Artificial Intelligence Driven Drug Discovery

Michelle Gill, PhD

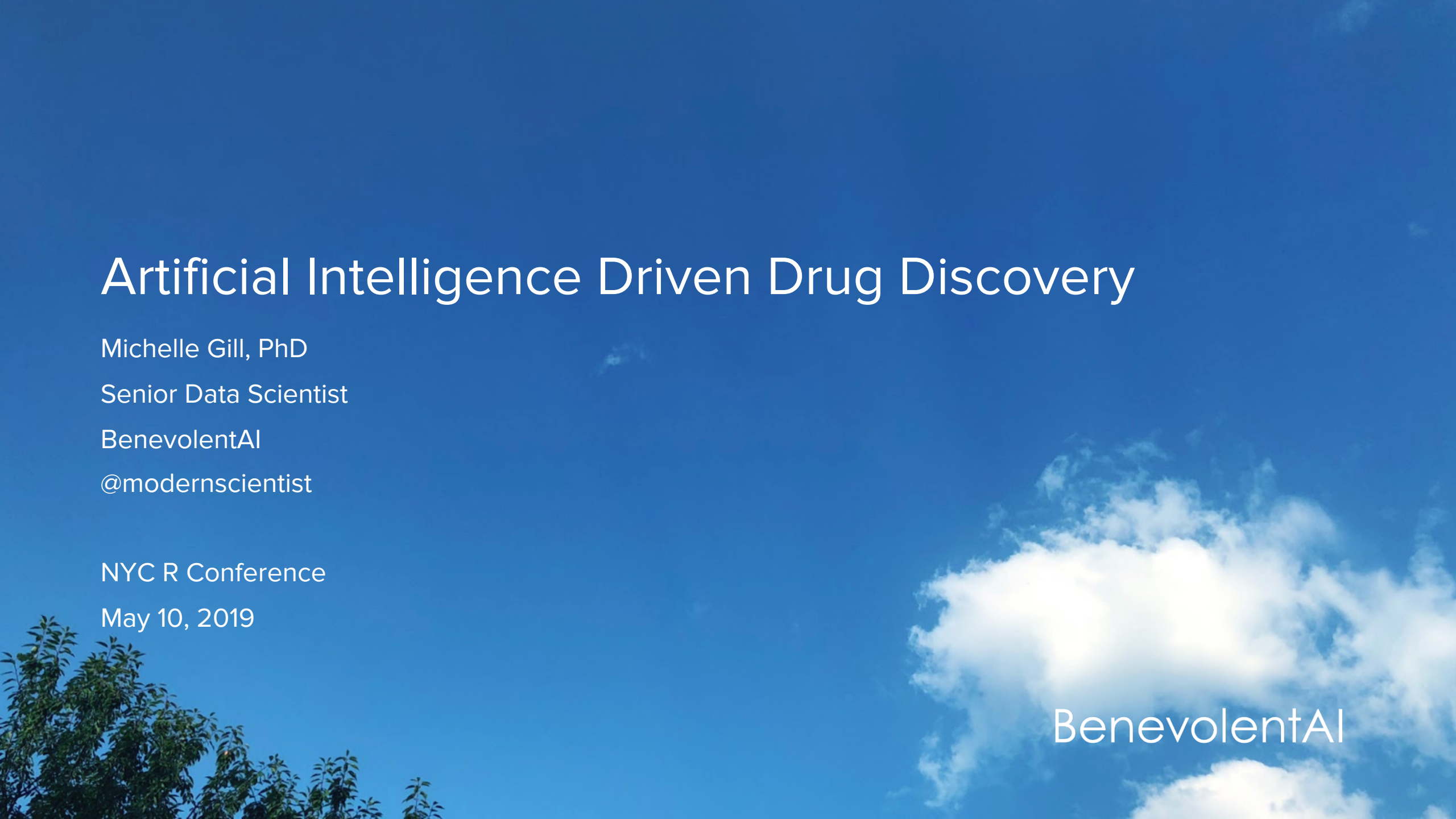
Senior Data Scientist

BenevolentAI

@modernscientist

NYC R Conference

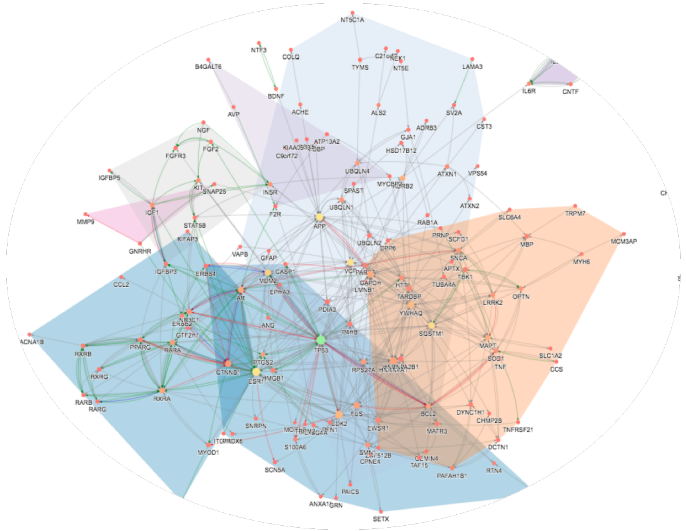
May 10, 2019

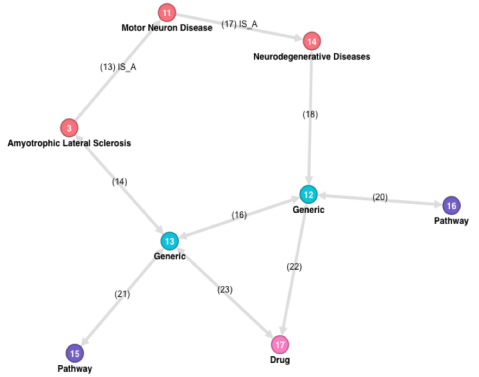


BenevolentAI

Because it matters.

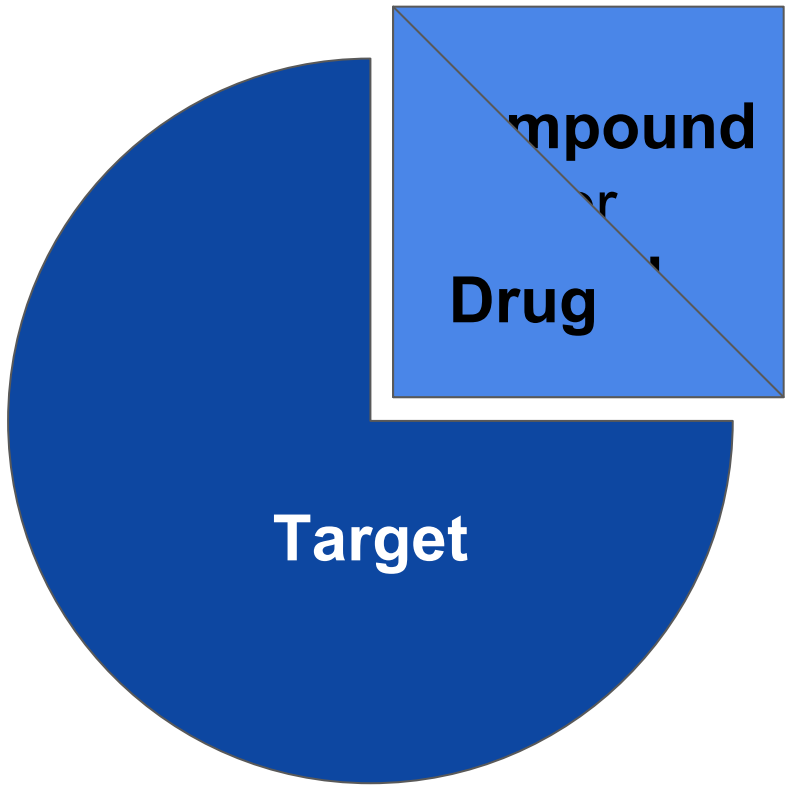
- **>200 life scientists, AI scientists, informaticians** work side-by-side
- Technology pipeline is **validated by scientific experimentation**
- Only AI company with expertise from **early to late stage drug development process**



