AstraZeneca
IMED Biotech Unit

Deep Learning (at AstraZeneca): Towards augmented design and automated design cycles

Christian Tyrchan
Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (RI), Biopharmaceuticals R&D, AstraZeneca, Gothenburg Sweden

2018 BigChem

December 2018

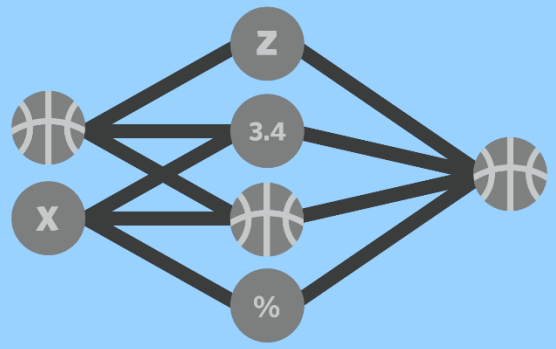
Evolution in AI

Artificial Intelligence

A computerized system that exhibits behaviour that is commonly thought of as requiring intelligence

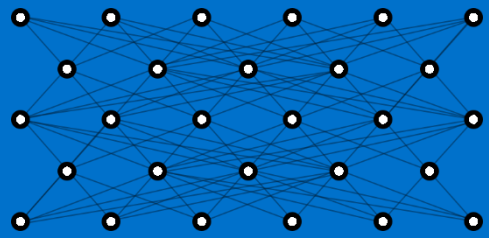
Machine Learning

A statistical process that starts with a body of data and tries to derive a rule or procedure that explains the data or can predict future data



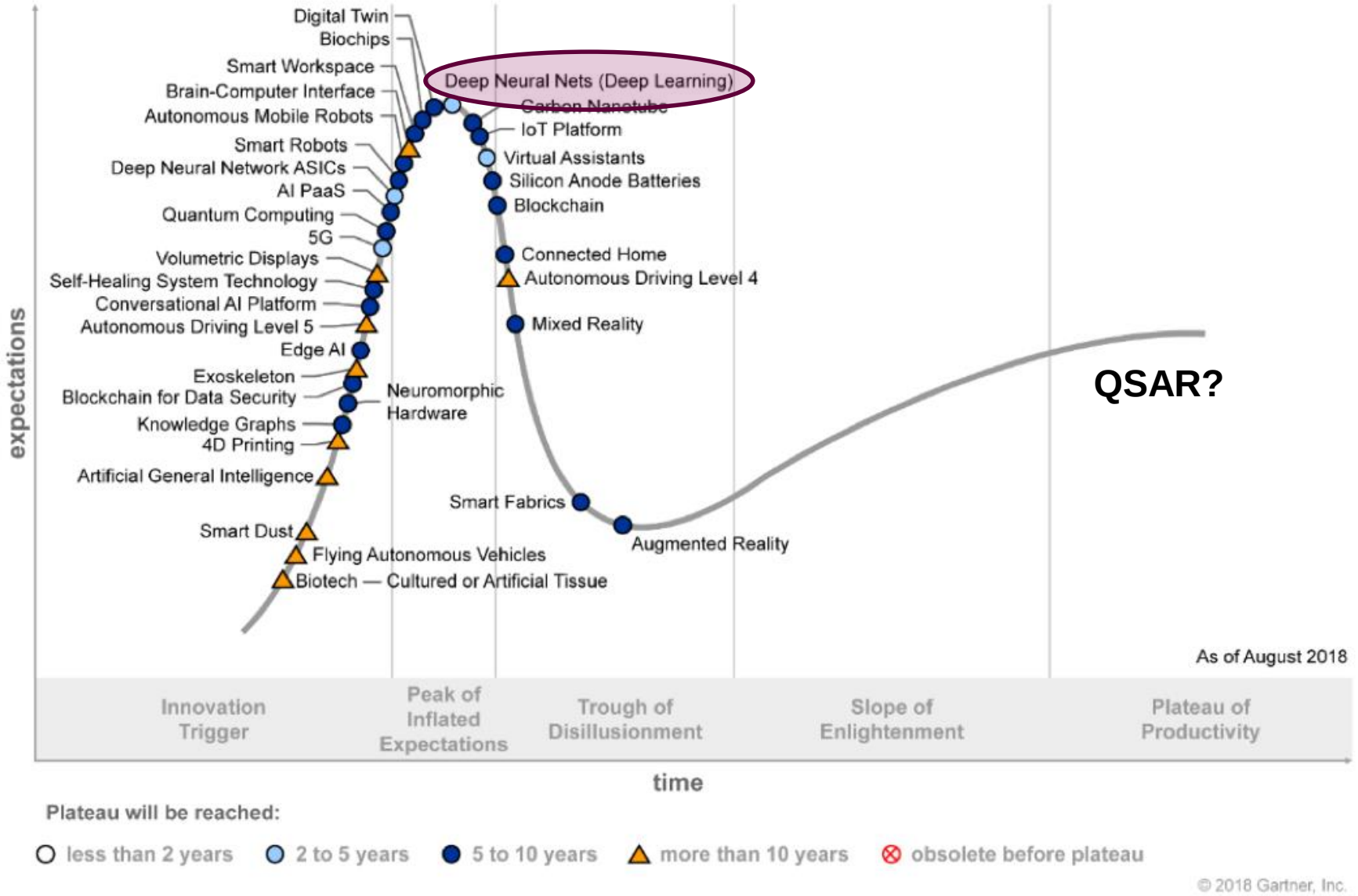
Deep Learning

A subfield of ML which uses structures loosely inspired by the human brain, consisting of a set of units (or “neurons”)

