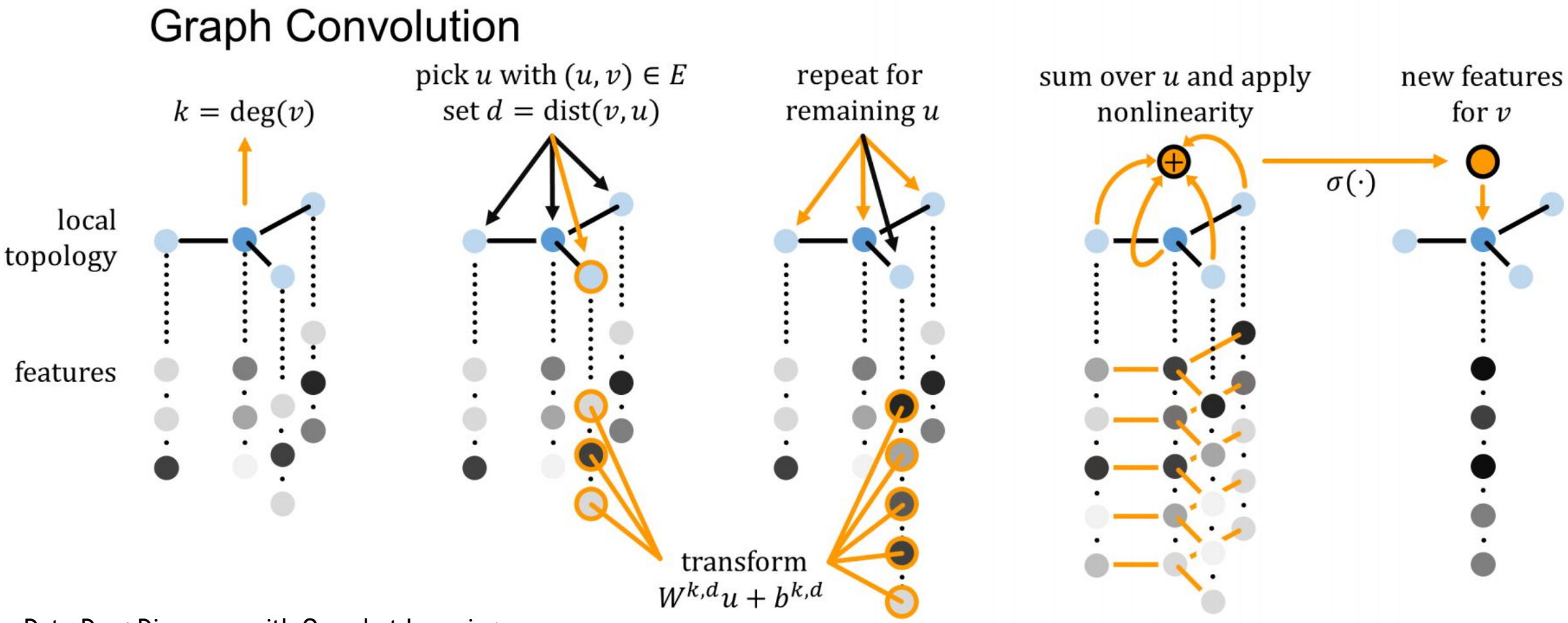
Massively Multitask Networks for Drug Discovery
Ramsundar et al, 2015, https://arxiv.org/abs/1502.02072

One-Shot Learning



Low Data Drug Discovery with One-shot Learning
Altae-Tran et al, 2016, https://arxiv.org/abs/1611.03199

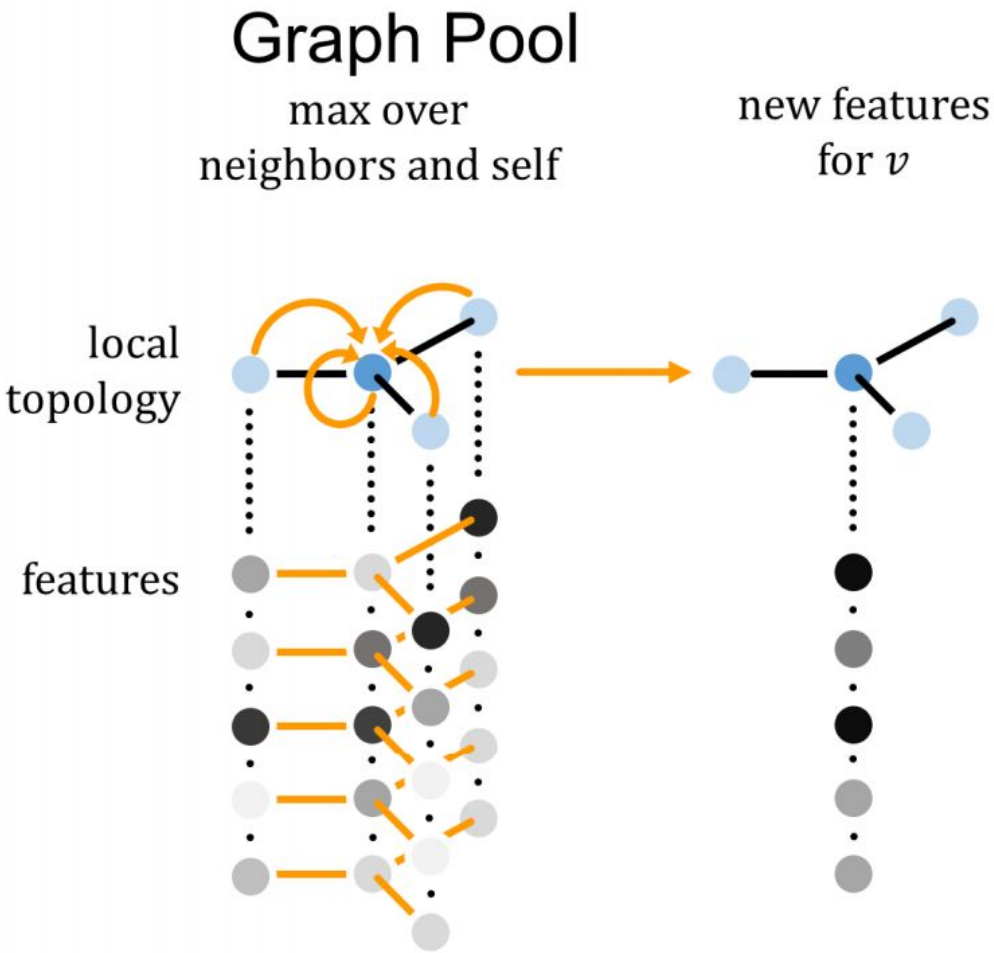
Graph Convolutions

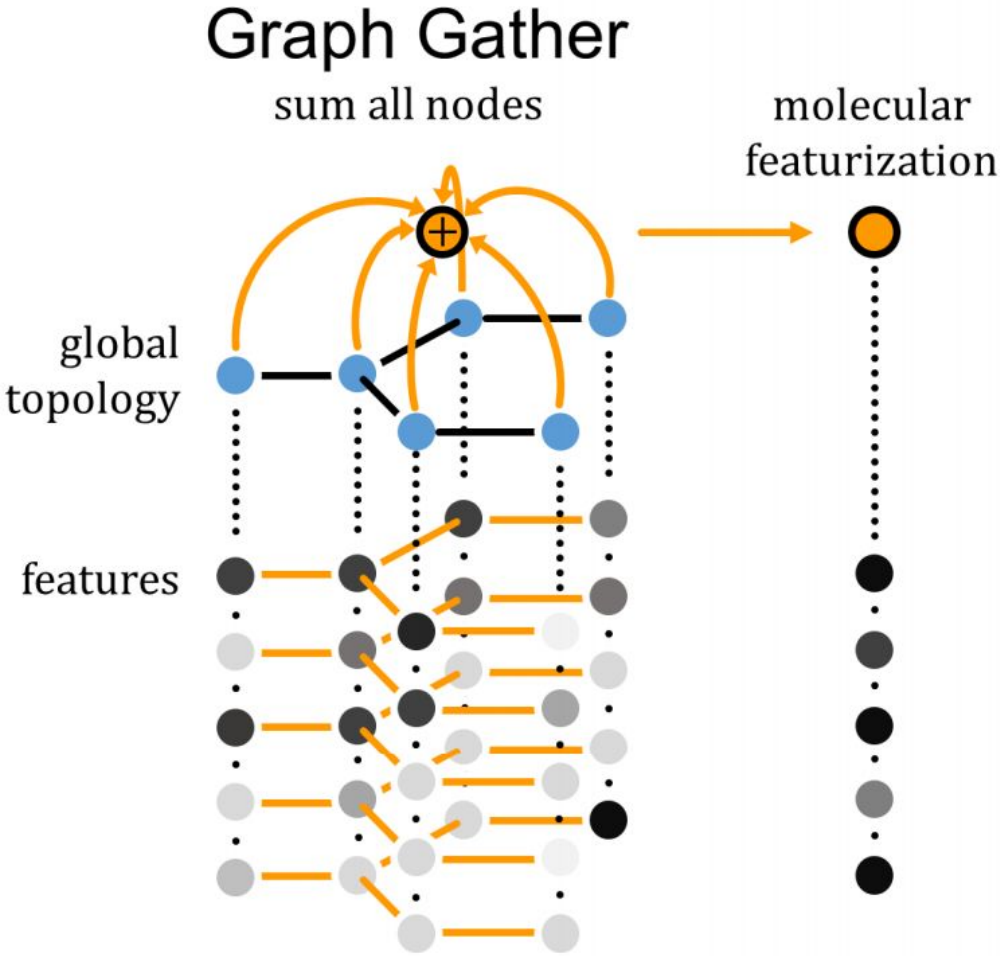


Low Data Drug Discovery with One-shot Learning
Altae-Tran et al, 2016, https://arxiv.org/abs/1611.03199

Convolutional Networks on Graphs for Learning Molecular Fingerprints
Duvenaud et al, 2015, https://arxiv.org/abs/1509.09292

Graph Convolutions





Low Data Drug Discovery with One-shot Learning
Altae-Tran et al, 2016, https://arxiv.org/abs/1611.03199

One-Shot Learning and Graph Convolutions

- One-shot
  - Non-trivial results in a low data setting
- Generalisation
  - Does your runtime task benefit from other tasks?
- Graph convolution methods
  - Learns graph transformation invariant representation
  - Acts as architectural regularisation