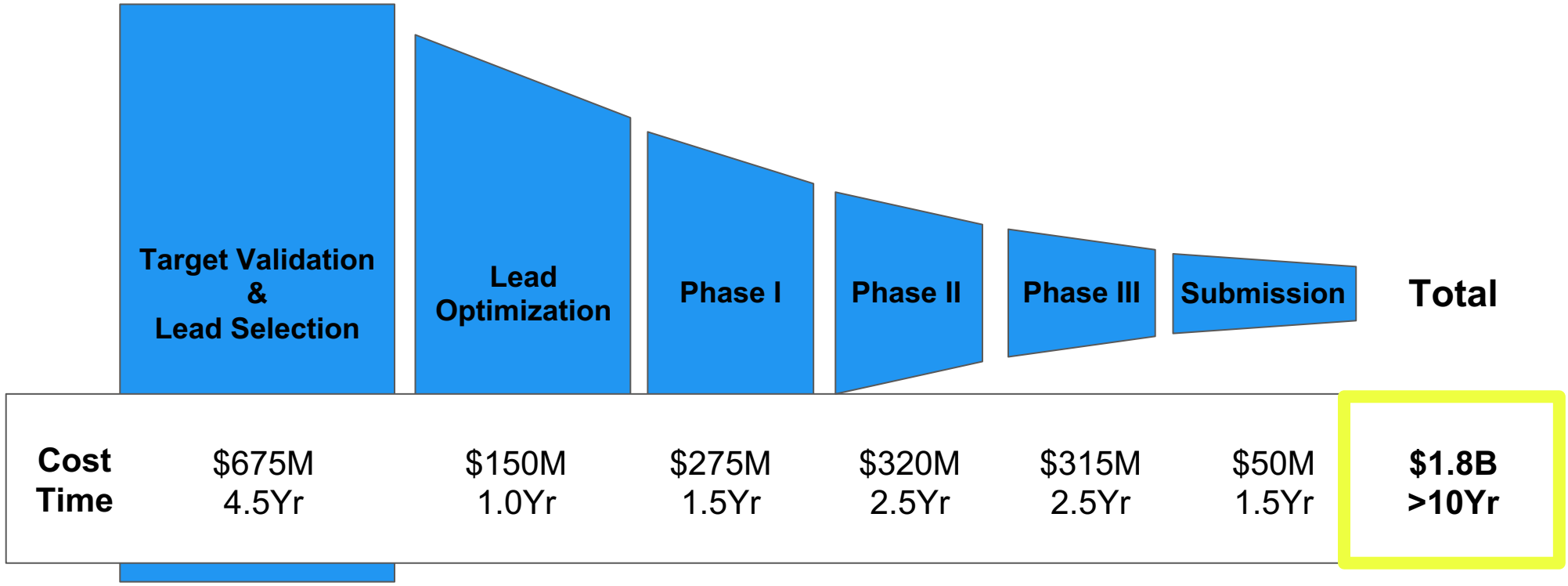


Common Nomenclature

- **Target:** a molecule within an organism that is associated with a disease and the intended destination for a therapy



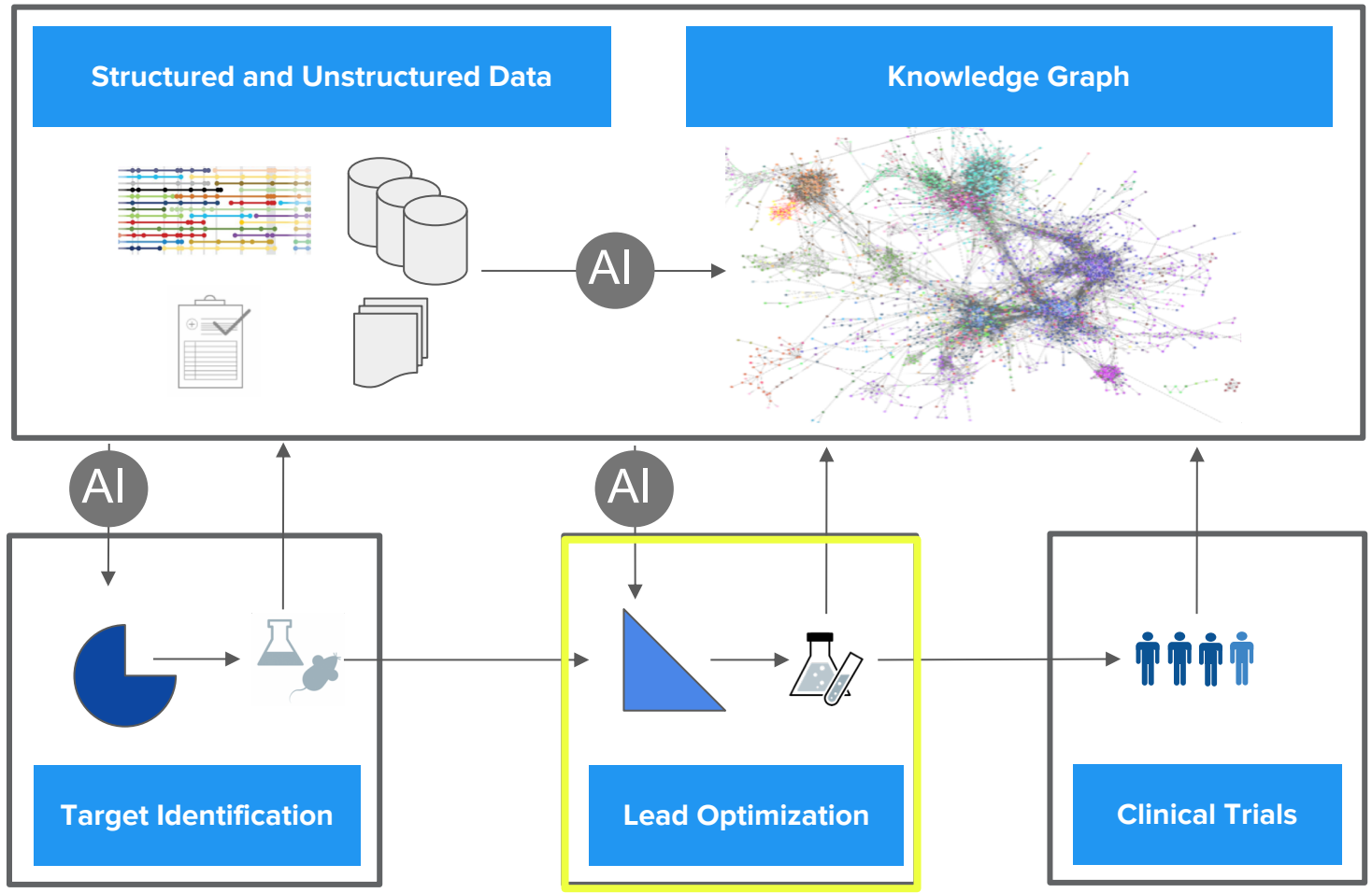
Drug Discovery is an Arduous and Expensive Process



$1.8B and >10 Years to Bring a Drug to Market

Source: *Developability assessment as an early de-risking tool for biopharmaceutical development*, J. Zurdo, 2013, DOI: 10.4155/pbp.13.3

Our Journey from Data to Drugs with Machine Learning



Exploring the Compound Universe is Challenging

- Compound space is large (10^{20} - 10^{60}, depending on definition) and discrete

Exploring the Compound Universe is Challenging

- Compound space is large (10^{20} - 10^{60}, depending on definition) and discrete
- Often interested only in regions of compound space

Exploring the Compound Universe is Challenging

- Compound space is large (10^{20} - 10^{60}, depending on definition) and discrete
- Often interested only in regions of compound space
- Identify compound which binds to target, then use local exploration to improve other properties
- **One solution**: learn search policies or **generative algorithms** to create novel and optimal compounds in regions of interest

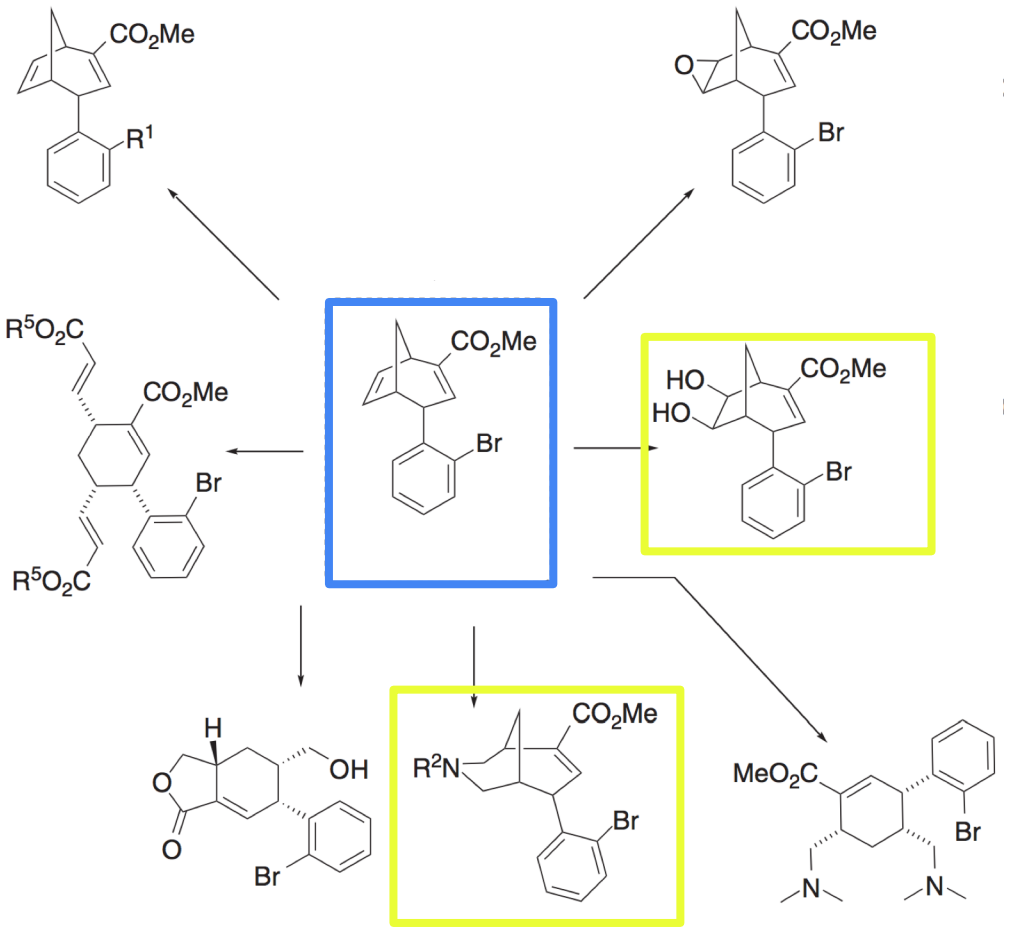
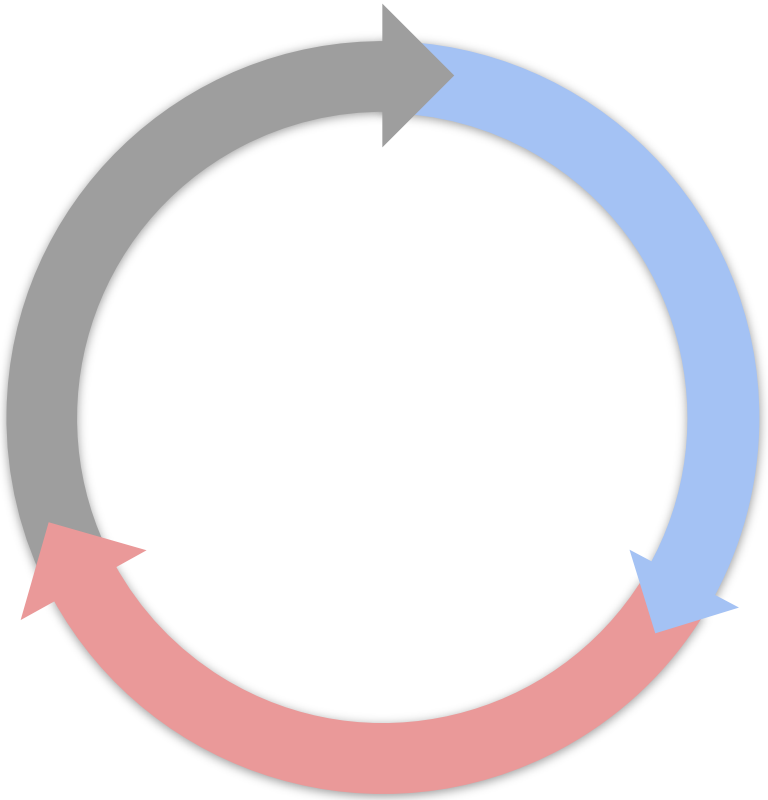