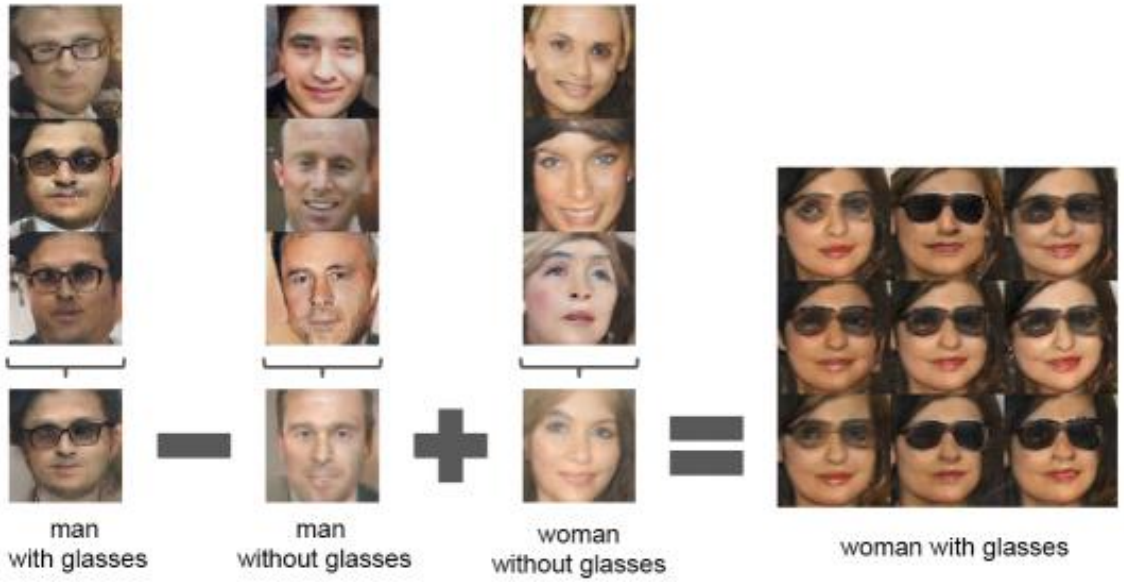


1950’s 1960’s 1970’s 1980’s 1990’s 2000’s 2010’s

Deep Learning - Gartner Hype Curve



Deep Learning - Image manipulation & generation

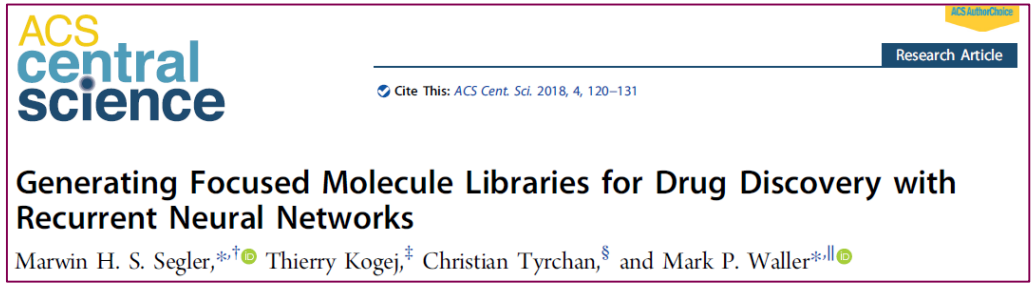


- From manipulating pictures to making up virtual people

Figure 5: 1024 × 1024 images generated using the CELEBA-HQ dataset. See Appendix F for a larger set of results, and the accompanying video for latent space interpolations.

A. Radford et al Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2016
T. Karras et al Progressive Growing of GANs for Improved Quality, Stability, and Variation, 2018

De novo molecular generation with deep learning has developed very rapidly



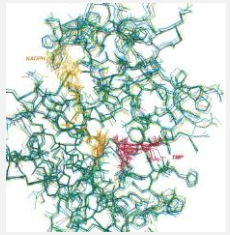
The role for machine learning in drug design

TODAY

TOMORROW?

Bioactivity only

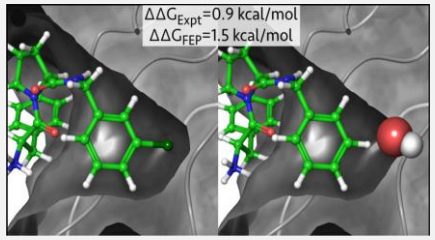
Structure-based approaches



Modify ligand based on interactions with the protein

Require a protein structure

Free energy perturbation

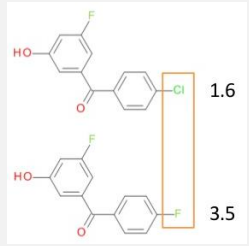


Compute Δfree energy from molecular dynamics

High computational cost of predictions limits use to LO

One-parameter

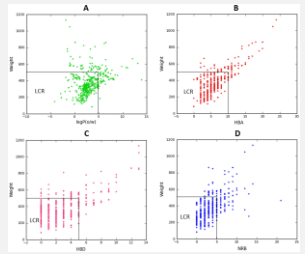
QSA(P)R



Single-endpoint predictions

Multi-parameter

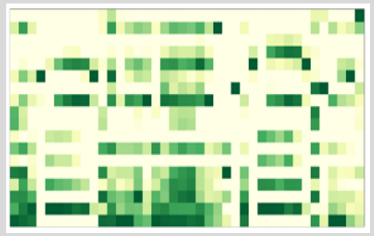
Desirability score



Define and score chemical (QSAR) or property (invQSAR) space

Constrained to specific areas of chemical or property space

AI



Learn, unconstrained, from previous human designs

Lack of experimental validation
Requires well-organised data

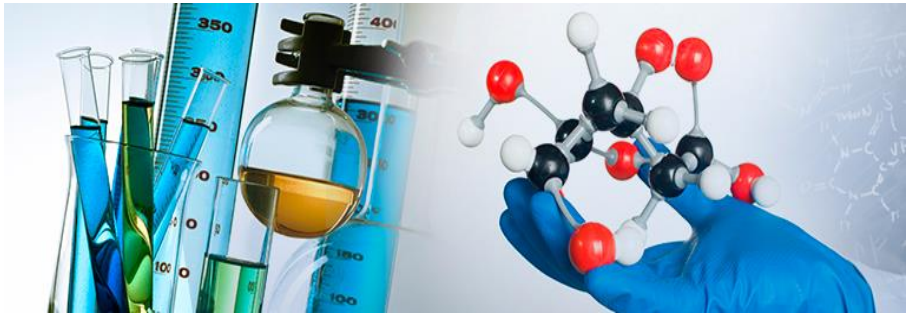
We can combine single parameter predictions in multi-parameter modelling

Medicinal chemistry: Our scientific interest

What to make next?

De novo design

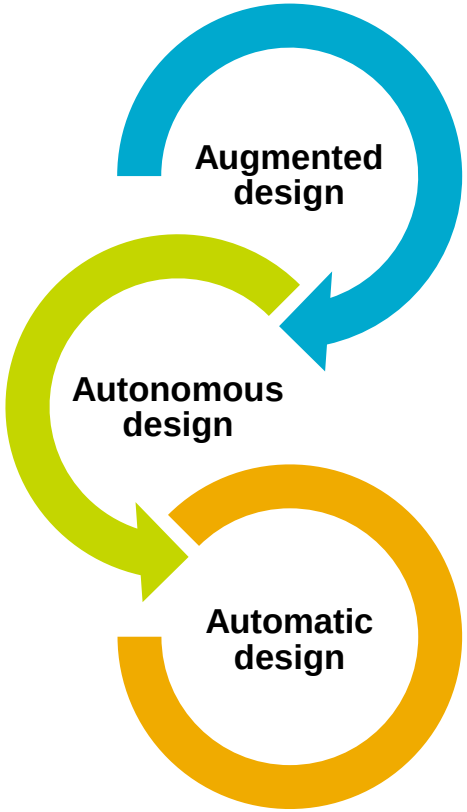
How to make it?



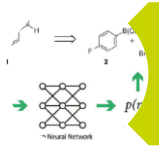
Retrosynthesis

Deep learning at AstraZeneca: Vision

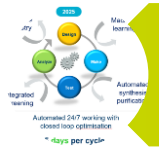
- Creating a world class leading AI platform for drug discovery projects




AI design platform


AI reaction platform

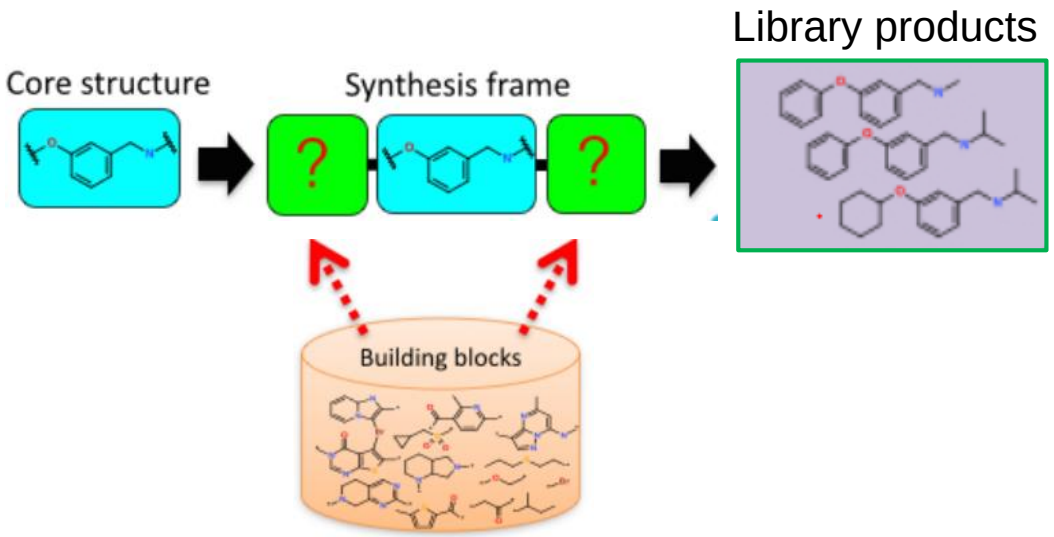

Automatic make and test (iLab)


Integration

Segler M.H.S. et al. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction, Chemistry, 2017, 23(25), 5966-5971
Segler M.H.S. et al. Planning chemical syntheses with deep neural networks and symbolic AI, Nature, 2018, 555, 604-610

De Novo molecule design using generative models

Library-model based molecule design



- Library based molecule design strategy: rule based methods based on predefined reaction rules and available building blocks.