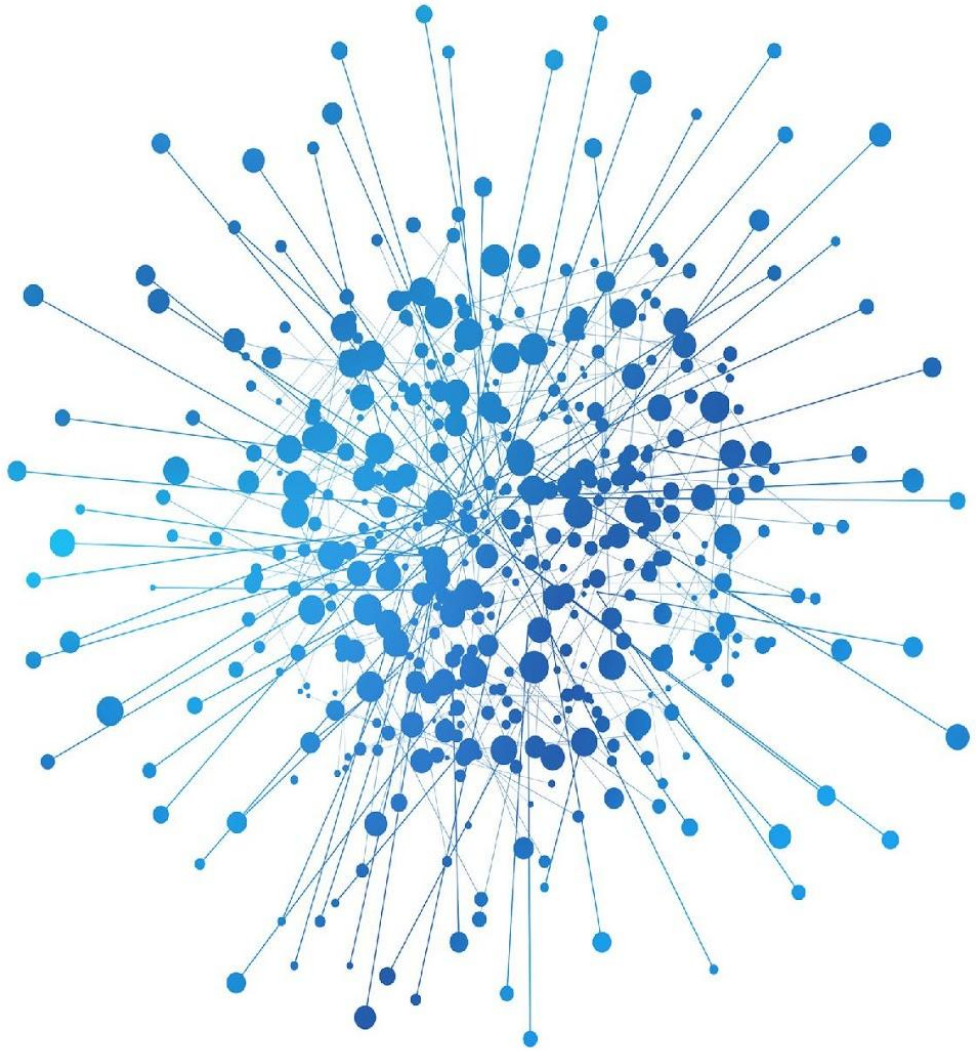
Table 1: Accuracies of models on held-out tasks for Tox21. Numbers reported are median on test-tasks. Numbers for each task are averaged for 20 random choices of support sets.

Tox21	RF (50 trees)	RF (100 trees)	Siamese	AttnLSTM	ResLSTM
10 pos, 10 neg	0.537	0.563	**0.831**	**0.834**	**0.840**
5 pos, 10 neg	0.537	0.579	0.790	**0.820**	**0.837**
1 pos, 10 neg	0.537	0.584	0.710	0.687	**0.757**
1 pos, 5 neg	0.571	0.572	0.689	0.595	**0.815**
1 pos, 1 neg	0.536	0.542	0.668	0.652	**0.784**

Table 3: Accuracies of models on held-out tasks for MUV. Numbers reported are median on test-tasks. Numbers for each task are averaged for 20 random choices of support sets

SIDER	RF (50 trees)	RF (100 trees)	Siamese	AttnLSTM	ResLSTM
10 pos, 10 neg	0.710	**0.741**	0.501	0.683	**0.712**
5 pos, 10 neg	0**.723**	**.751**	0.708	0.674	0.672
1 pos, 10 neg	0.586	**0.624**	0.567	0.583	**0.619**
1 pos, 5 neg	0.561	0.579	0.546	0.565	**0.634**
1 pos, 1 neg	0.558	0.573	0.498	0.501	0.512

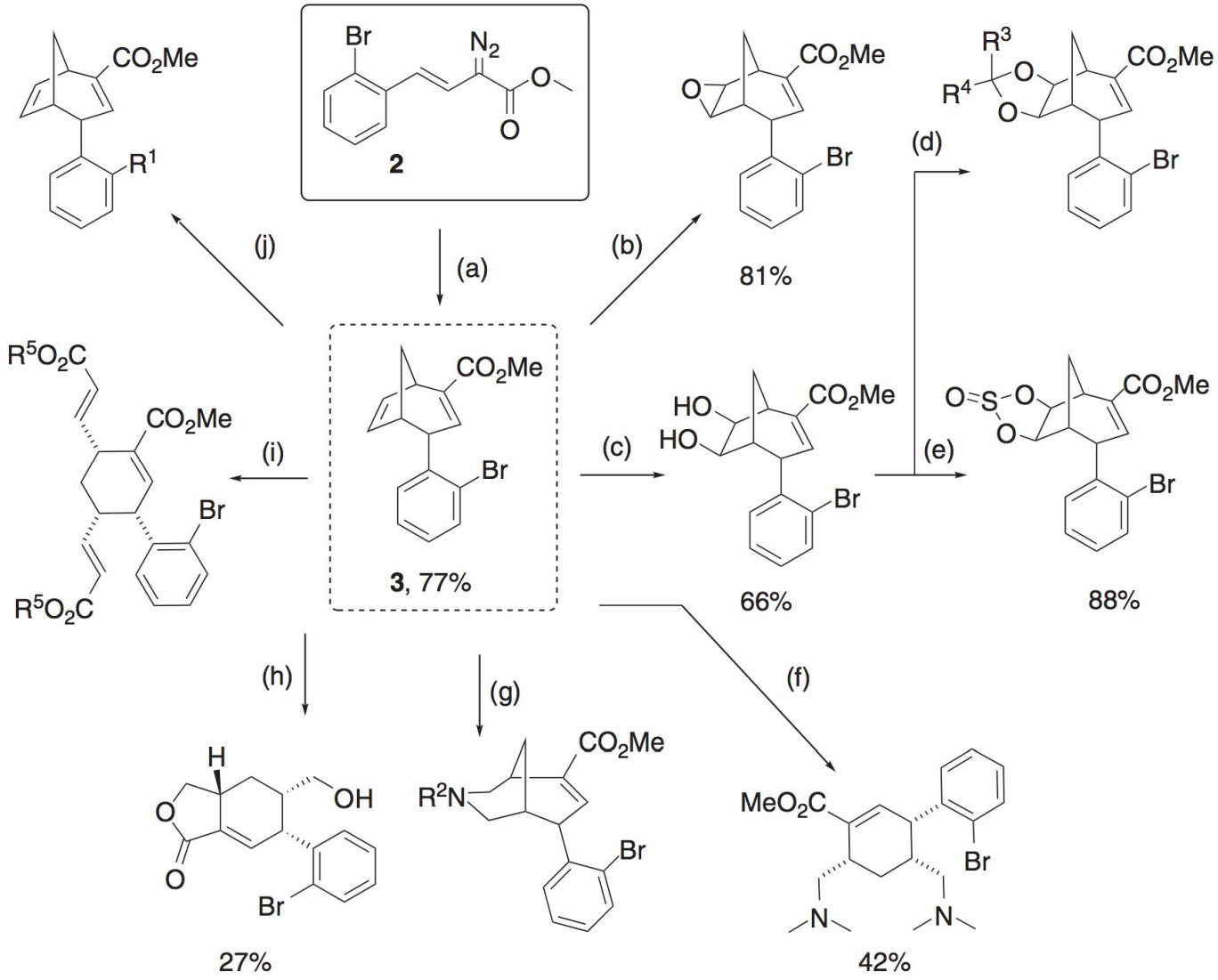
Low Data Drug Discovery with One-shot Learning
Altae-Tran et al, 2016, https://arxiv.org/abs/1611.03199



Molecule Generation

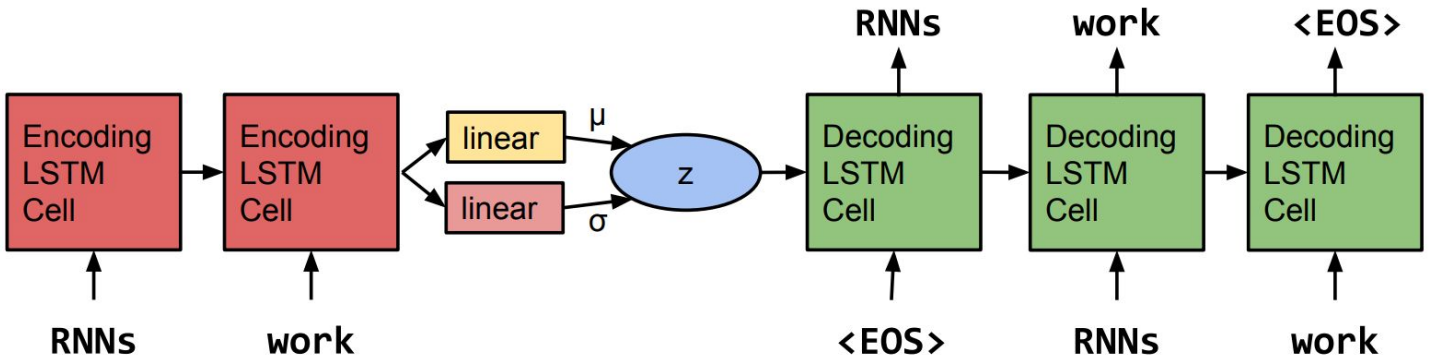
Diversity and Goal Oriented Molecular Generation Optimisation

- Molecular space is discrete
  - There are enumerations of drug-like small molecules (e.g. GDB-13 ~ 1B, GDB-17 ~ 166B)
  - No gradients over the input space
  - Evolutionary algorithms (e.g. De Novo Design at the Edge of Chaos, Schneider and Schneider, 2016)
- Given a set of goals, evaluating all molecules is infeasible
- Learn generative algorithms or search policies to create novel and optimal molecules



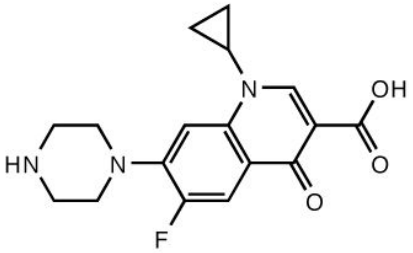
Diversity-Oriented Synthesis: Developing New Chemical Tools to Probe and Modulate Biological Systems
Galloway et al, 2014, http://www-spring.ch.cam.ac.uk/publications/pdf/2014_DOS_379.pdf

Variational AutoEncoder Models and Text Generation

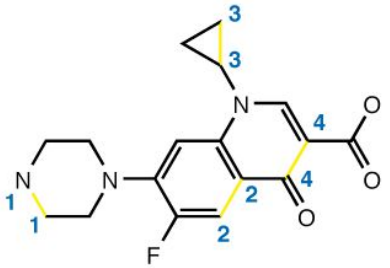


“ i want to talk to you . ”
“i want to be with you . ”
“i do n’t want to be with you . ”
i do n’t want to be with you .
she did n’t want to be with him .
he was silent for a long moment .
he was silent for a moment .
it was quiet for a moment .
it was dark and cold .
there was a pause .
it was my turn .