Image credit: Diversity-Oriented Synthesis: Developing New Chemical Tools to Probe and Modulate Biological Systems, Galloway *et al*, 2014, http://www-spring.ch.cam.ac.uk/publications/pdf/2014_DOS_379.pdf

How can we efficiently explore the compound universe in a property driven way?

The Compound Design Cycle

Design -- based on knowledge

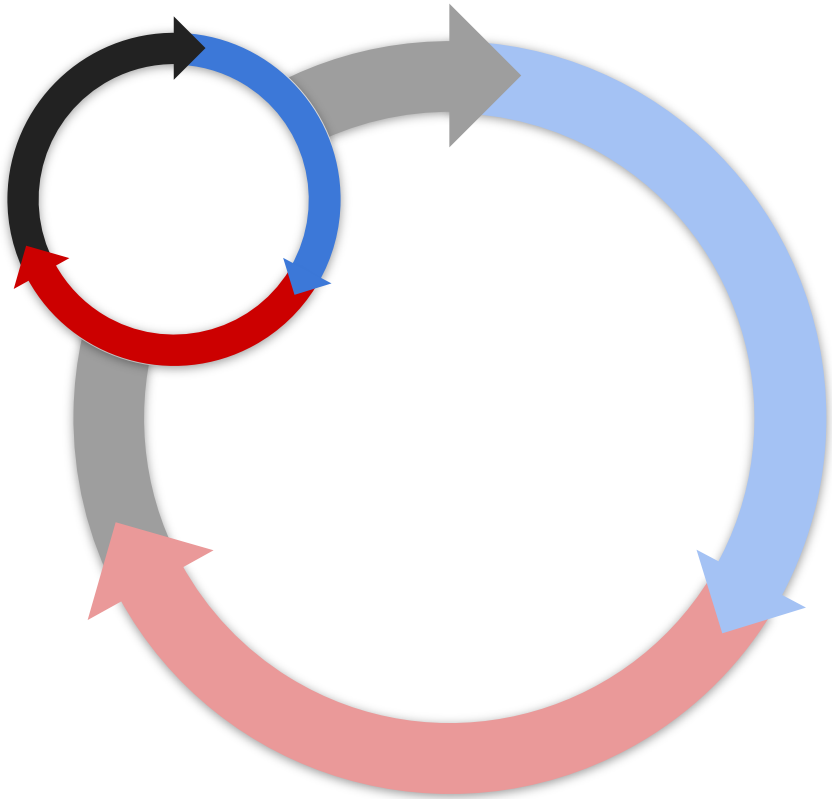


Make -- synthesize in the lab

Test -- with experimental assays

De novo Design mimics the Compound Design Cycle

De novo Design



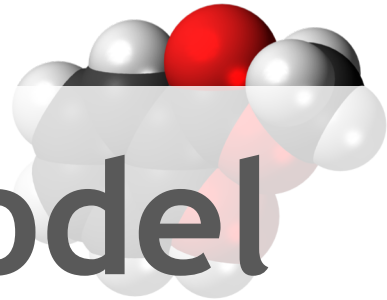
Make

Test

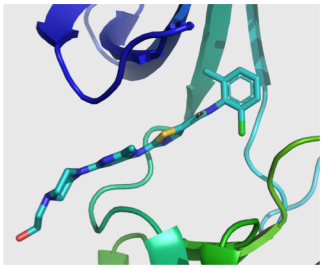
De novo Design

Design
- Search
- Optimization

Make
- Computational modeling



Generative Model

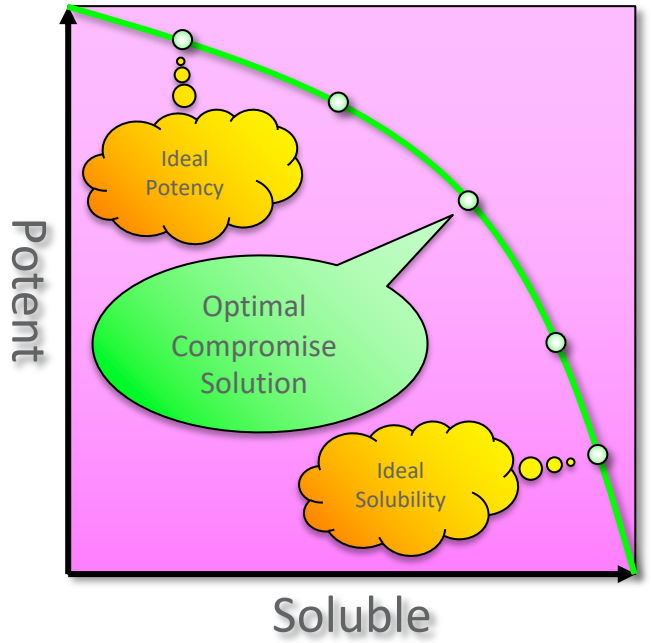


Test
- Model and predict properties with machine learning
- Computer simulations

Multi-Parameter Objective Optimization

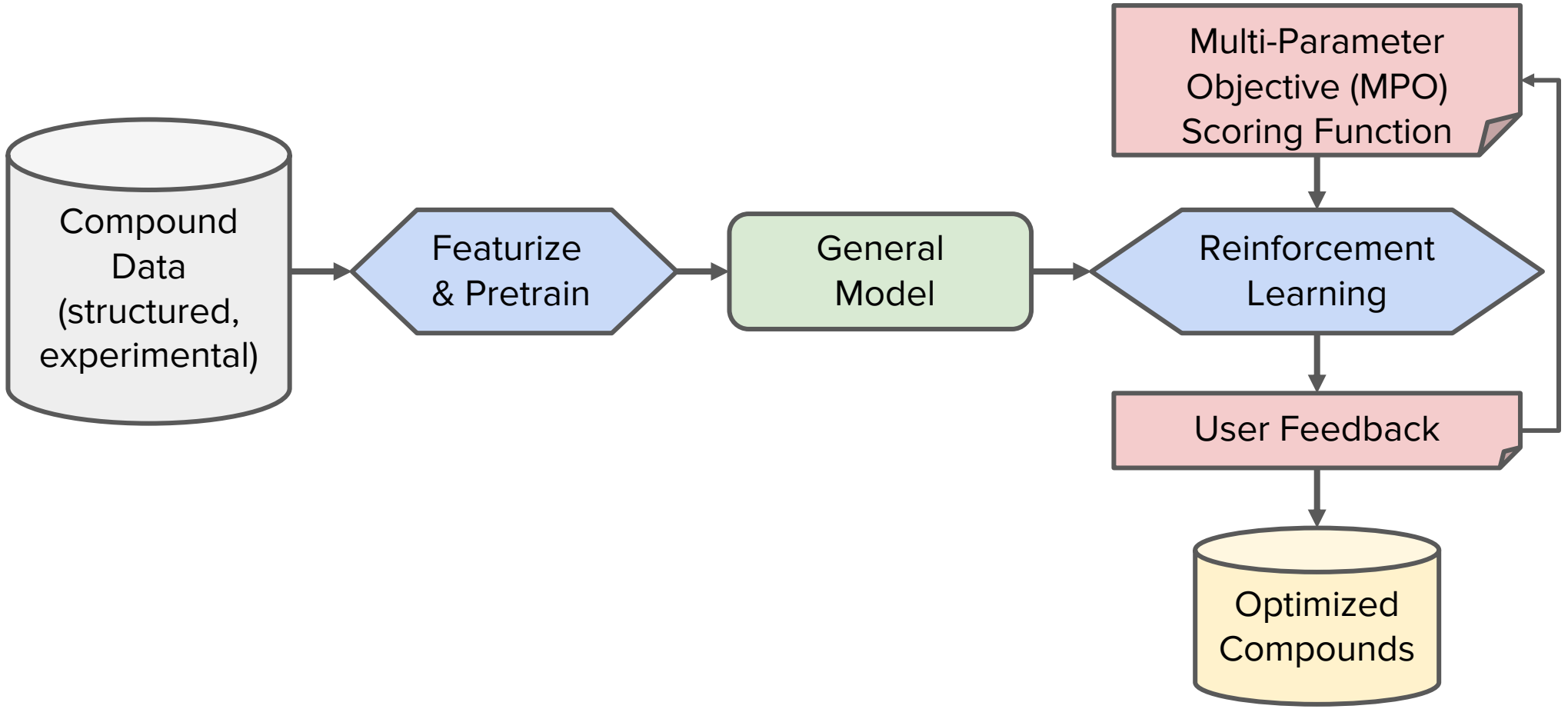
Focus on multiple properties:

- Affinity -- binds to target well
- Toxicity -- isn't harmful to organism
- Selectivity -- binds to only the desired target



Drug design is inherently a multi-objective optimization problem

Learning to Generate the Best Compounds



Aside: machine learning has a long history in chemistry

Machine Learning & Chemistry: Long Time Acquaintances

Analytica Chimica Acta, 248 (1991) 1–30
Elsevier Science Publishers B V , Amsterdam

Review

Neural networks: A new method for solving chemical problems or just a passing phase?

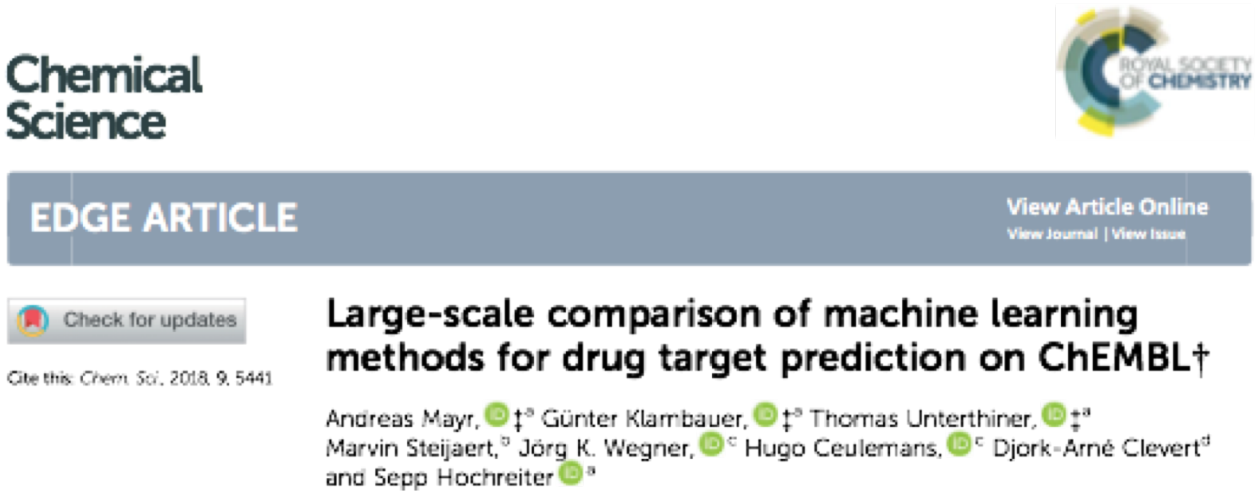
J. Zupan *,1 and J. Gasteiger
Organisch-chemisches Institut, Technische Universität München, D-8046 Garching (Germany)

(Received January 1991)

Feed-forward neural networks have decades of history in computational chemistry

Zupan, Gasteiger, *Analytica Chemica Acta*, **1991**, *248*, 1-30.

Machine Learning & Chemistry: Long Time Acquaintances



Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2018, 9, 5441

Large-scale comparison of machine learning methods for drug target prediction on ChEMBL†

Andreas Mayr, ‡[a] Günter Klambauer, ‡[a] Thomas Unterthiner, ‡[a] Marvin Steijaert,[b] Jörg K. Wegner,[c] Hugo Ceulemans,[c] Djork-Arné Clevert[d] and Sepp Hochreiter[a]

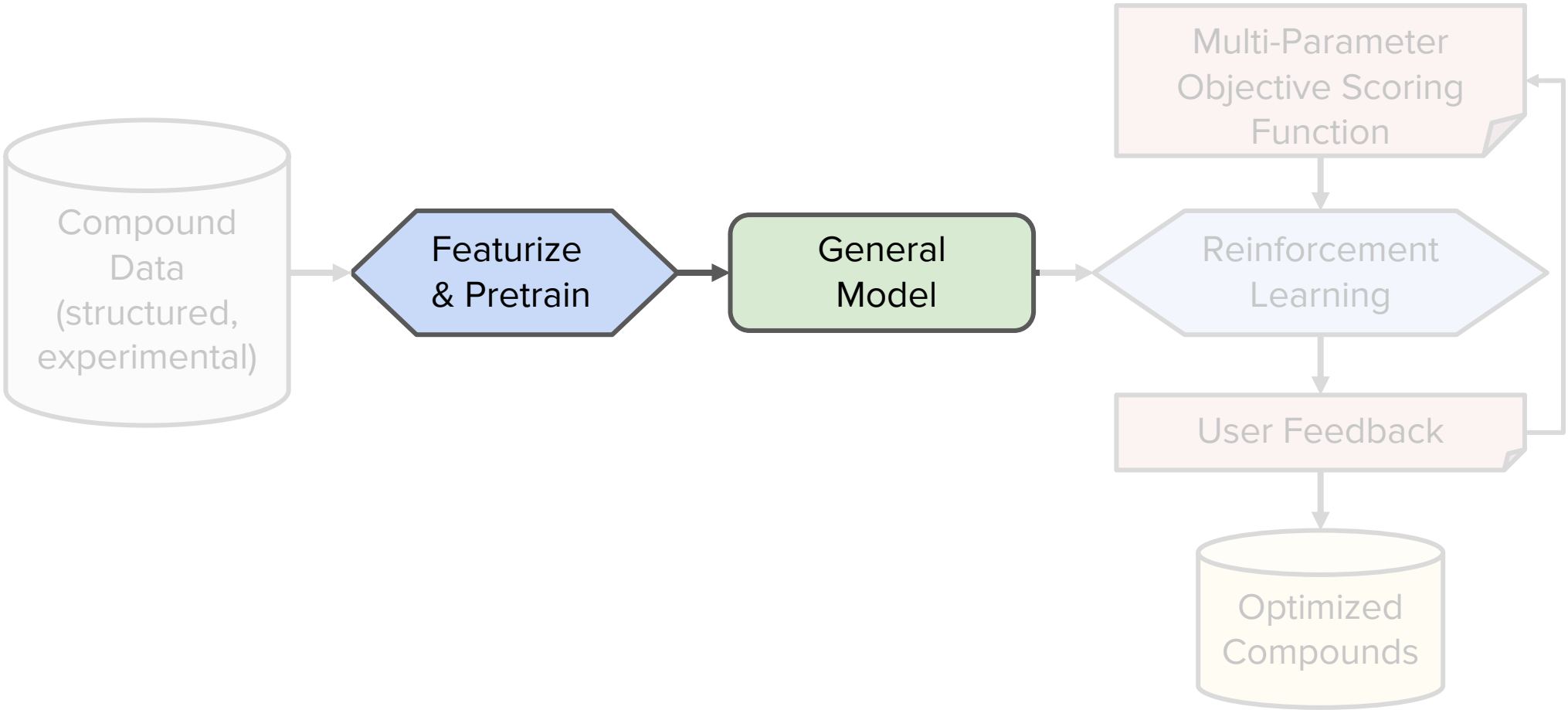
(Received January 2018)

Feed-forward neural networks have decades of history in computational chemistry

Large chemical datasets have fueled recent successes with deep neural networks

Mayr, Klambauer, Unterthiner, Steijart, Wegner, Ceulemans, Clevert, Hochreiter, *Chem. Sci.*, **2018**, *9*, 5441-5451

Learning to Generate the Best Compounds



How can neural networks be used to predict compounds?