Generative model based molecule design

- Making *de novo molecule design using a (probabilistic generative) model*
- Data-driven – learns from molecules already synthesized by human experts, not rule-driven (predefined building blocks, reactions or rules).
- Can be fine-tuned using a scoring function based on molecular desirability (Druglikeness, ADMET, target activity etc.)
- After training it should generate highly desirable (based on score) structures

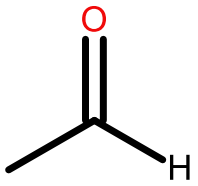
Natural language generation and molecular structure generation

- Can we borrow concepts from natural language processing and apply to SMILES description of molecular structures to generate molecules?

The → grass → is → ?

- Conditional probability distributions given context
- $P(green \mid is, grass, The)$

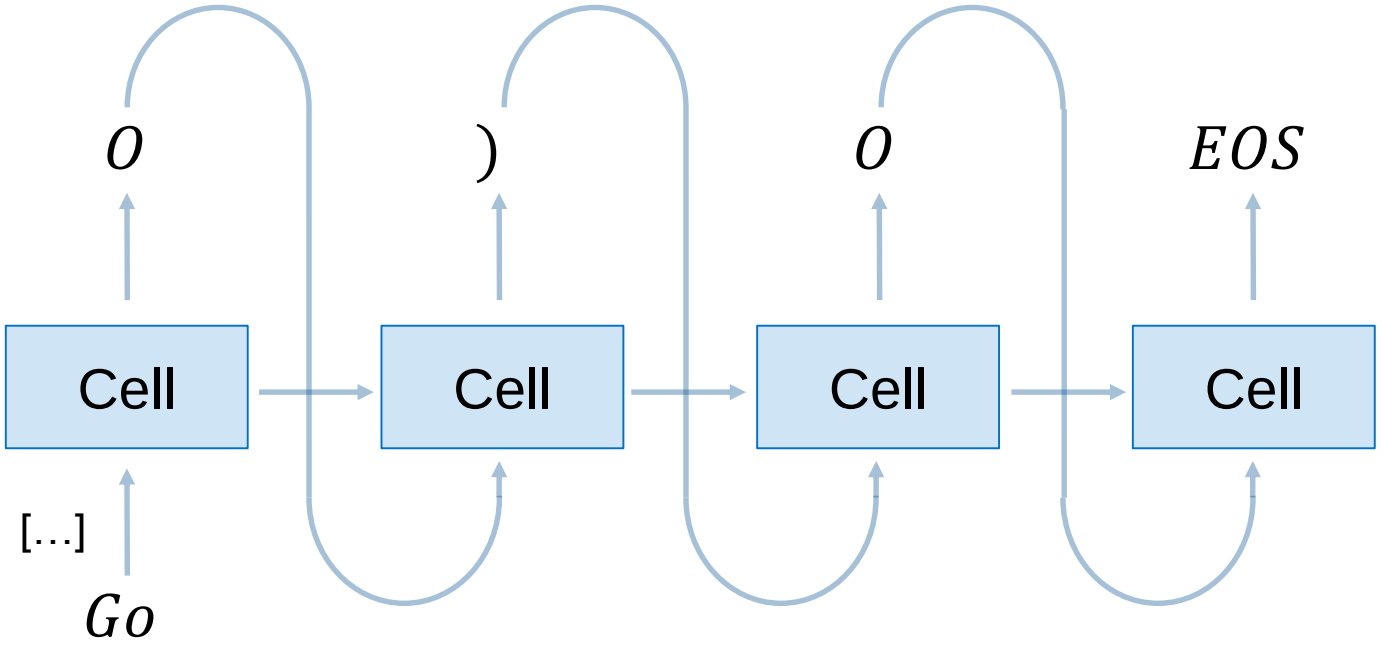
C → C → = → ?

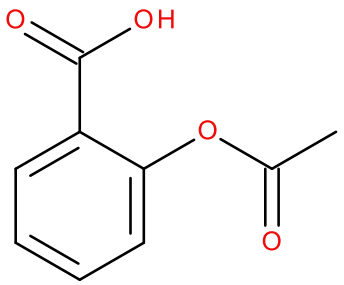


- $P(O \mid =, C, C)$

Generative model: Recurrent Neural Networks

- When trained, can be used to generate new sequences (e.g. SMILES)
- Sample from probability distribution at every step. Use sampled character as next input
- Trained using Maximum Likelihood Estimation to maximize the likelihood of next character

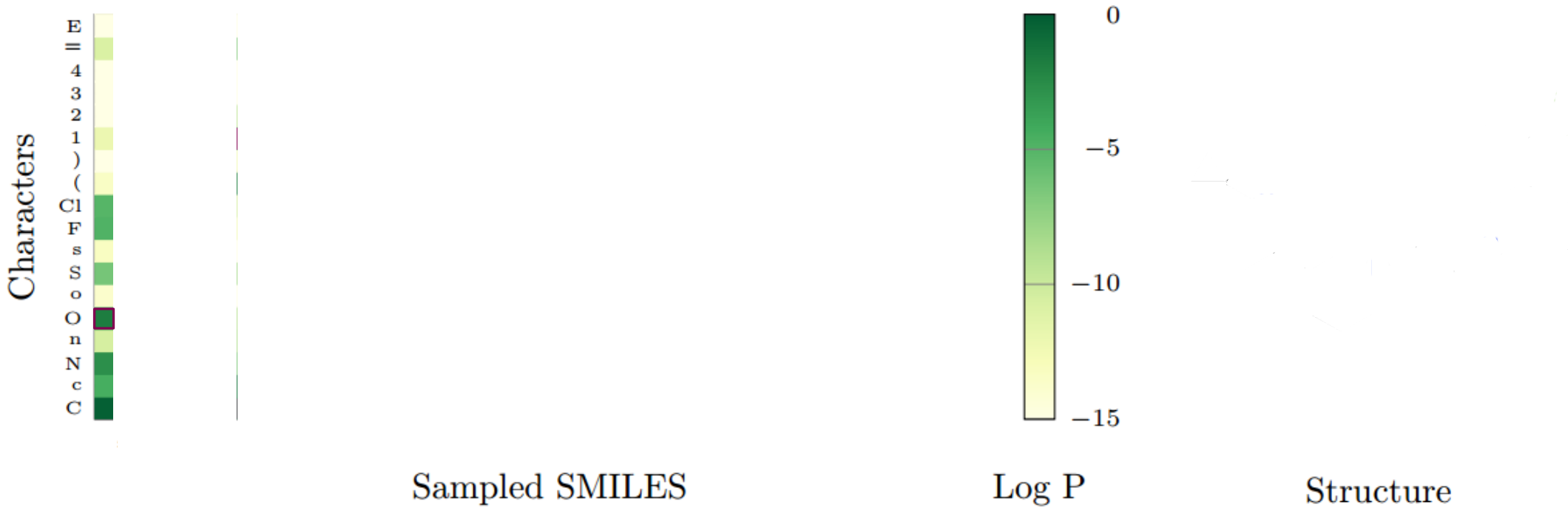




CC(=O)Oc1ccccc1C(=**O)O**

Generative model

- **1.5 million SMILES** from ChEMBL
- **Conditional probability distributions** from natural language processing