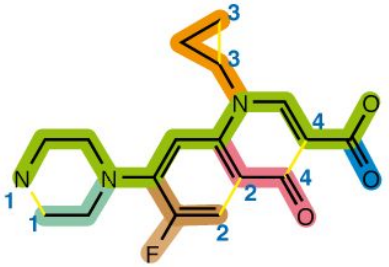
A



B



C

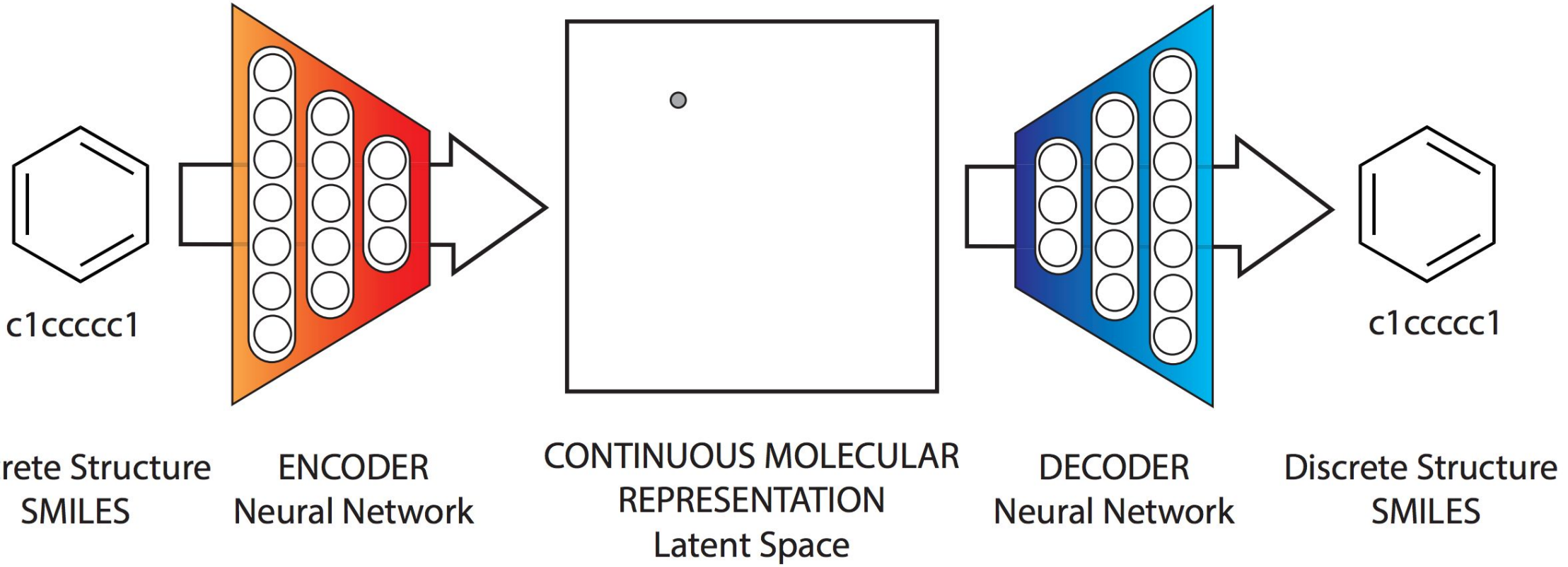


D

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

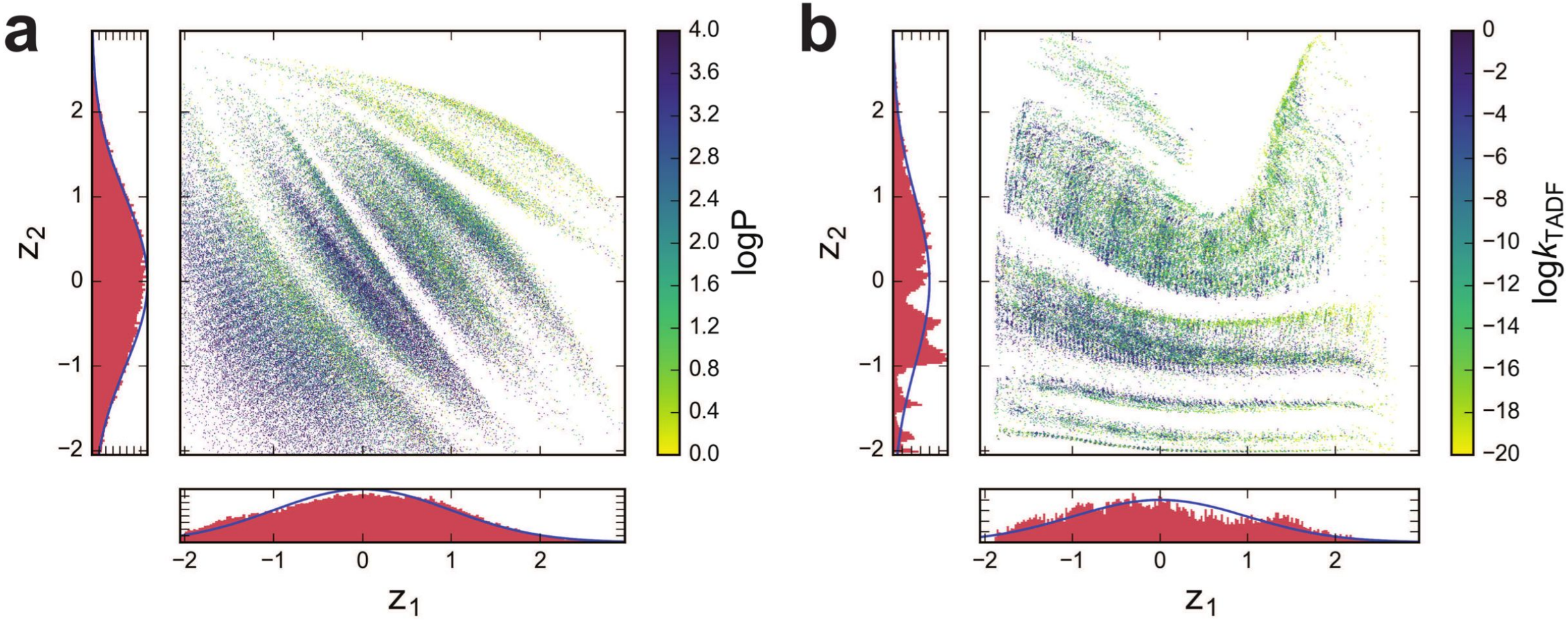
Generating Sentences from a Continuous Space
Bowman et al, 2016, http://www.aclweb.org/anthology/K16-1002

VAE for Compound Generation



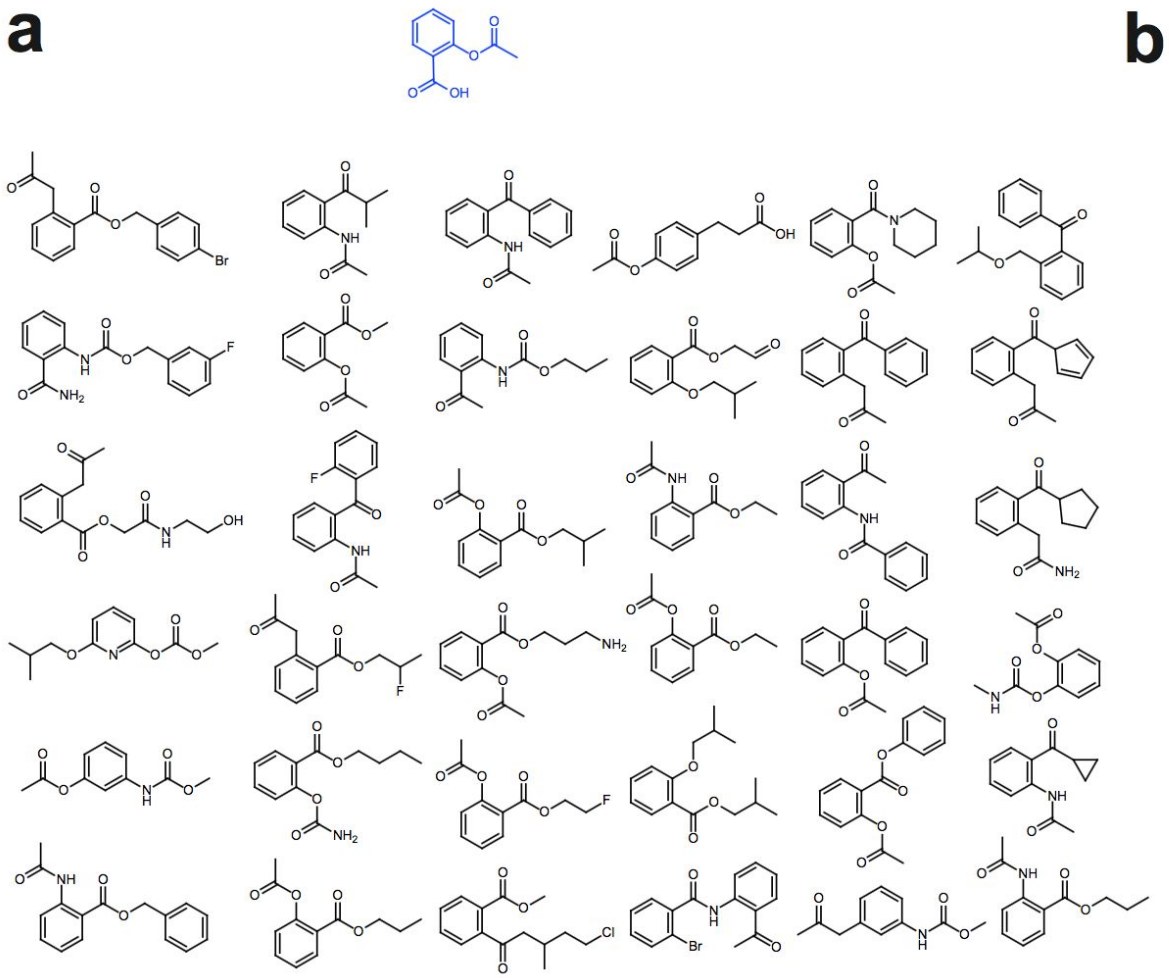
Automatic chemical design using a data-driven continuous representation of molecules
Gomez-Bombarelli et al, 2016, https://arxiv.org/abs/1610.02415