Language Models

Predict probability of next word in a sentence

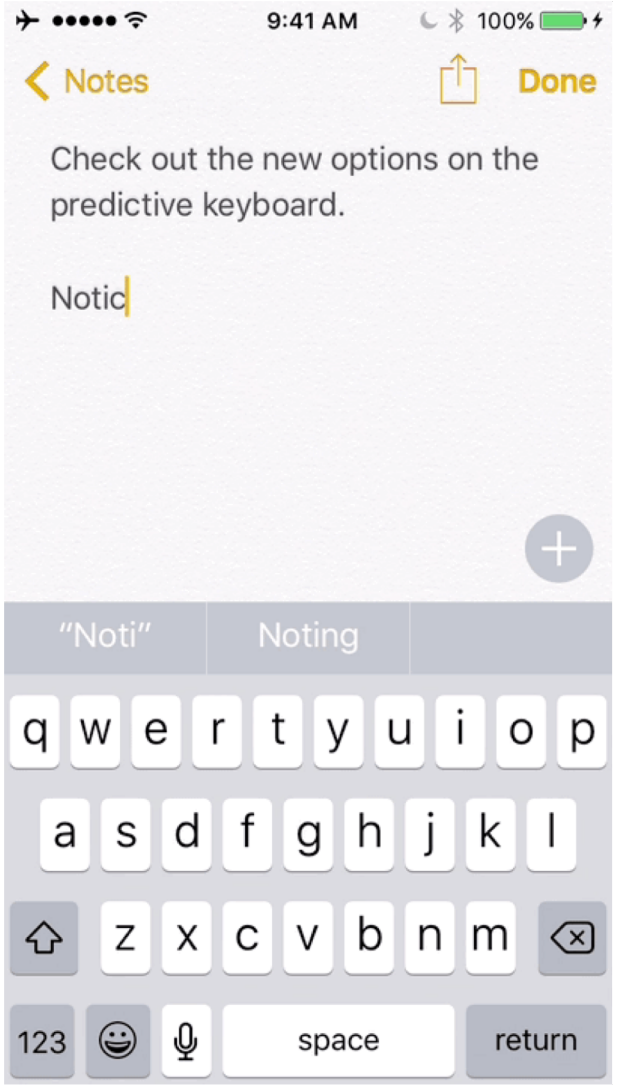
Chemistry is _____?

amazing

difficult

potato

compares

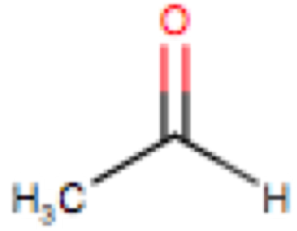


Language Models for Chemistry?

Chemistry → is → ?

Use SMILES!

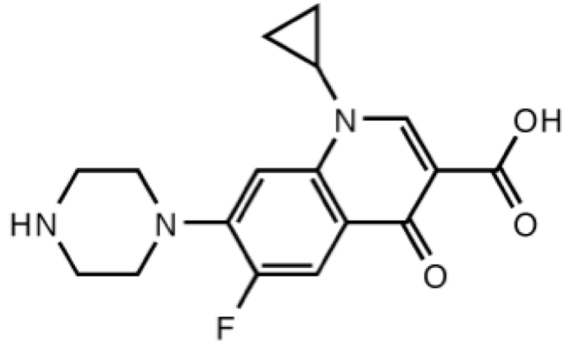
C → C → = → ?

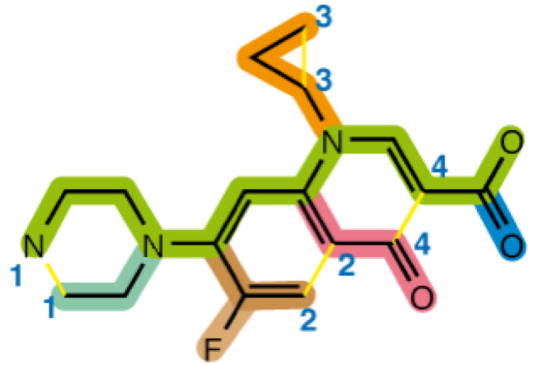


Predict the probability of next character

Featurizing Compounds with SMILES

- Simplified Molecular Input Line-Entry System (SMILES)
- Symbolic string obtained from depth-first traversal of a compound graph
- Multiple variations -- standardization is challenging
- Chemical rules must be obeyed for validity

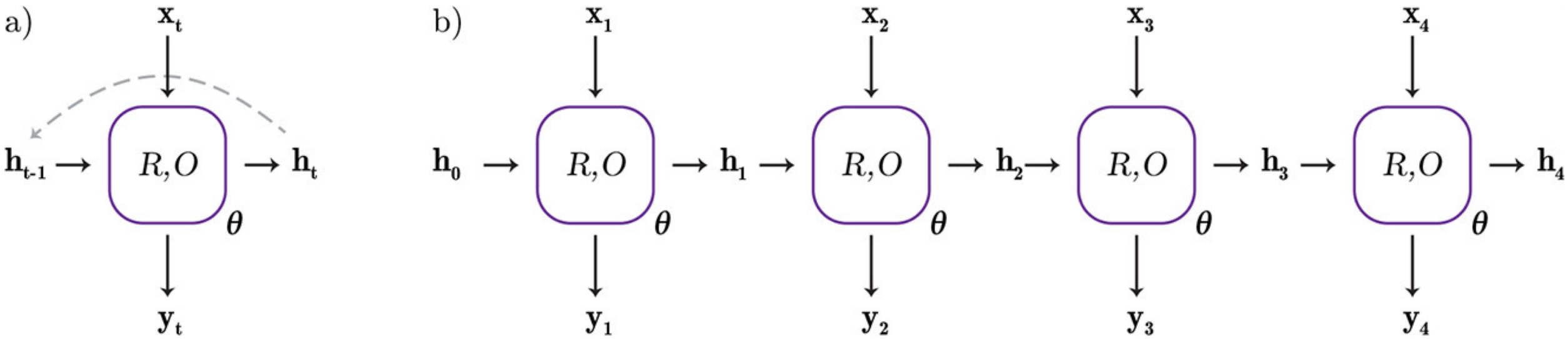




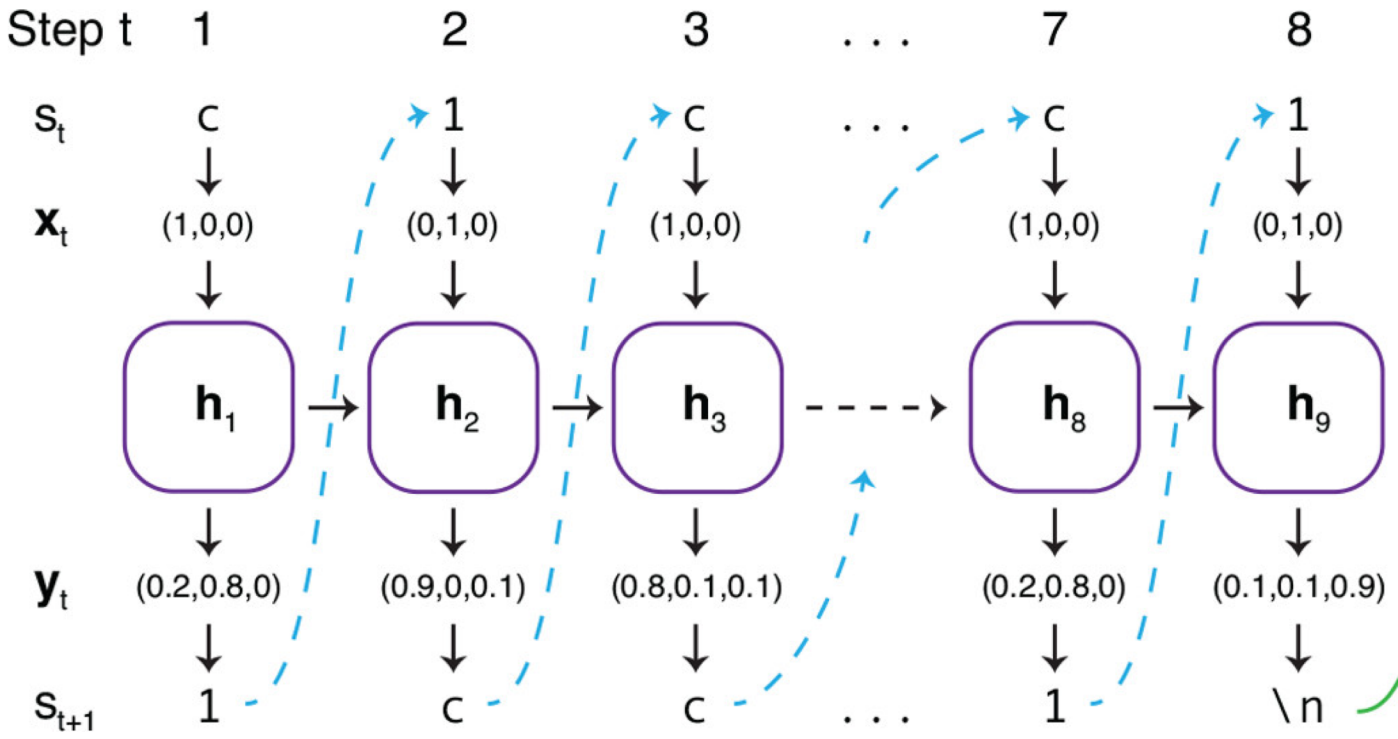
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

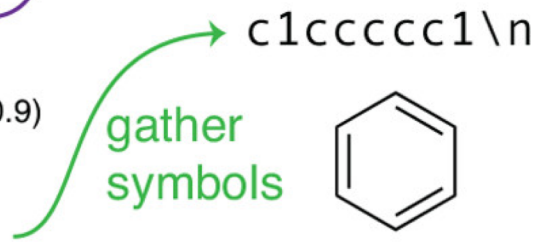
Image credit: SMILES Wikipedia entry, https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

Recurrent Neural Networks (RNNs)