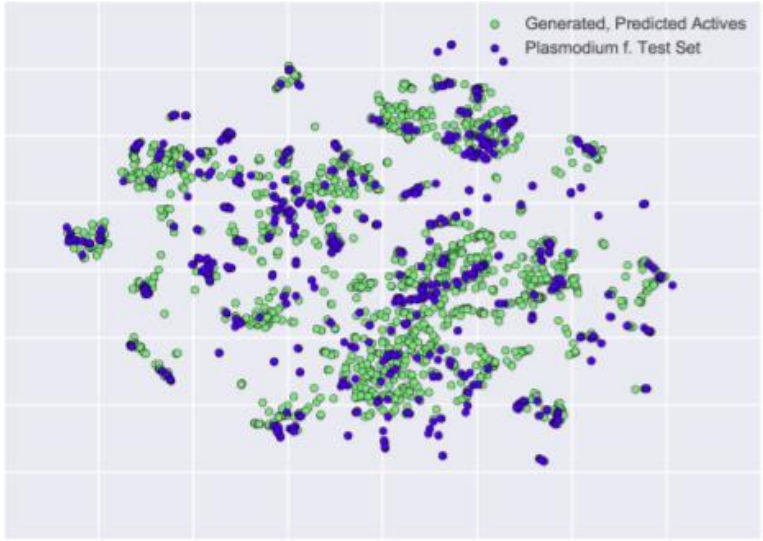


Molecular de-novo design through deep reinforcement learning; Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen; Journal of Cheminformatics, 2017, 9:48 https://doi.org/10.1186/s13321-017-0235-x

Some misconceptions about de novo RNN generated molecules

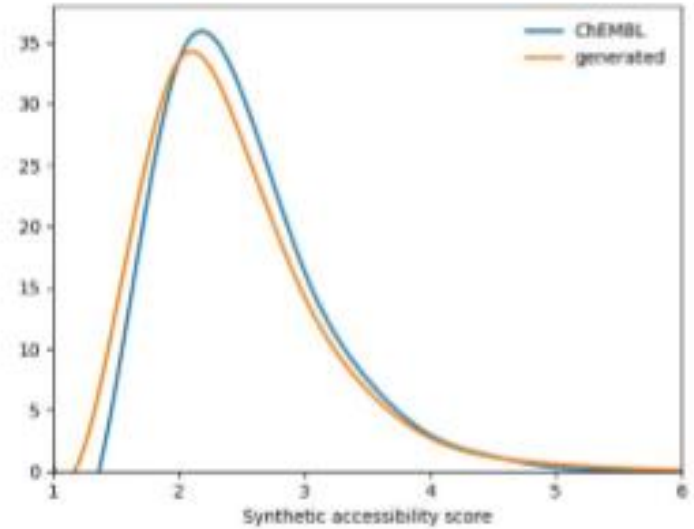
“The molecules are not diverse”
“The molecules are not synthetic feasible”

Answer: The generated molecules follows the properties of the dataset used as prior



Segler et al ACS Central Sci. 2018, 4, 120-131

Diversity

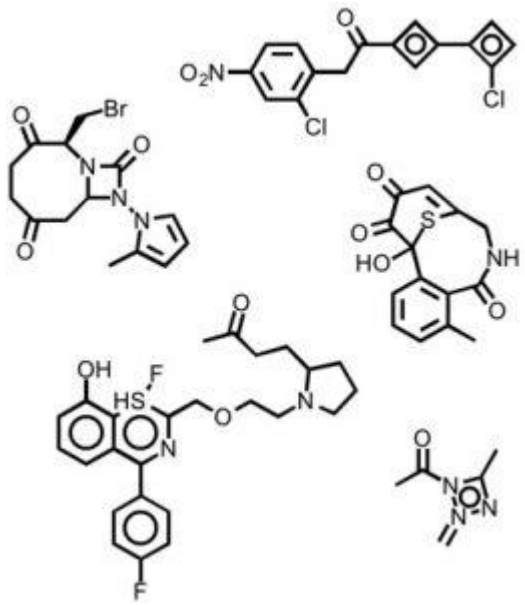


Ertl et al arXiv:1712.07449

Synthetic feasibility

Generative model

- State of the art not 2 years ago

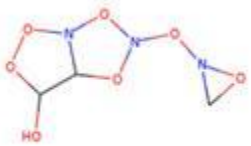


Calculating A Few Too Many New Compounds

By Derek Lowe | 8 November, 2016

[−5.91, 181.104]

[−6.01, 195.1]



[−6.22, 195.08]

February 2018, 100% novelty and 98% chemical valid structures

What to make next?

Possible to generate billions of reasonable molecules, ignoring the relevant questions:

MedChem perspective: What to make next?

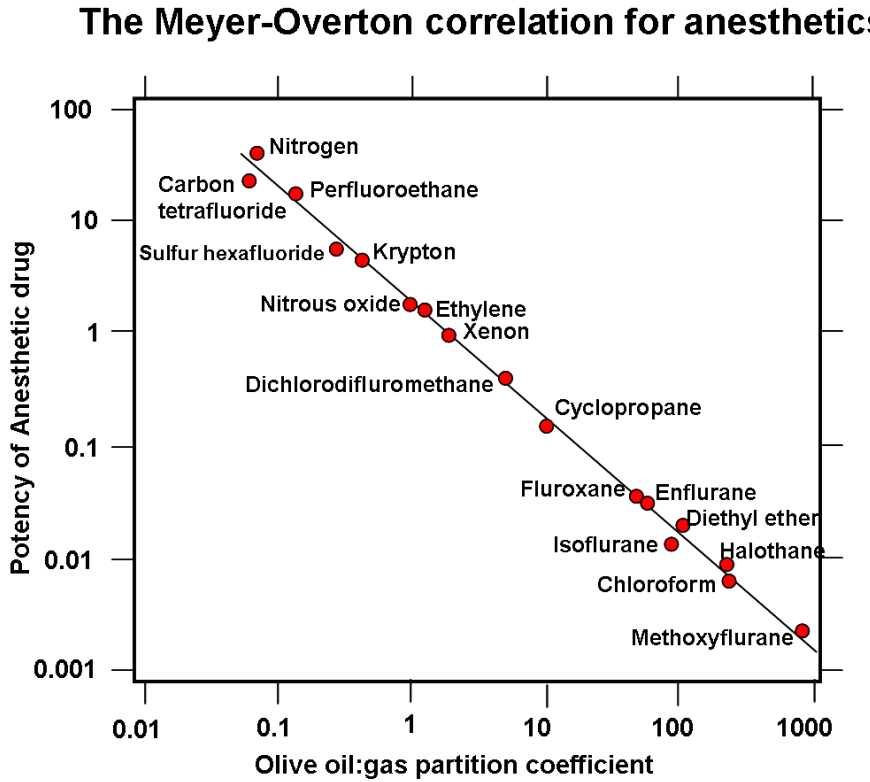
Model improvement: What to make next?

Machine Learning -> QSA/(P)R to help!?