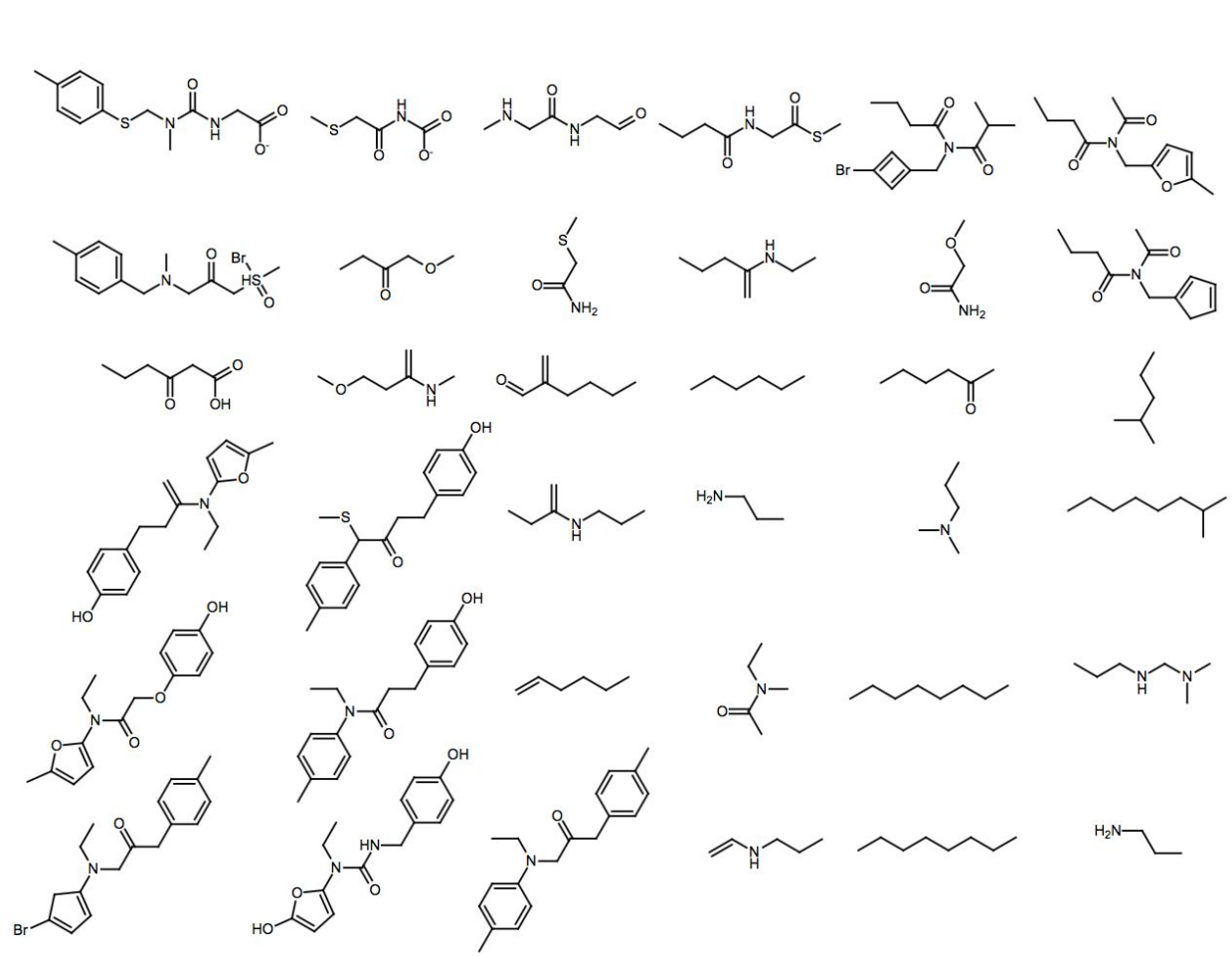
Latent Space and Properties



Automatic chemical design using a data-driven continuous representation of molecules
Gomez-Bombarelli et al, 2016, https://arxiv.org/abs/1610.02415

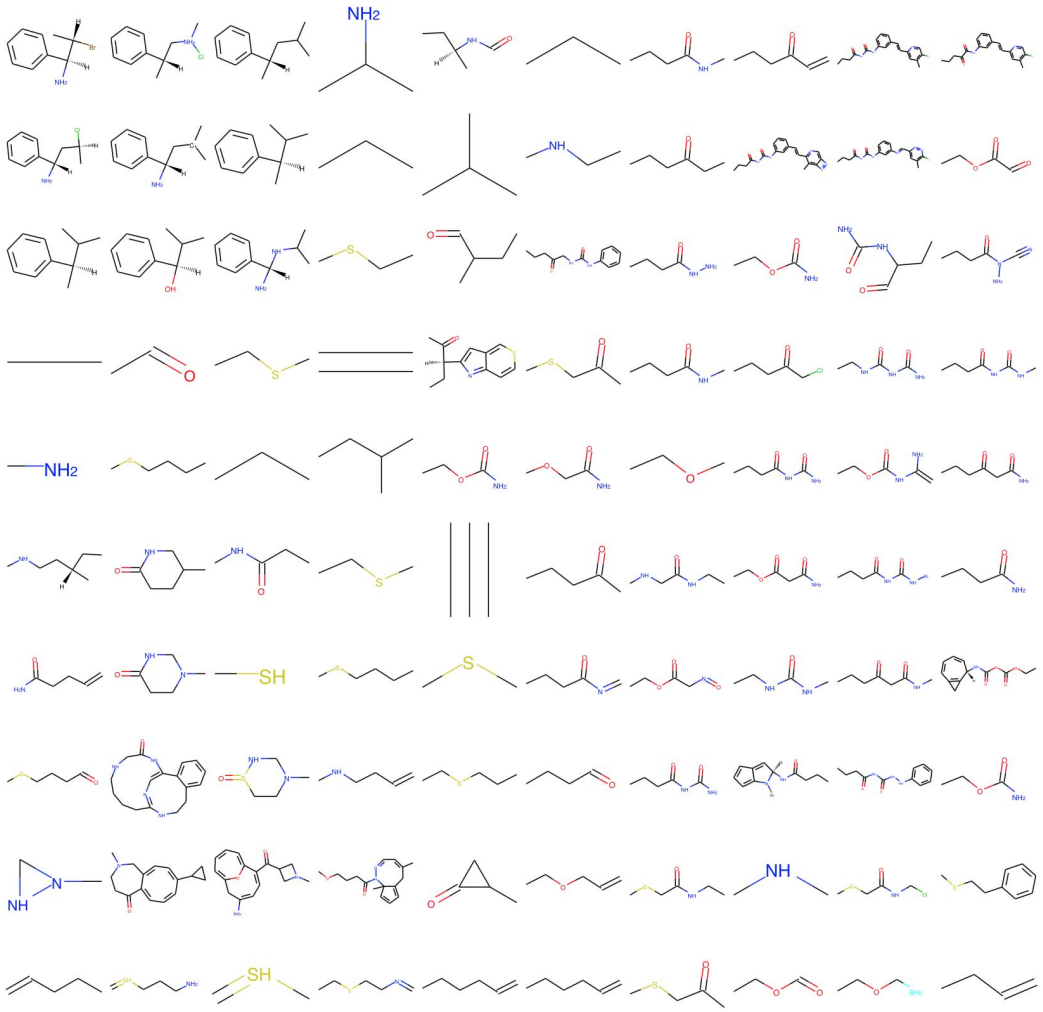
Latent Space Exploration





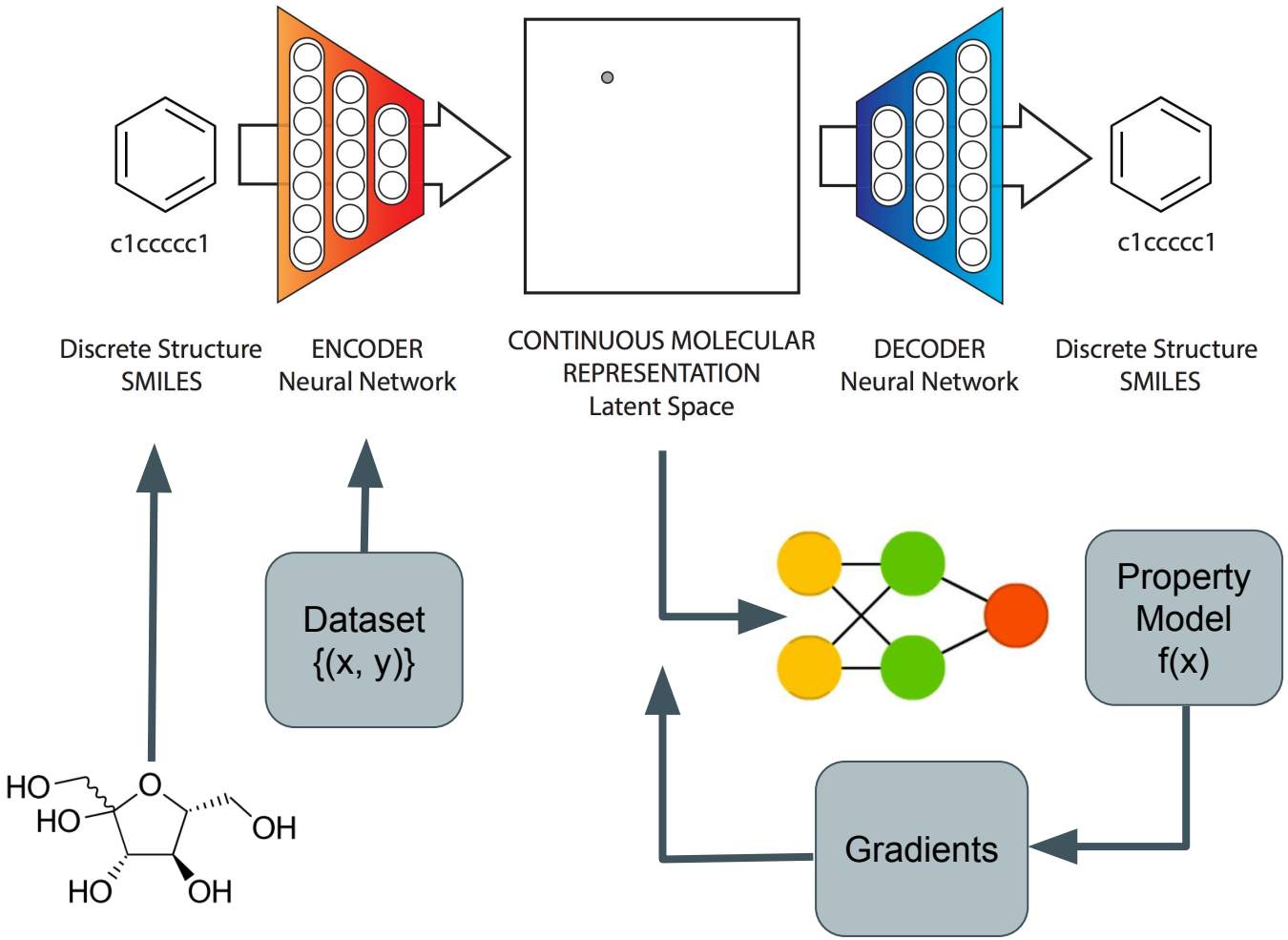
Automatic chemical design using a data-driven continuous representation of molecules
Gomez-Bombarelli et al, 2016, https://arxiv.org/abs/1610.02415