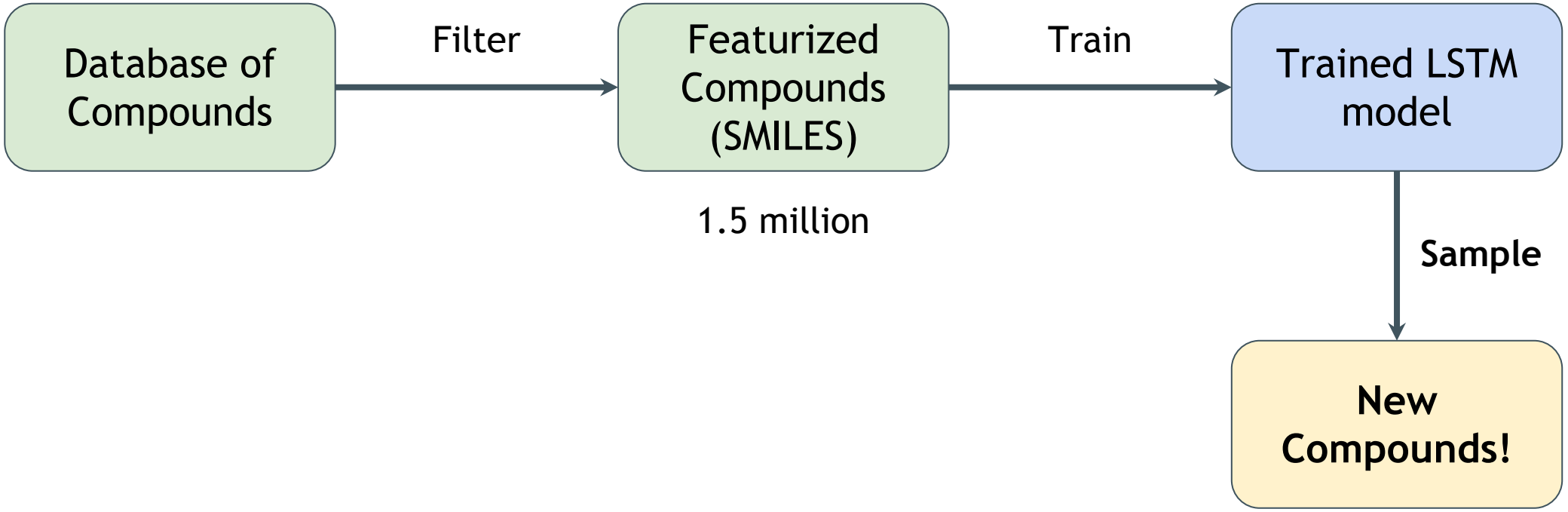
Probing Compound Space with RNNs





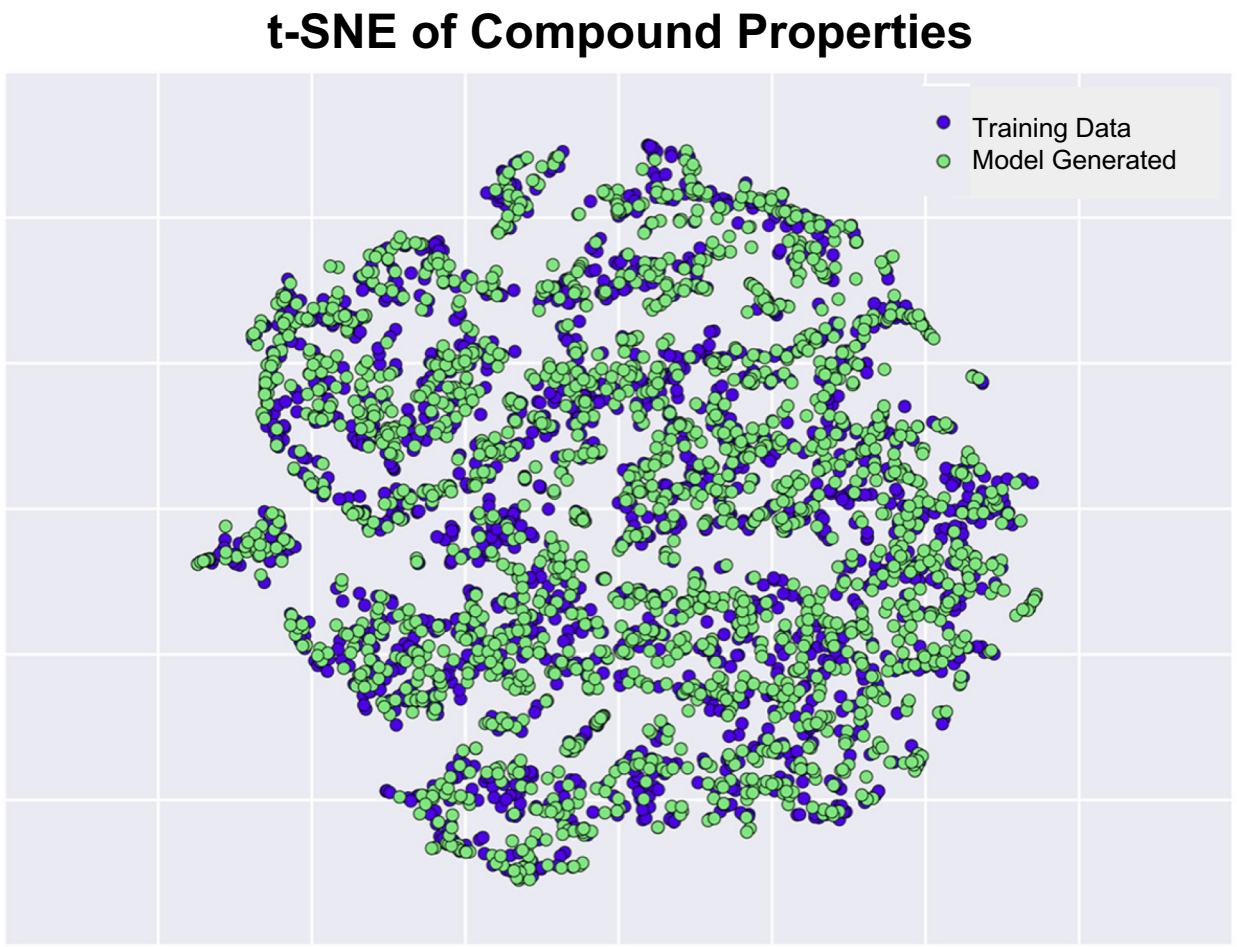
- Network is a character-level LSTM with initial embedding layer and teacher forcing
- Trained with SMILES representations of compounds from database

Probing Compound Space with RNNs



Neil *et al*. *Exploring Deep Recurrent Models with Reinforcement Learning for Molecule Design,* ICLR Workshop track *https://openreview.net/forum?id=Bk0xiI1Dz*

Probing Compound Space with RNNs



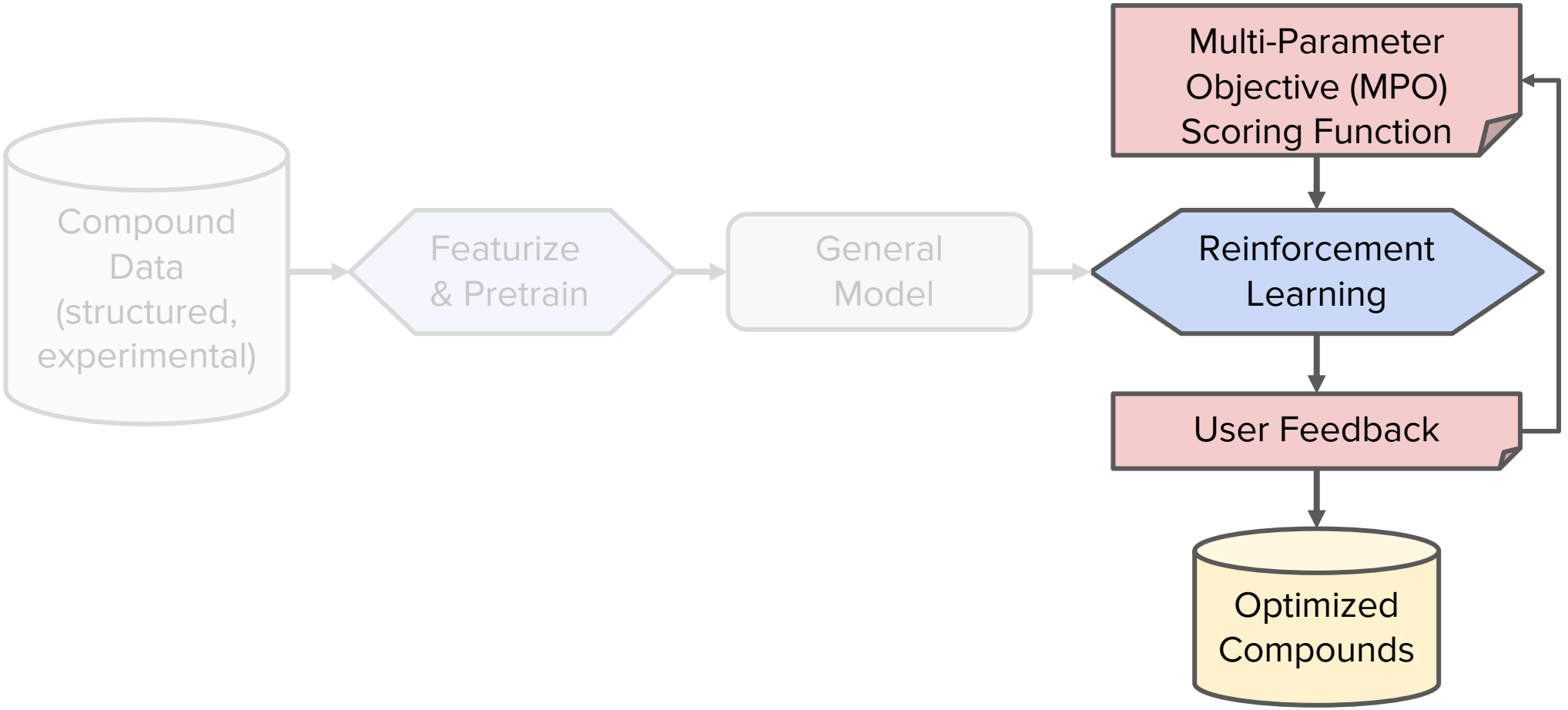
- Distribution of physical properties in training data as reproduced by sampling from model
- SMILES strings produced are:

95% valid

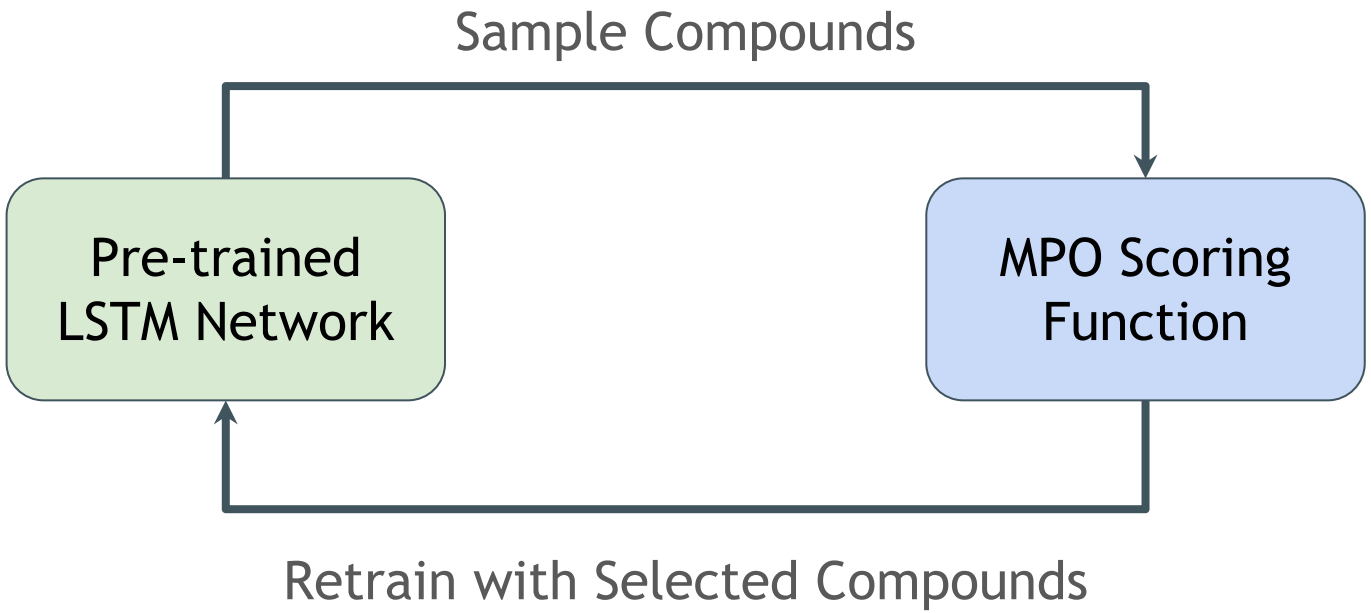
90% novel

Neil *et al*. *Exploring Deep Recurrent Models with Reinforcement Learning for Molecule Design,* ICLR Workshop track *https://openreview.net/forum?id=Bk0xiI1Dz*

Learning to Generate the Best Compounds



Reinforcement Learning (RL) Rounds



Model Refinement with RL and User Feedback

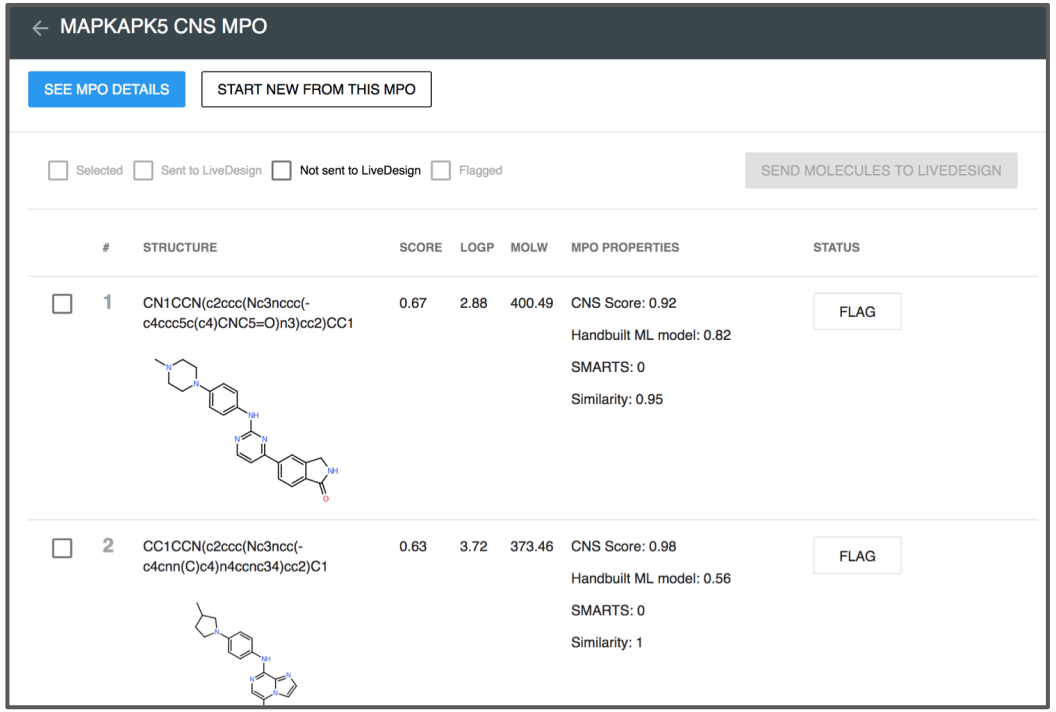
Compounds from Model

Score

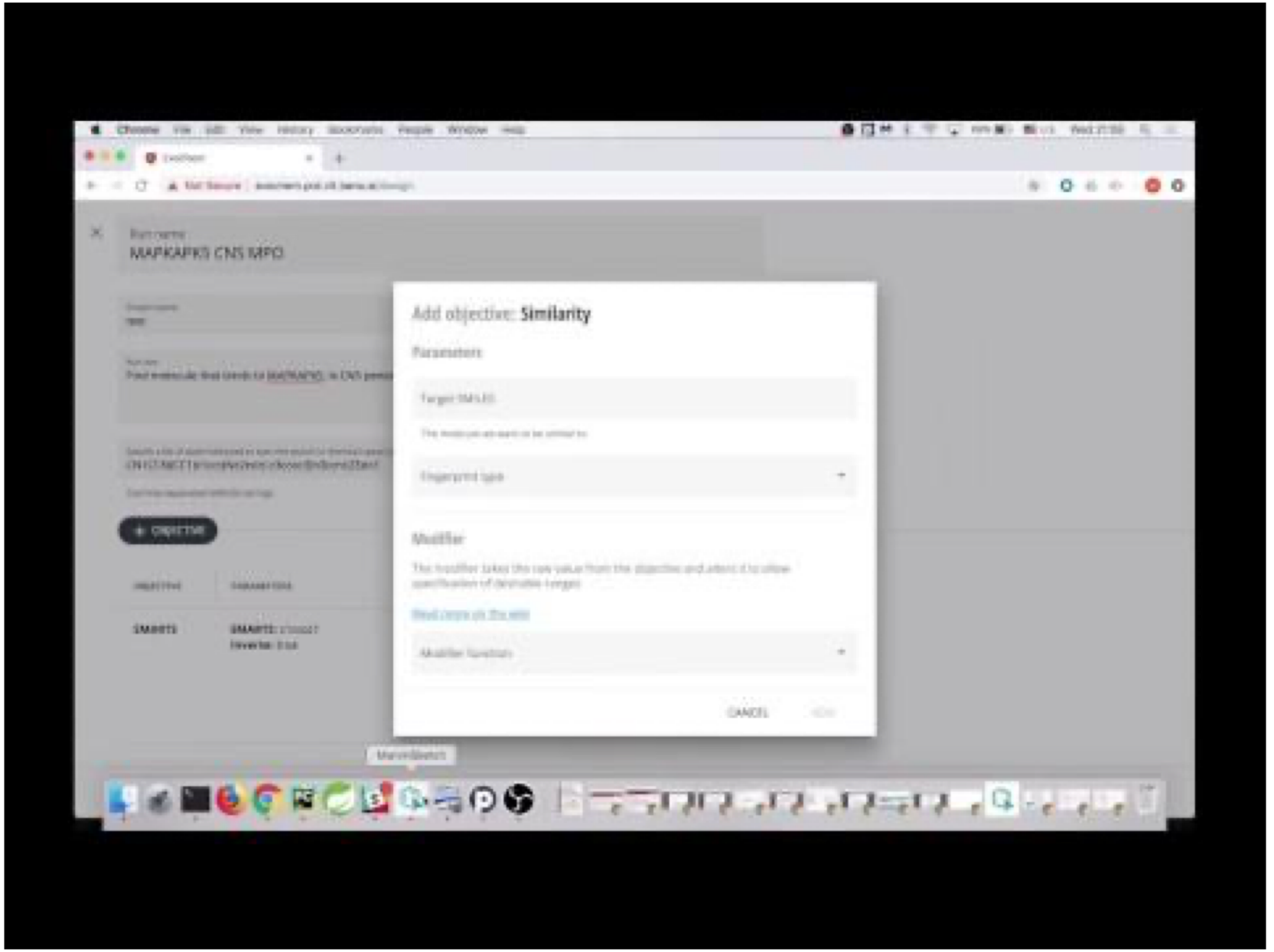
C 1 = C C = C O = C O = C 1	0.3
C 1 = C C = C C = C C = C C	0.5
C 1 = C C = C C = C C	0.6
C 1 = C C = C C = C C = C 1	0.9