QSAR: Since the 19th century

- Meyer-Overton-Rule

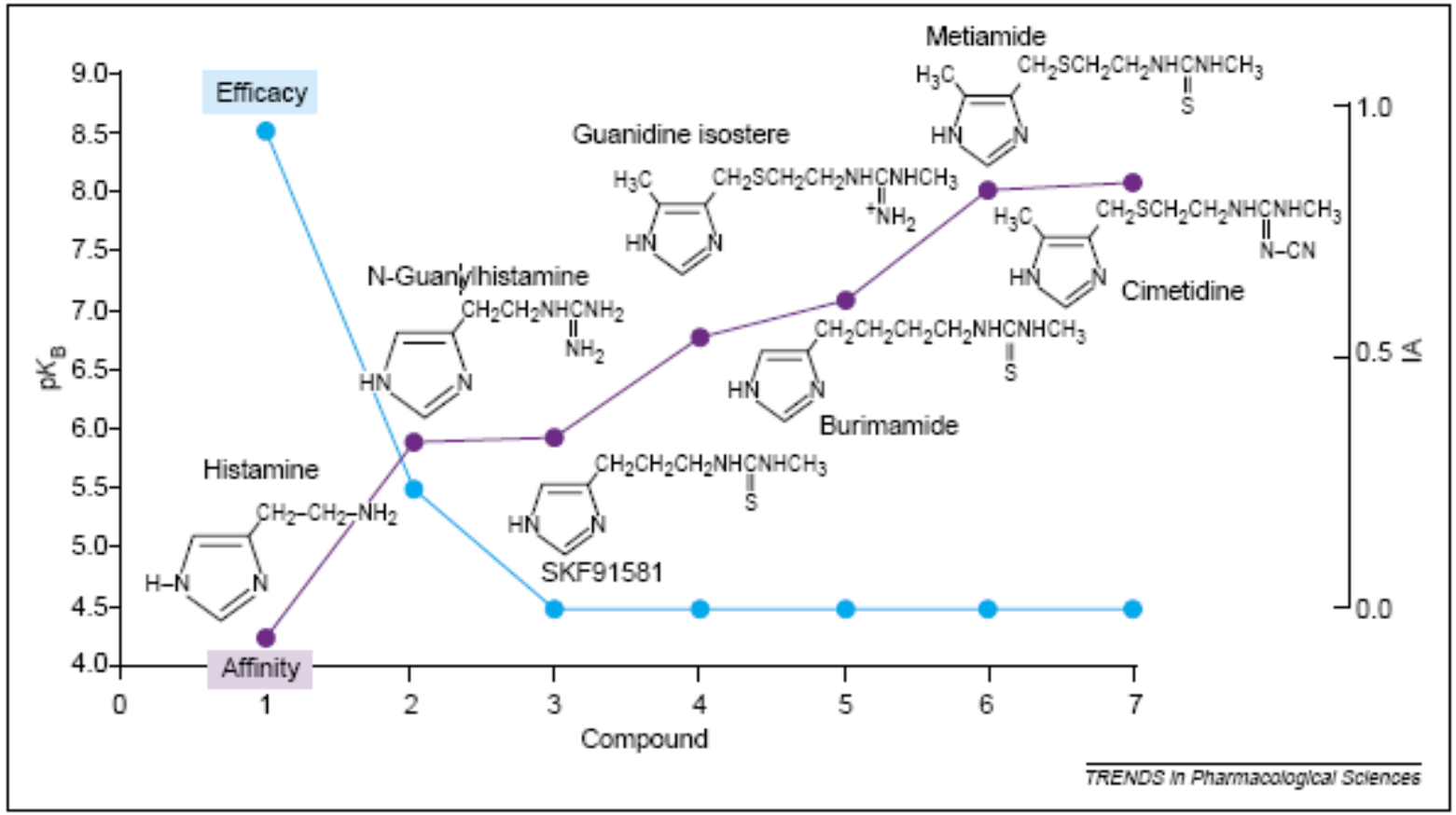
The *permeability coefficient* of a solute is **linearly** related to its *partition coefficient* between oil and water.



https://de.wikipedia.org/wiki/Meyer-Overton-Korrelation

The correlation between lipid solubility and potency of general anaesthetics is a necessary but not sufficient condition

QSAR: Inverse correlation between efficacy and affinity



H2 dataset

Trends in Pharm. Sci. 2002, 23, 275

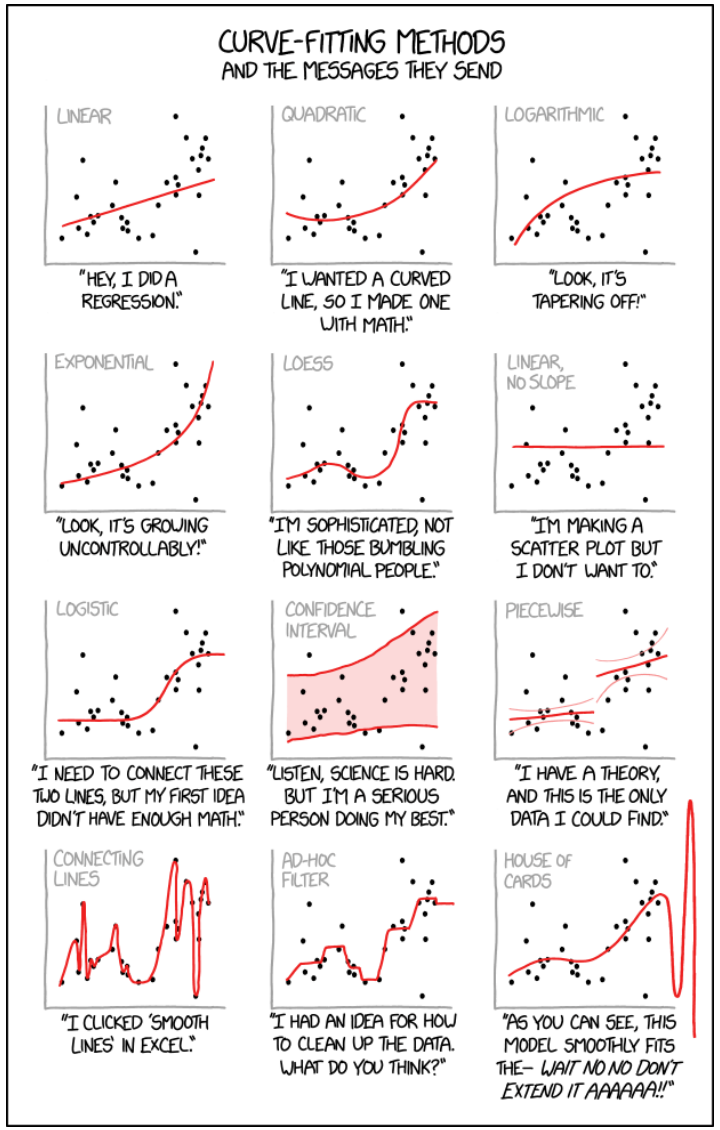
Affinity ≠ potency ≠ efficacy

QSAR

Relating a (complex) property y to observables x (descriptors)

- Correlation – causation
- Complex endpoints
- Unlimited amount of descriptors purely theoretical (PSA) to experimental (logD)

What is the relevant endpoint and what are the relevant descriptors to use?



https://xkcd.com/2048/

QSAR: Hansch, Muir and Fujita

$$\log\left(\frac{1}{C}\right) = k_i\pi + k_2\sigma + k_3$$

pi: lipophilicity

sigma: Hammentt's electronic parameter

C is the molar concentration of compound that produces a standard response (e.g., LD50, ED50)

Further it was noted that the prediction of the biological activity <u>didn't improved while using logP alone</u>, but in combination with Hammett' s sigma.