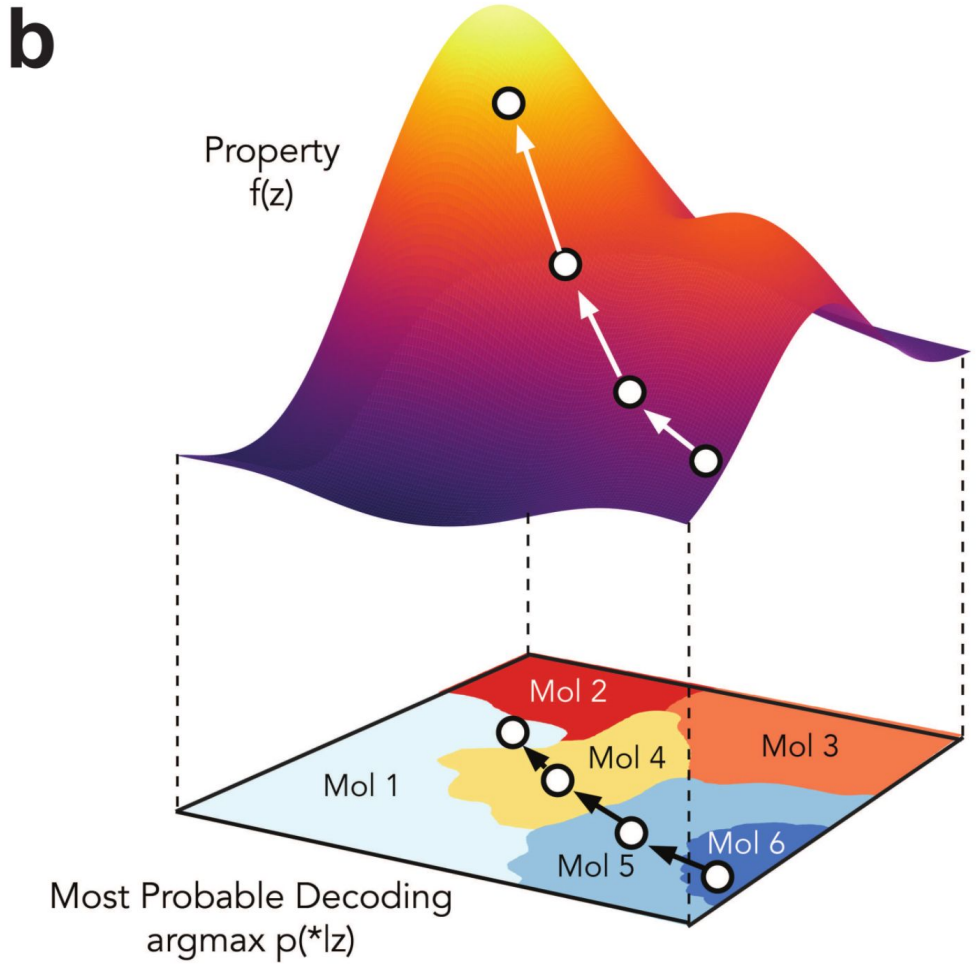
Latent Space Exploration



Automatic chemical design using a data-driven continuous representation of molecules
Gomez-Bombarelli et al, 2016, https://arxiv.org/abs/1610.02415

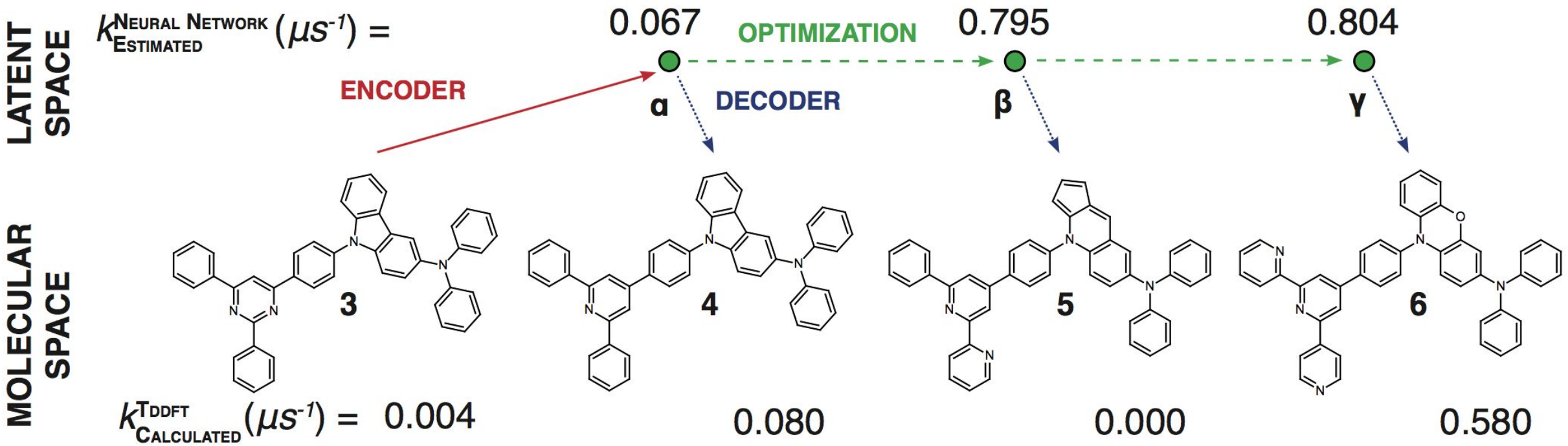
Gradient-based Optimisation





Automatic chemical design using a data-driven continuous representation of molecules
Gomez-Bombarelli et al, 2016, https://arxiv.org/abs/1610.02415

Gradient-based Optimisation



Automatic chemical design using a data-driven continuous representation of molecules
Gomez-Bombarelli et al, 2016, https://arxiv.org/abs/1610.02415

Grammar Variational AutoEncoders: Encoder

- SMILES are not natural language
  - It has a context free grammar
  - SMILES can be represented as a parse tree
- Instead of generating character by character, generate production rules

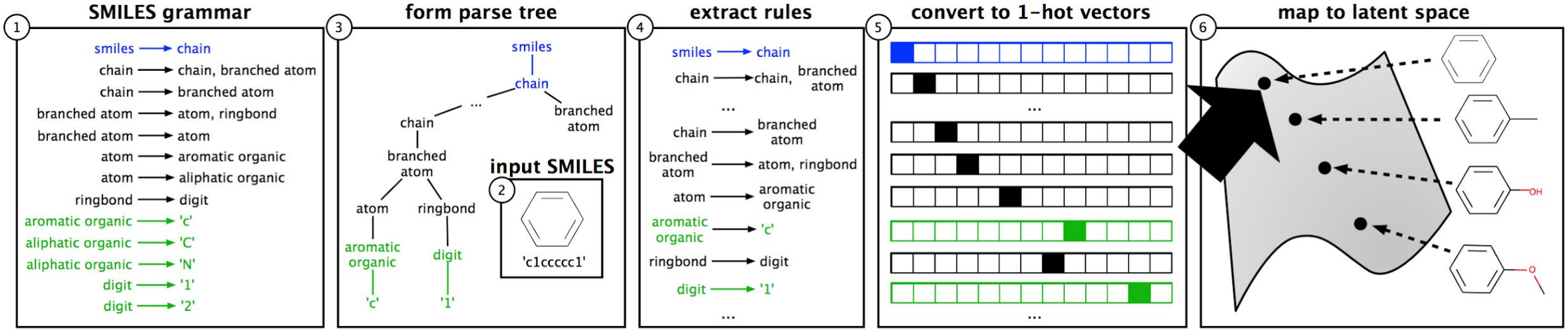


Figure 1. The encoder of the GVAE. We denote the start rule in blue and all rules that decode to terminal in green. See text for details.

Grammar Variational Autoencoder
Kusner et al, 2017, https://arxiv.org/abs/1703.01925