1. Sample compounds from retrained model and score according to MPO
2. Select best compounds for another round of retraining or to present to user

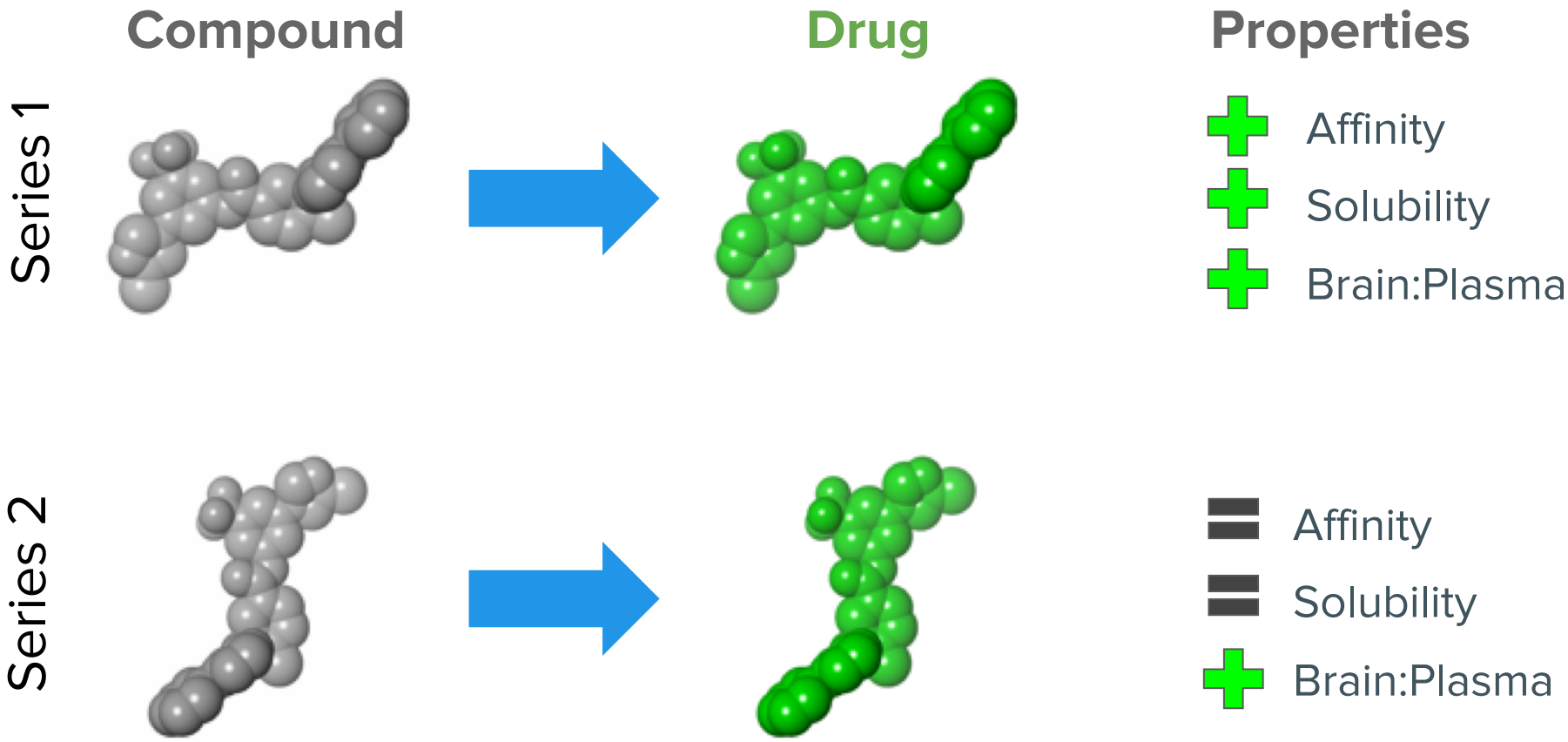
User Feedback Sessions



MPO Setup & User Feedback



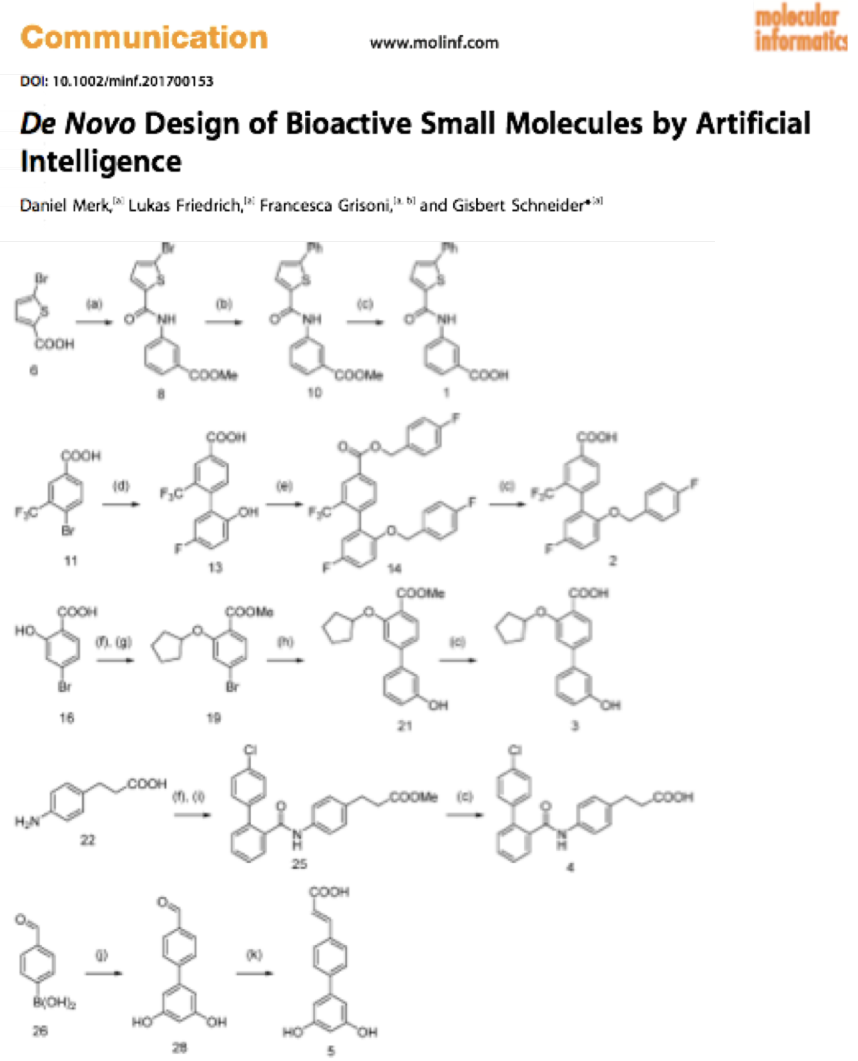
Finding Drugs for ALS with Generative Models



< 3 months *vs* industry average of 1.5 - 3 years

External Demonstration of Drugs Designed by ML

- First published validation of compounds optimized by machine learning
- Models fine tuned with transfer learning
- Five highly active drugs produced



Merk, Friedrich, Grisoni, Schneider, *Mol. Inf.* **2018**, *37*, 1700153

Benchmarking: How Good is a Generative Model?

Assesses two dimensions:

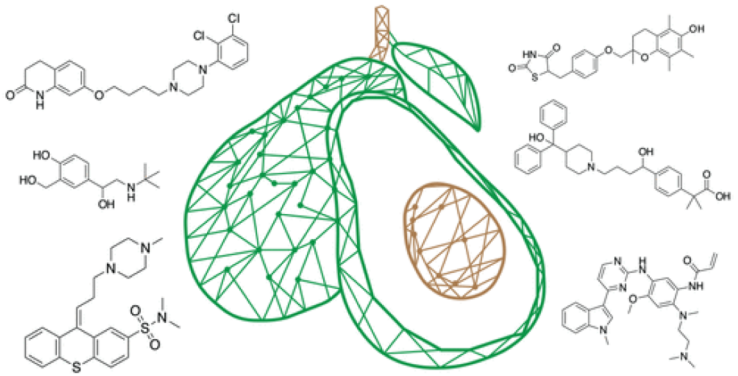
- **Distribution based:** how well a model can learn the chemical distribution of data
- **Goal based:** model generates molecules to satisfy pre-defined goal(s)

GuacaMol: Benchmarking Models for De Novo Molecular Design

Nathan Brown, Marco Fiscato, Marwin H.S. Segler, Alain C. Vaucher

https://arxiv.org/abs/1811.09621

Code: https://github.com/BenevolentAI/guacamol



Acknowledgements
Marwin Segler, Nathan Brown, Dan Neil, Alix Lacoste

Michelle Gill: michelle.gill@benevolent.ai
Twitter: @modernscientist

Website: https://benevolent.ai