Table 1. "Classic" Parametrization

Lipophilic	Electronic	Steric
Nernst – 1891		
Overton – 1895		
Meyer – 1899		
	Hammett – 1940	
		Taft – 1956
	Taft-Lewis – 1958	
Hansch	Hansch	Hansch
	1962	
	Charton – 1964	
	Swain-Lupton – 1968	
Rekker – 1973		
Goldstein – 1974		
Seiler – 1974		
Hansch – 1975		
		Charton – 1975
		Verloop – 1976

•R.F. Rekker, The history of drug research: From Overton to Hansch, 1992

•http://www.netsci.org/Science/Compchem/feature12.html

Drug Discovery Endpoint

$$\downarrow Dose \sim \frac{IC50_u \bullet Cl\,\mathrm{int}_u}{f_{abs}}$$

Die Dosis macht das Gift!
The dose makes the poison

(Paracelsus)

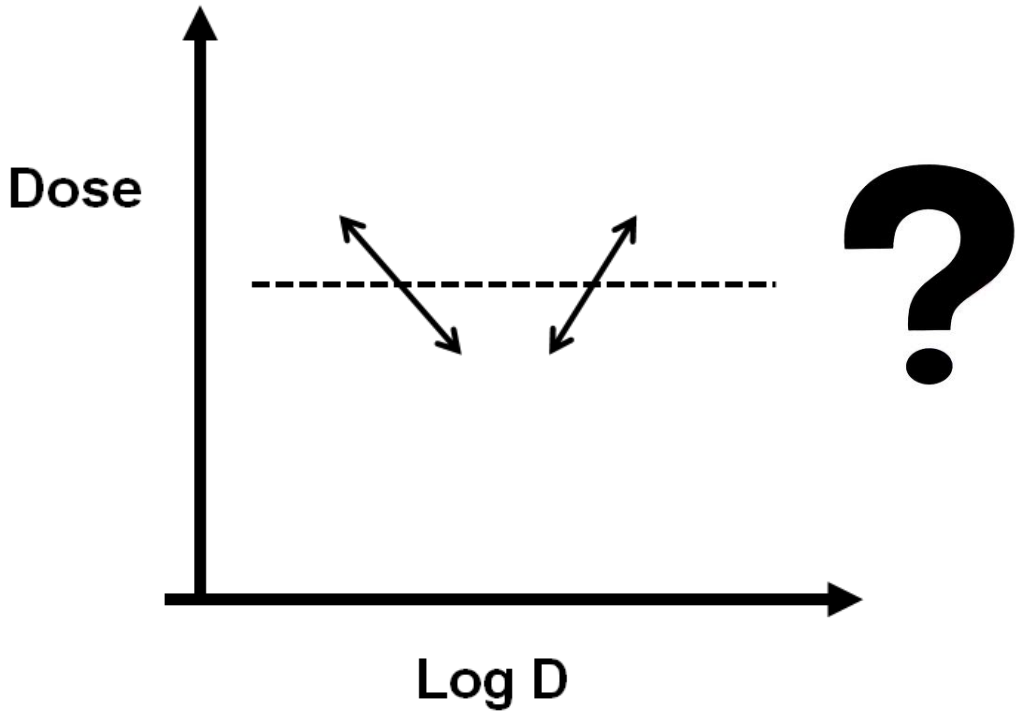
$$Halflife: C = C_o \cdot e^{-\frac{CL}{V}\cdot t} \rightarrow t_{1/2} = \ln(2) \cdot \frac{V}{CL}$$

Lowest efficacious dose with largest therapeutic index

QSAR: Surrogate descriptors and dose

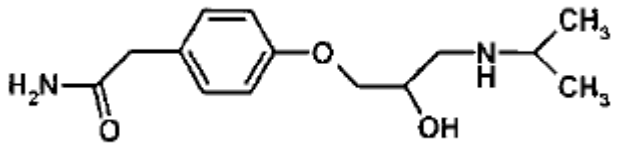
Influence of lowering logD on

- potency ⬇
- clearance ⬆
- absorption (permability ⬇ & solubility ⬆)
- Volume of distribution (Plasma Protein Binding ⬇)

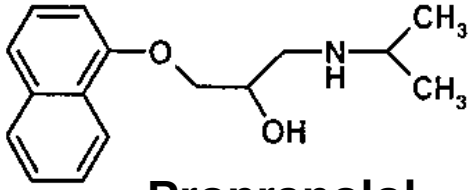


No clear understanding of how to predicit/model dose, efficacy, clearance or absorption with relevant descriptors!

logD on Dose and half-life



Atenolol



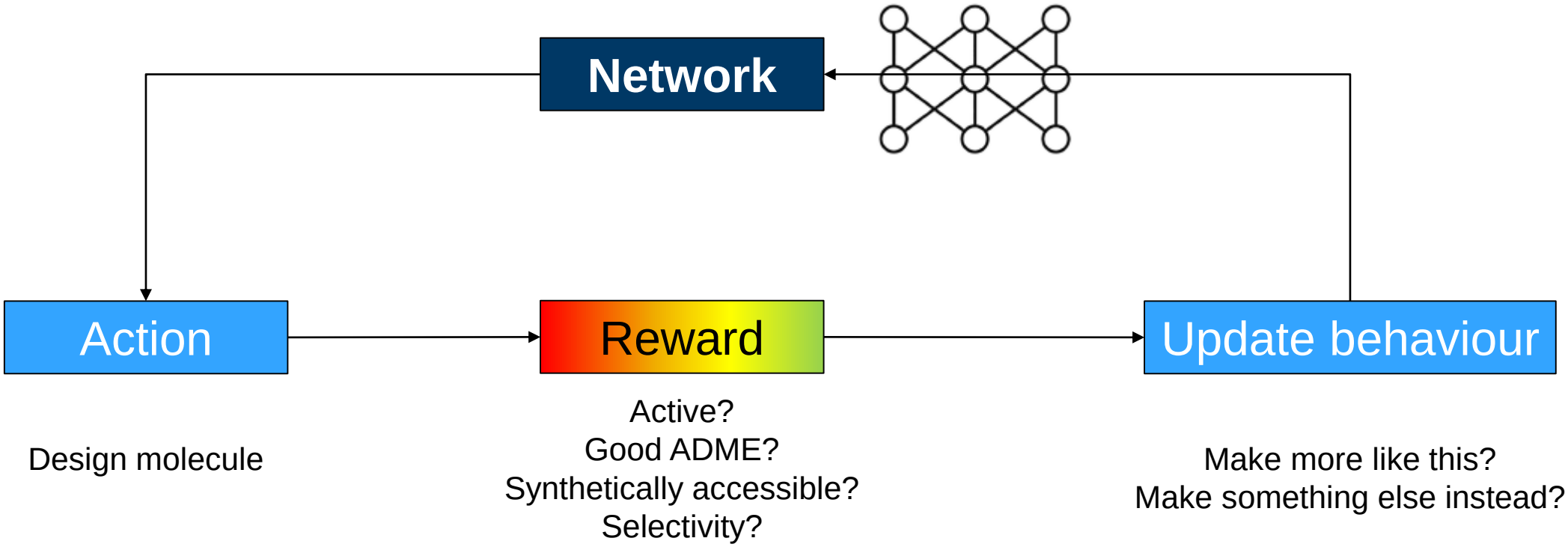
Propranolol

	log $D_{7.4}$	Affinity (pA_2)	Absorption (%)	Renal Cli(u) (ml/min/kg)	Metabolic Cli(u) (ml/min/kg)	Vol(u) L/kg	Half-life (h)	Dose (mg)
Atenolol	−1.7	6.5	50	2	–	1	3–5	50–100
Propranolol	1.2	8.3	100	–	470	50	3–5	30–90

- increasing lipophilicity raises Vdu, Cli(u) and potency, effectively cancelling out any changes in half-life or steady state concentration and dose