Grammar Variational AutoEncoders: Decoder

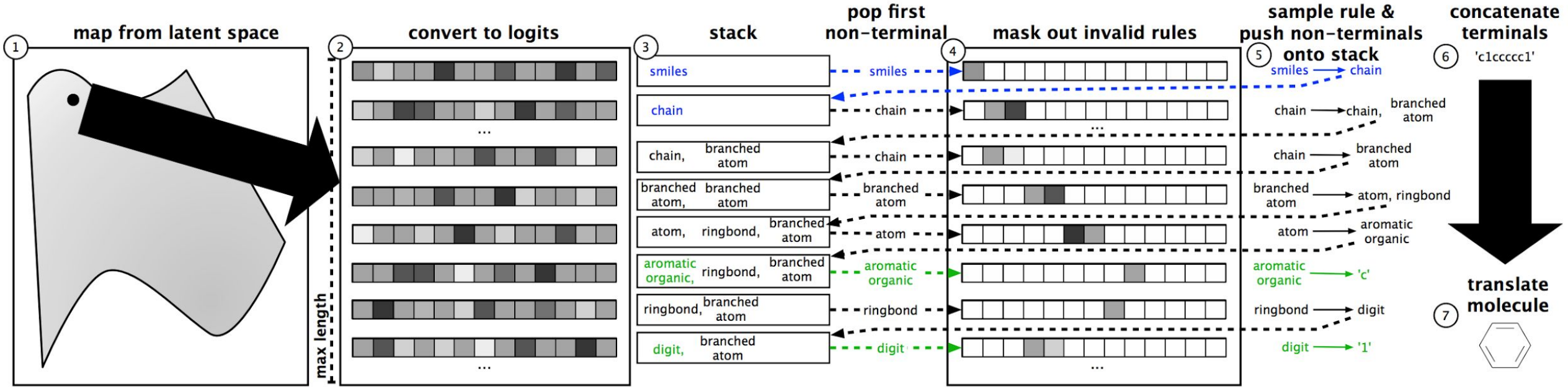
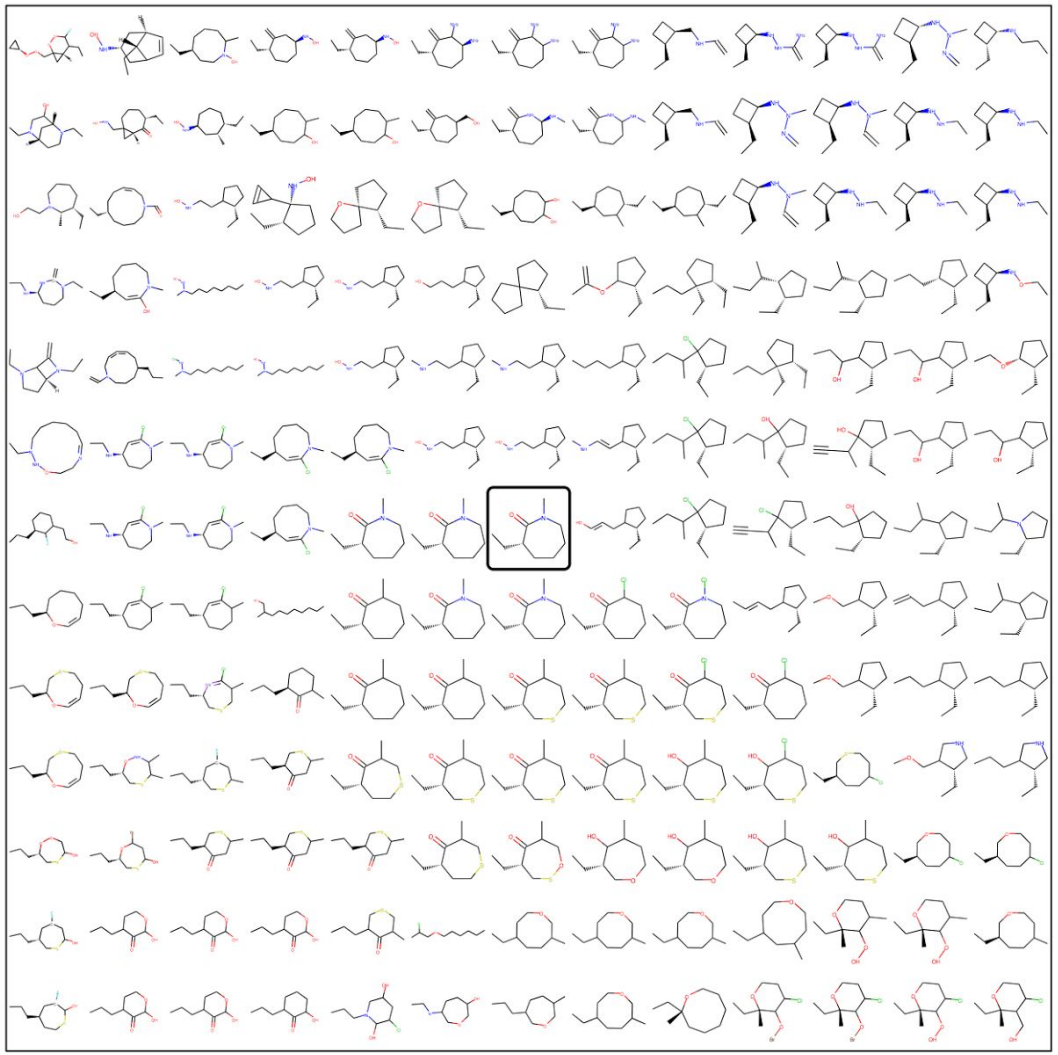
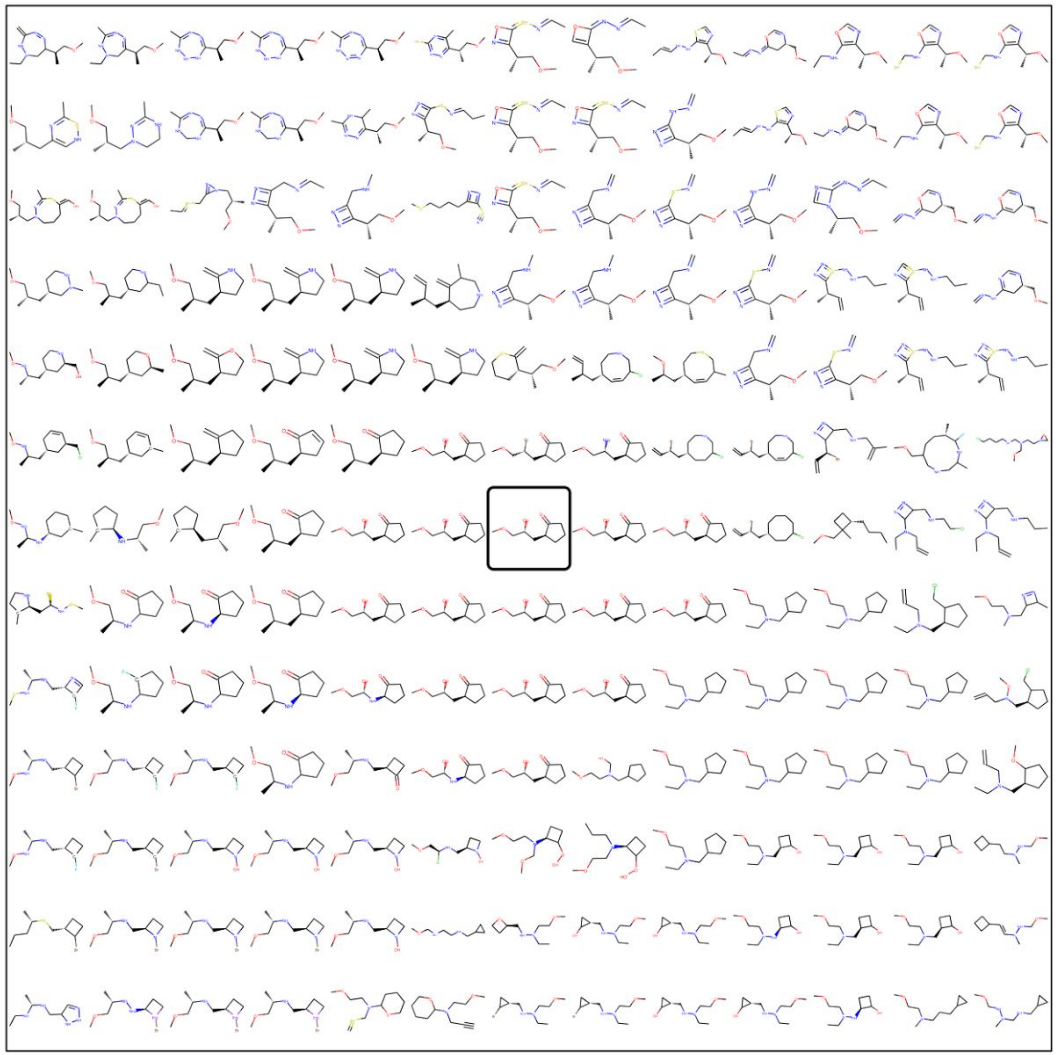


Figure 2. The decoder of the GVAE. See text for details.

Grammar Variational Autoencoder
Kusner et al, 2017, https://arxiv.org/abs/1703.01925

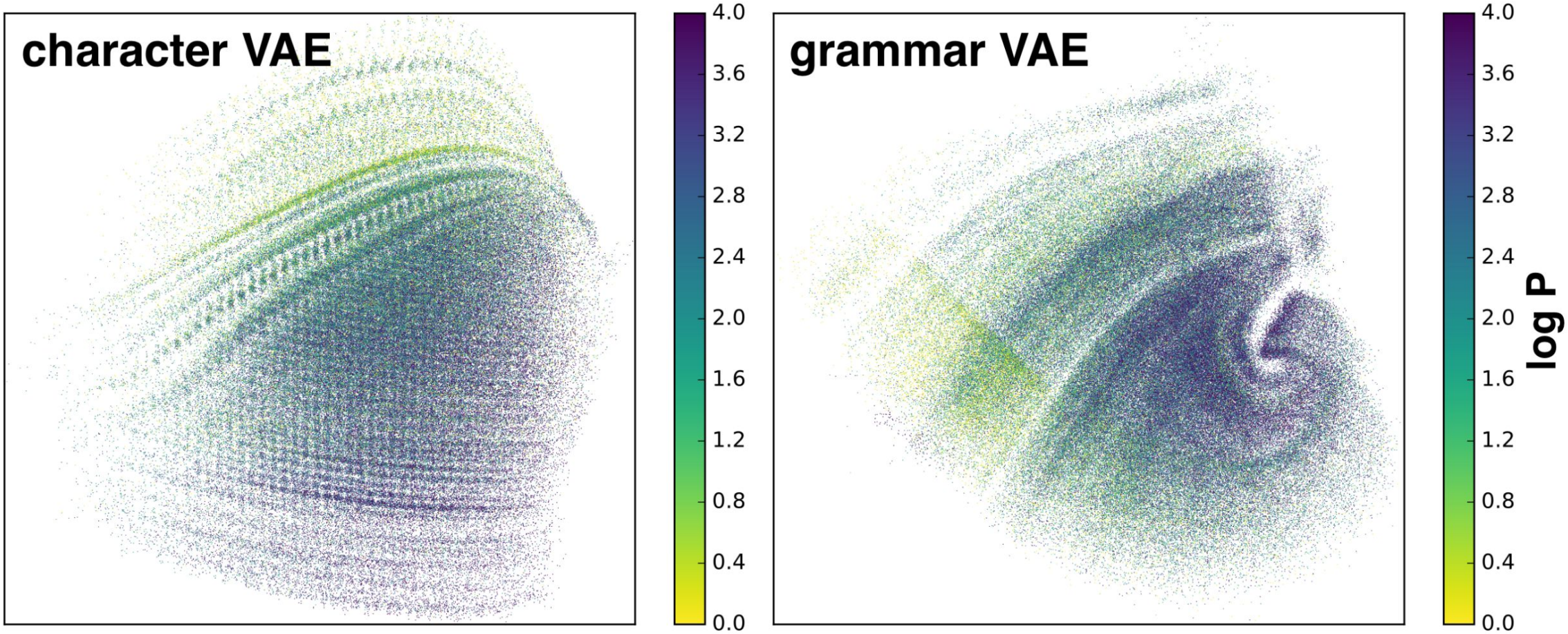
GVAEs: Latent Space





Grammar Variational Autoencoder
Kusner et al, 2017, https://arxiv.org/abs/1703.01925

GVAEs: Latent Space and Properties



Grammar Variational Autoencoder
Kusner et al, 2017, https://arxiv.org/abs/1703.01925

GVAEs: Bayesian Optimisation and Valid Decoding

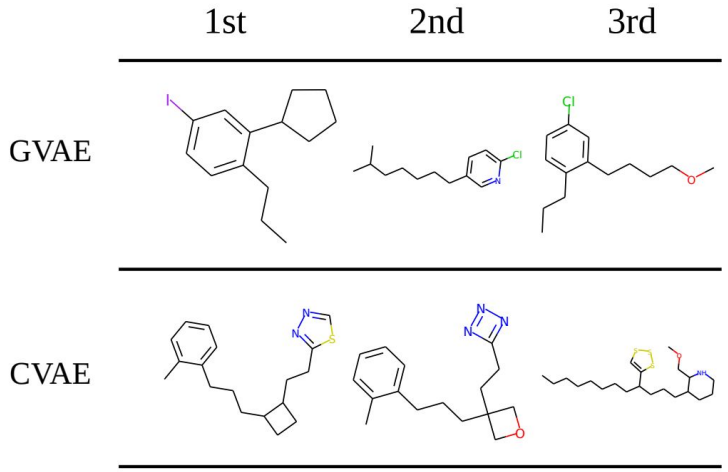


Figure 5. Plot of best molecules found by each method.

Table 4. Best molecules found by each method

Method	#	SMILE	Score
GVAE	1	CCCc1ccc(I)cc1C1CCC-c1	**2.94**
	2	CC(C)CCCCCc1ccc(Cl)nc1	**2.89**
	3	CCCc1ccc(Cl)cc1CCCCOC	**2.80**
CVAE	1	Cc1ccccc1CCCC1CCC1CCc1nncs1	1.98
	2	Cc1ccccc1CCCC1(COC1)CCc1nnn1	1.42
	3	CCCCCCCCCC(CCCC212CCCnC1COC)c122csss1	1.19

Table 2. Results finding best expression and molecule

Problem	Method	Frac. valid	Avg. score
Expressions	GVAE	**0.99±0.01**	**3.47 ±0.24**
	CVAE	0.86±0.06	4.75±0.25
Molecules	GVAE	**0.31±0.07**	**-9.57 ±1.77**
	CVAE	0.17±0.05	-54.66±2.66

Grammar Variational Autoencoder
Kusner et al, 2017, https://arxiv.org/abs/1703.01925

RL for Sequence Generation

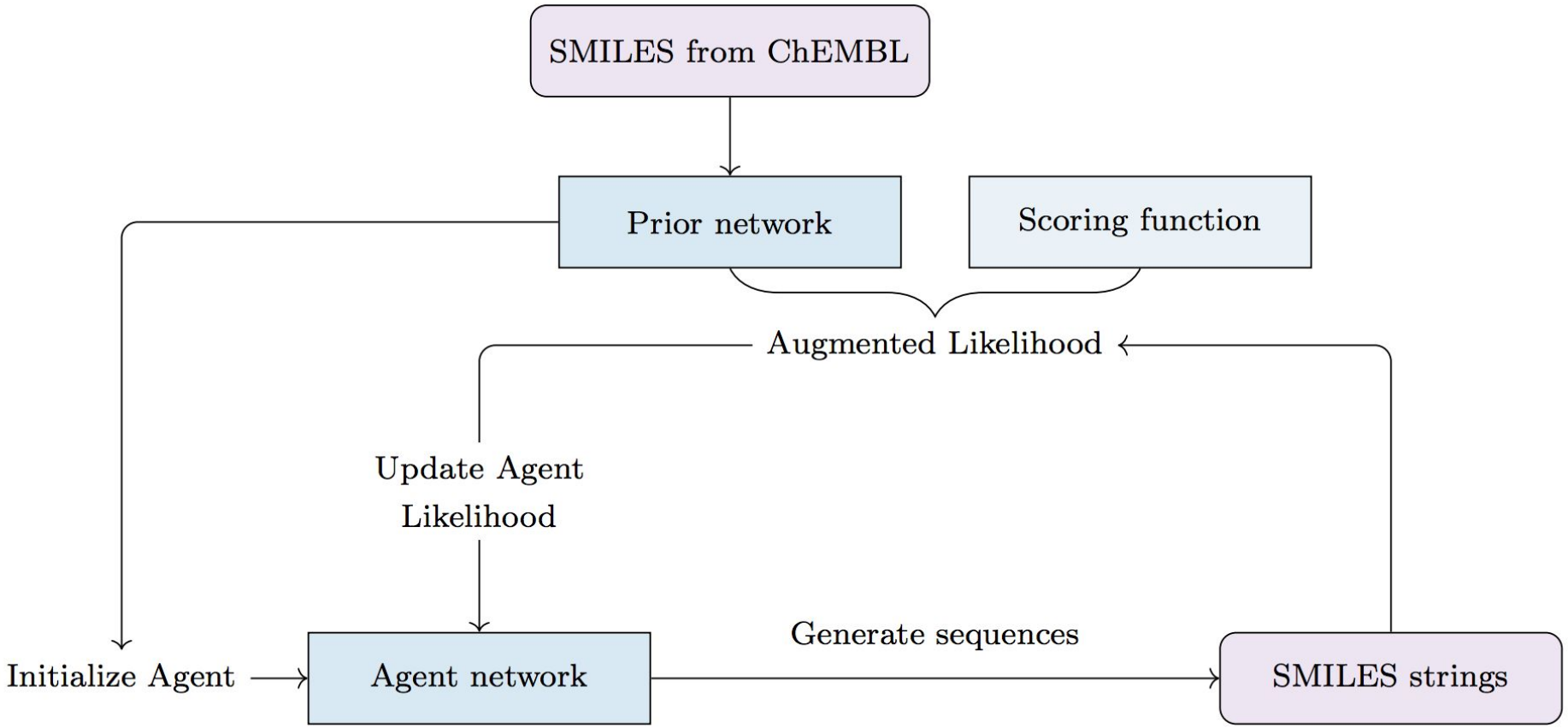


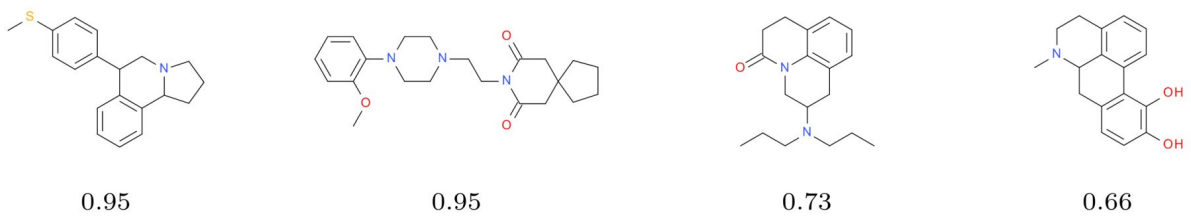
Figure 4 The Agent. Illustration of how the model is constructed. Starting from a Prior network trained on ChEMBL, the Agent is trained using the augmented likelihood of the SMILES generated.

Molecular De-Novo Design through Deep Reinforcement Learning
Olivecrona et al, 2017, https://arxiv.org/abs/1704.07555

RL for Sequence Generation

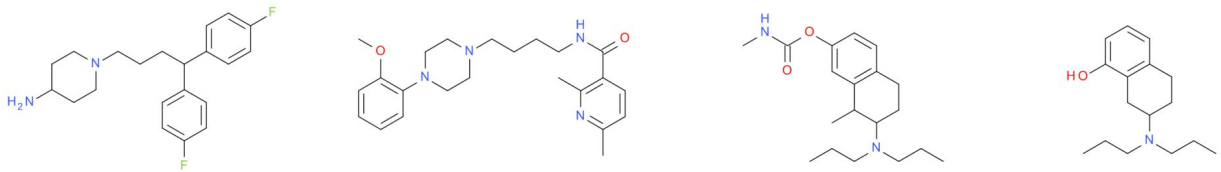
Model	Prior	Agent	Prior†	Agent†
Fraction predicted actives	0.03	0.97	0.02	0.96
Fraction similar to train active	0.02	0.79	0.02	0.75
Fraction similar to test active	0.01	0.46	0.01	0.38
Fraction of test actives recovered	0.00	0.13	0.00	0.07
Probability of generating a test set active ($\times 10^{-3}$)	0.17	40.2	0.05	15.0

†DRD2 actives witheld from the training of the Prior



P_{active} 0.95 0.95 0.73 0.66

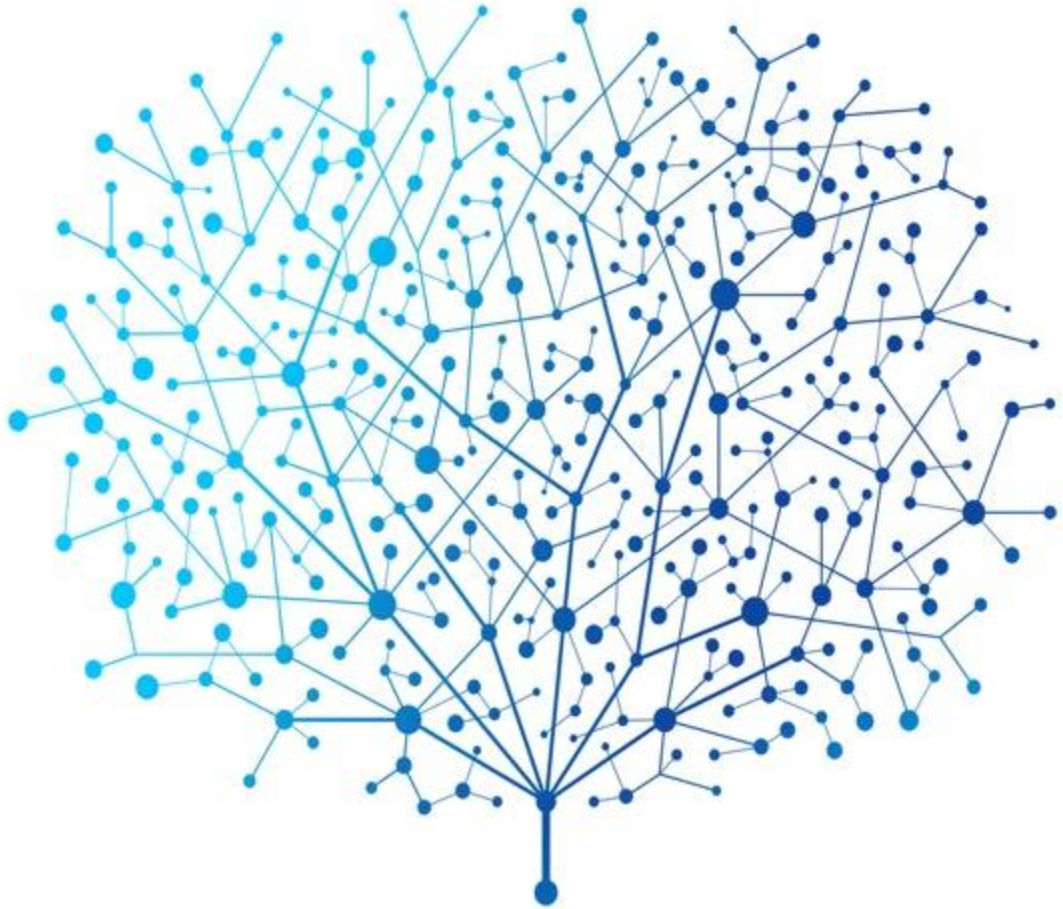
Recovered test actives



P_{active} 1.00 0.99 0.98 1.00

Randomly selected

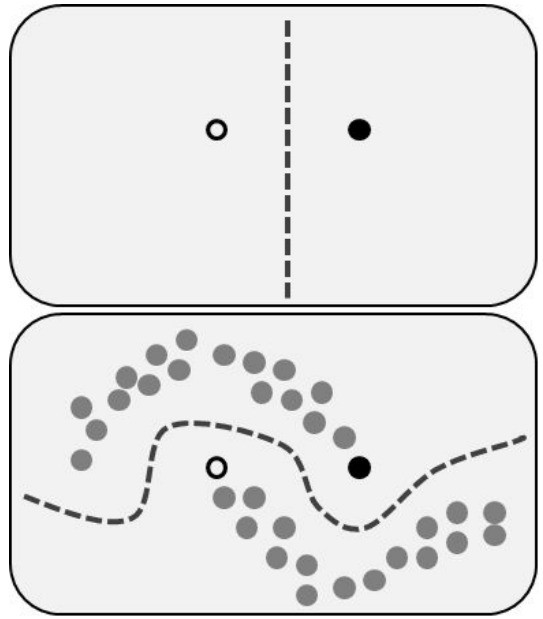
Molecular De-Novo Design through Deep Reinforcement Learning
Olivecrona et al, 2017, https://arxiv.org/abs/1704.07555