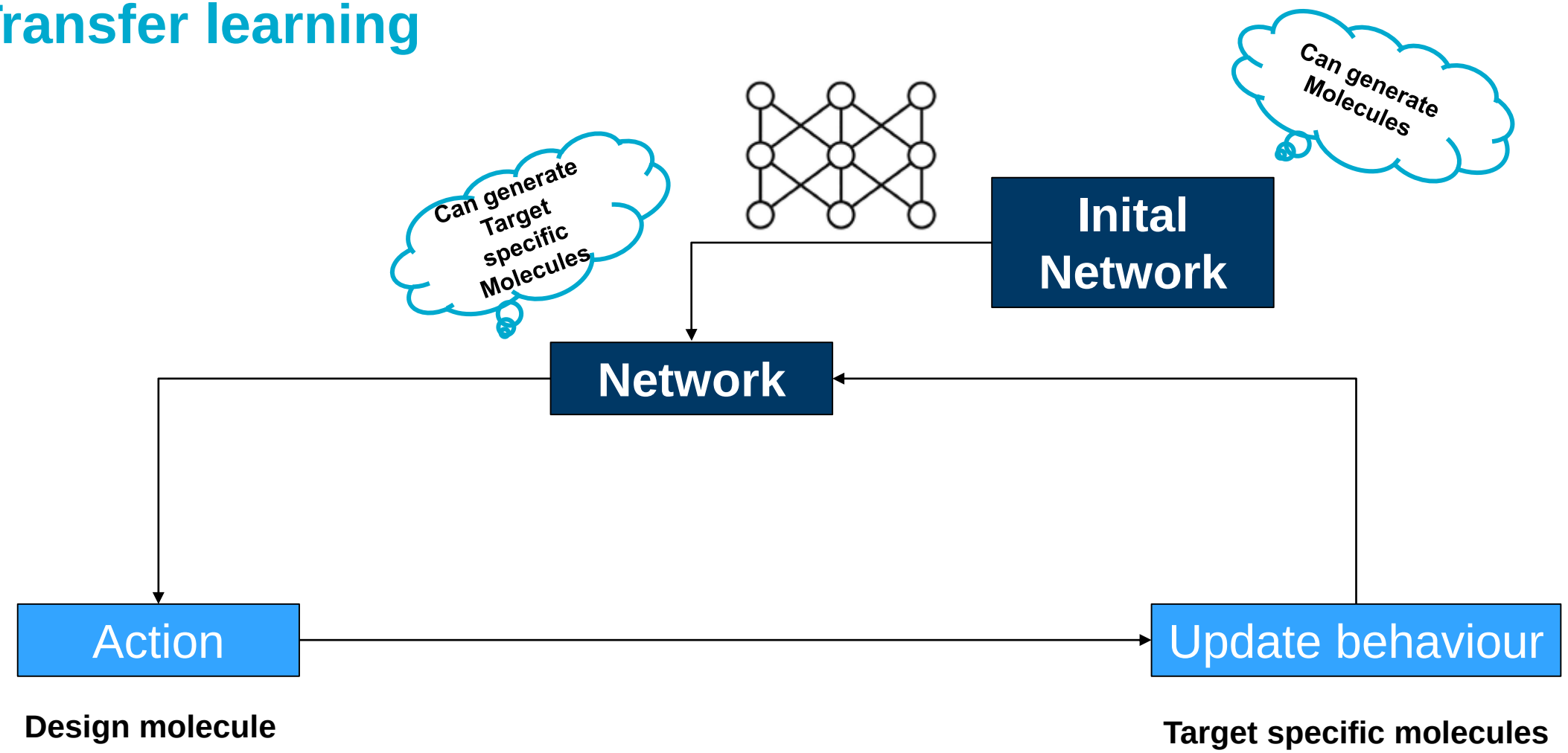
H. van de Waterbeemd et al Lipophilicity in PK design: methyl, ethyl, futile, 2001

AI: Reinforcement learning



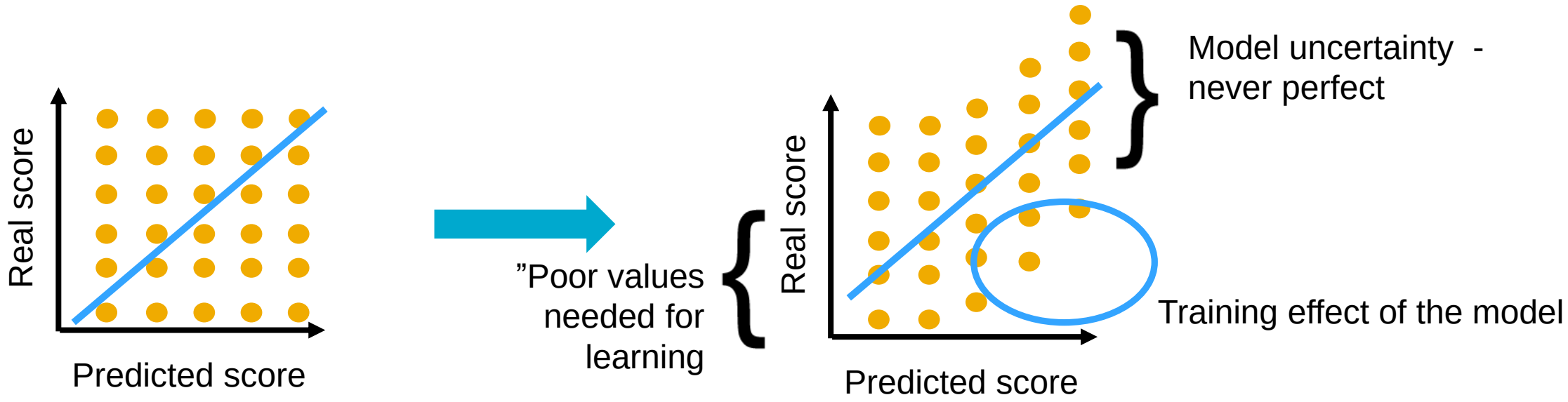
- Learning from doing
- Often use pre-trained model as a starting point
- Reward function? Weakness

AI: Transfer learning



- Molecule generator will be retrained to be task specific
- Need of a high quality set of relevant compounds (late LO)

Deep network (or MPO/QSAR) training

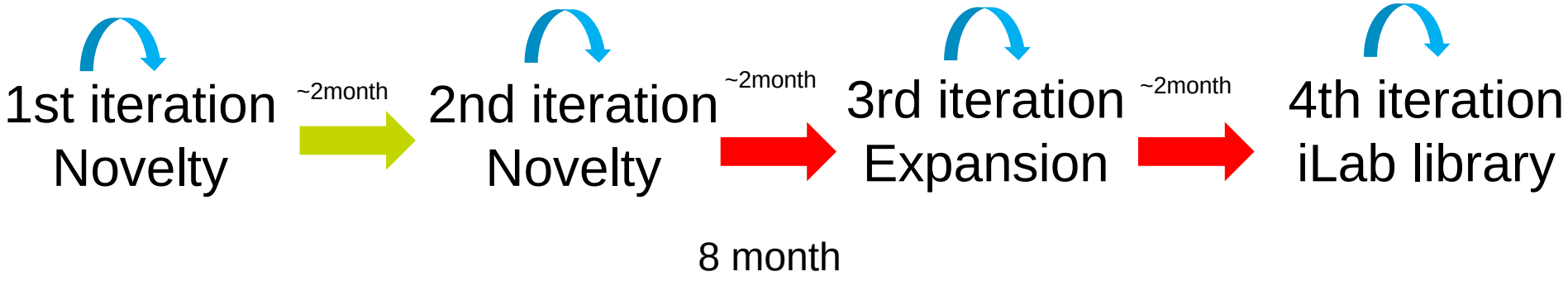


Incorporation of positive & negative data in training for DL?
QSAR model quality?
False positive problem – What do next?