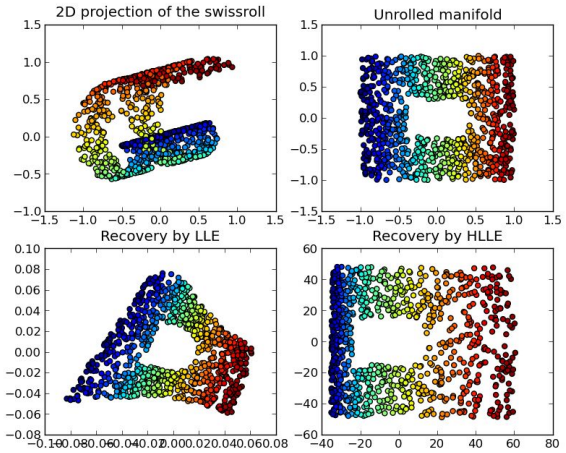


Discussion and Future Directions

Learning with Auxiliary Tasks

- Lack of enough labelled data
- Availability of very large scale unlabelled data
  - GDB-13 and GDB-17
- Regularised learning with auxiliary tasks to learn better representations
  - Pretraining and fine-tuning
  - Semi-supervised learning
  - Multitask learning
  - Transfer learning
  - One-shot learning
  - GANs
- Automatic discovery of adaptive auxiliary tasks
  - How to automatically learn auxiliary problems, architectures, and, loss functions?
- How to efficiently work with very large scaled data



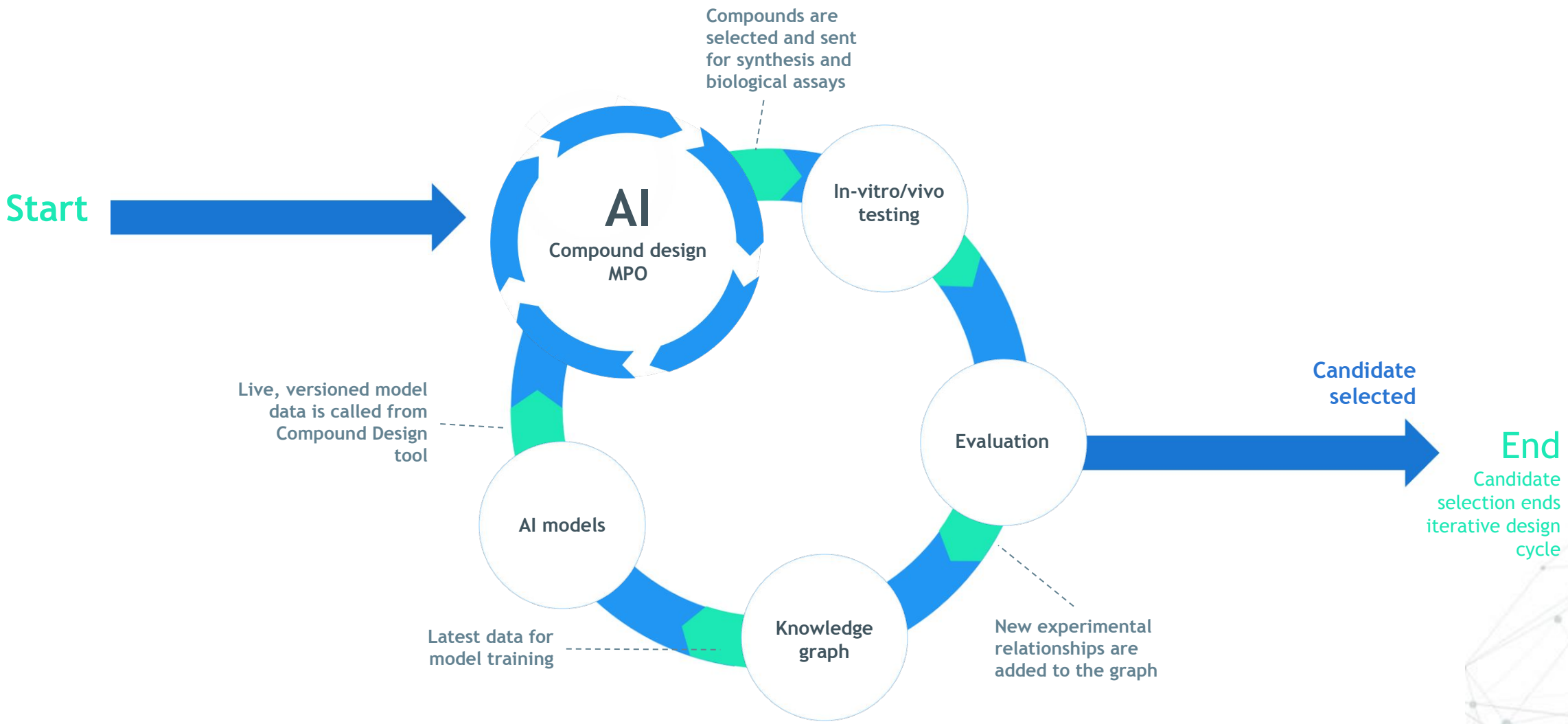


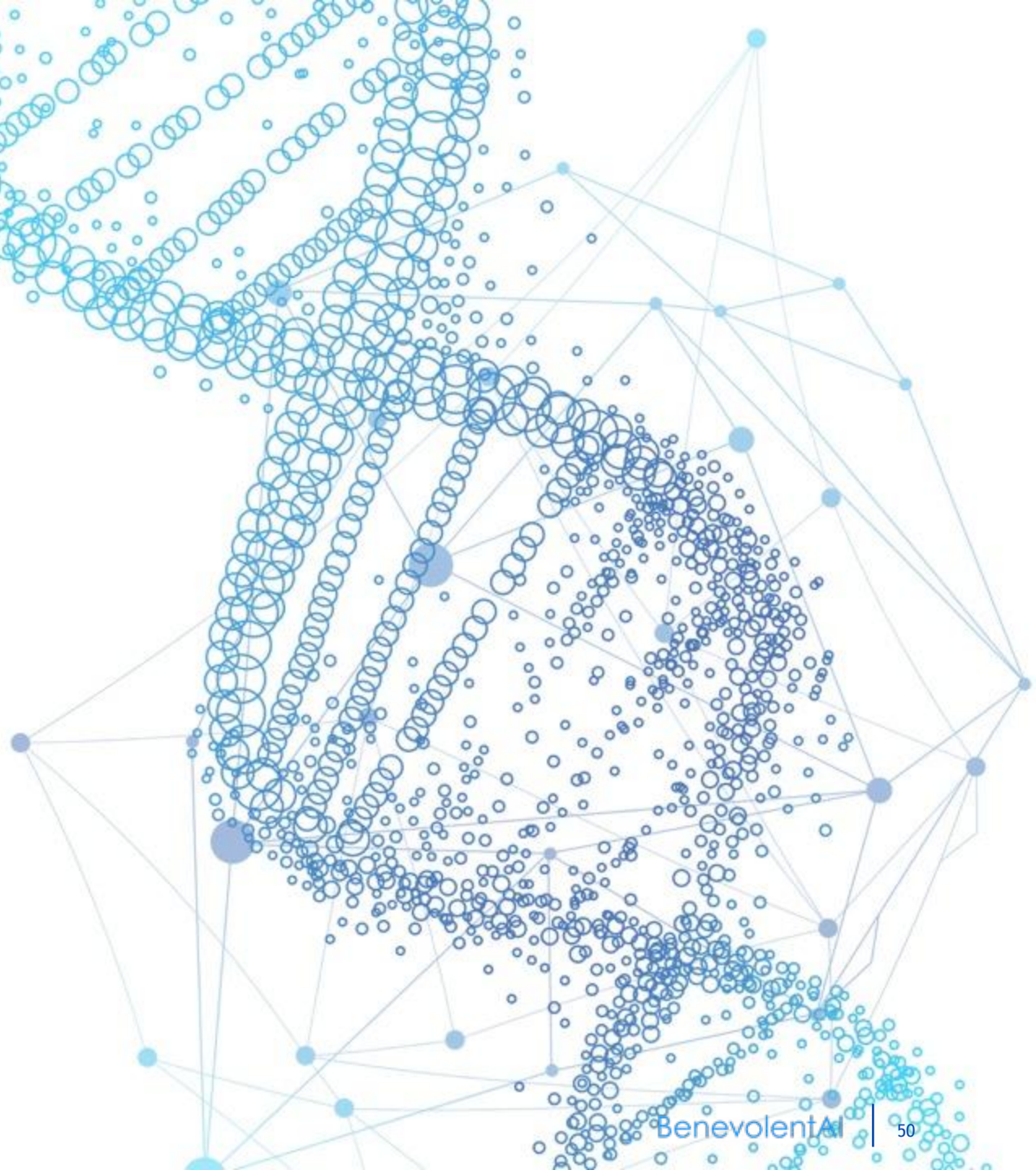
https://en.wikipedia.org/wiki/Semi-supervised_learning

Generative Optimisation

- Learning reward functions
  - Many property prediction models
  - Molecular simulations
  - Dealing with uncertainty
- Learning representations
  - Discriminative representations learnt by property prediction models might not be suitable
  - How to automatically learn shared and specialised representation
- Generative models
  - Learning search and optimisation policies
  - Multi-objective optimisation
  - Learning to generate valid, diverse, and, novel compounds
  - Learning intrinsic rewards
  - Interpretable generative models
- Benchmarking
  - Methods for evaluating generative models
  - Learnt reward functions are proxy to real-world
  - How to assess quality of latent spaces?
  - Similarity based perturbation and discovery of known (hidden) sample

Interactive Learning with Users and Real-World Test Labs





Thanks!

Get in touch if interested,

we are hiring!

@amirsaffari

amir.saffari@benevolent.ai