Real World

Three test cases prospective, augemented design in on-going projects

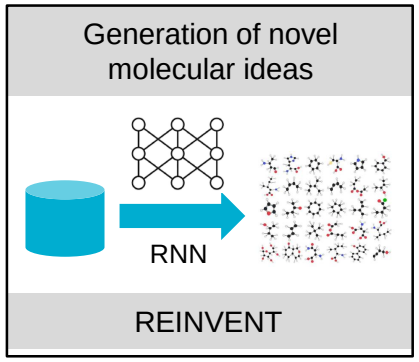
- Lead identification –based on large amount of late stage data (>5000 data points) BUT **novelty with LO quality**,
- Lead identification – **selectivity LO quality**, based on bigger corpus of target class data
- Lead identification – **novel LI series with LO quality**, based on Hit Finding readouts (e.g. HTS)

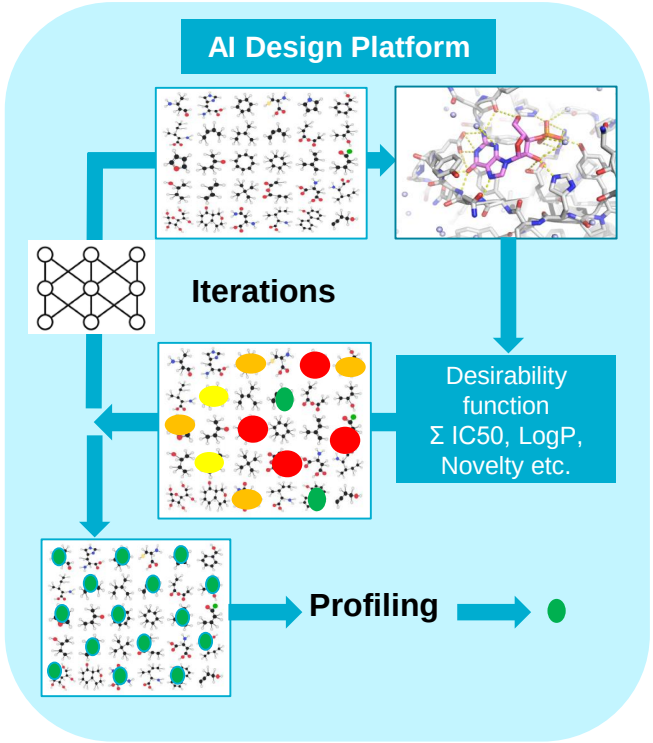


Constant re-learning and training

Real World

Achievement: Automated AI work flows identified novel, potent compounds





Target1	Novel compounds with ~100nM potency
Target2	Novel, potent series (<10nM)
Target3	Novel, potent compounds with good DMPK

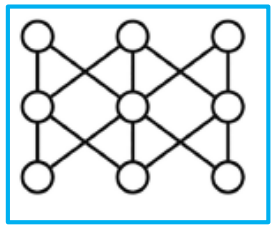
Deep learning use case: The models

Activity Prediction	Property Prediction – Desirability Score	RNN
• SVM model with ECFP6 • Actives pIC50 >7 • ~10.000 compounds (~2000 test set) • ROC AUC testset is 0.99 • Validation with 226 new compounds (ROC AUC 0,97, Accuracy 0.89)	• Solubility • Clearance • Permeability • AZLogD • MW, cLogP, PSA, RotB • AZfilters **Dscore** = $(P^{a}_{1} \times P^{b}_{2} \times P^{c}_{3}\ldots)^{1/(a+b+c+..)}$ a, b, c,… = weight	• **Transfer learning** • **Reinforcement learning** • **Scaffold biased reinforcement learning**

D.J. Cummins and M.A. Bell, J. Med. Chem, 2016,59, 6999-7010

Filtering down the AI generated compounds

AI machine



800K cpds

Reinforcement learning 500K cpds

Reinforcement Learning, scaffold 80K cpds

Transfer learning 200K cpds

368K cpds (TI ECFP6 sim. <=0.4)

65K cpds
(Activiy >=0.8 & Dscore >=0.7)

58K cmpds Glide docking >-7

Similarity Filter

Act. + Desirability score

Docking score

Novelty & Synthesis

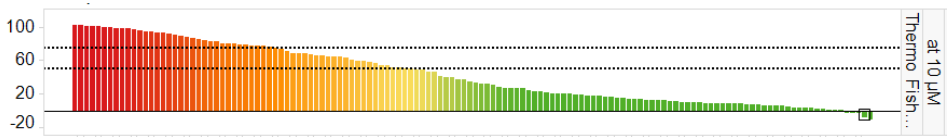
Diversity selection

Design set for Med. Chem.

First Iteration: Results

	Structure	Structure	Structure	Structure	Structure	Structure
	AZ1	AZ2	AZ3	AZ4	AZ5	AZ6
Target pIC50 (enzyme)	7.2	6.7	6.2	<5	7.0	6.9
logD	3.0	2.5	3.1	1.9	2.7	2
Clint (HuMics)	<3	9.5	29.8	92	26	44
Solubility (DMSO)	5	127	294	1000	12	116

Kinase selectivity profile

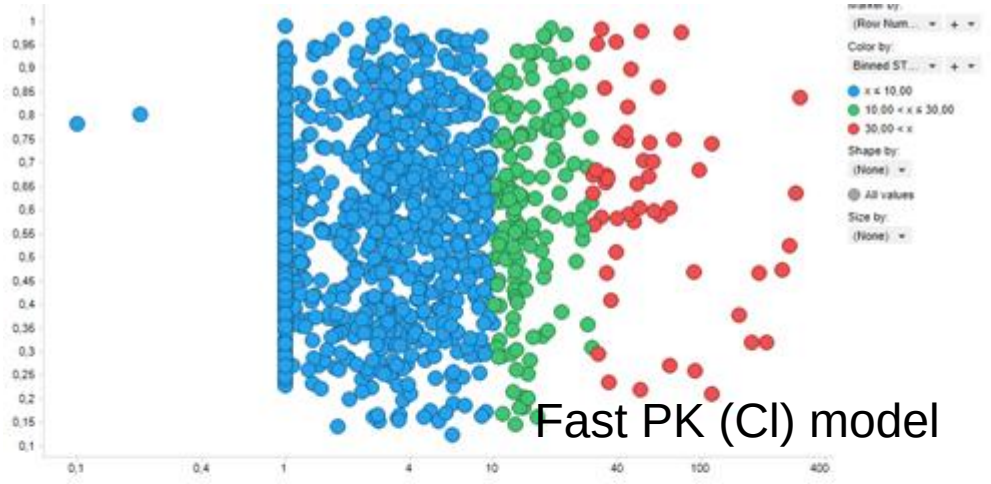


Only compound in training set with scaffold

Closest known analogue, not in the training set

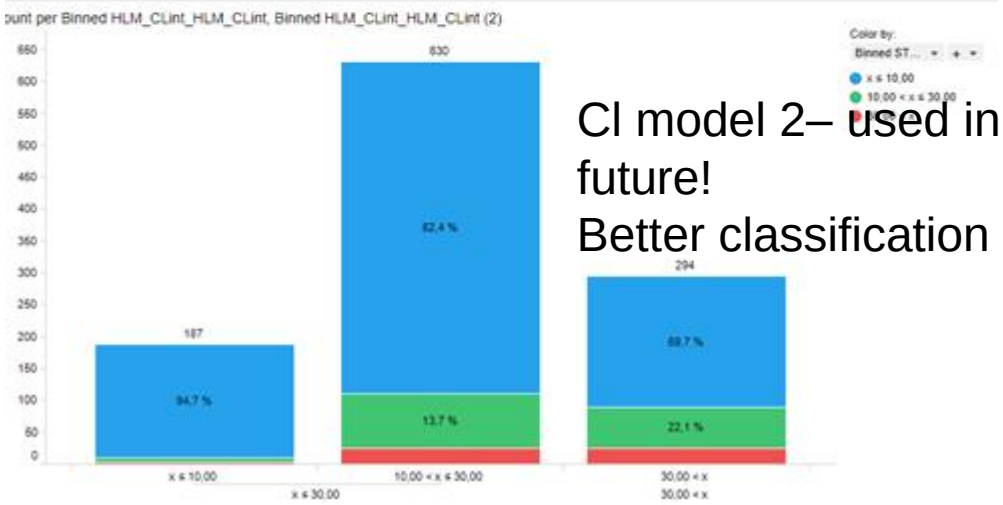
Our PK models

Predicted Cl



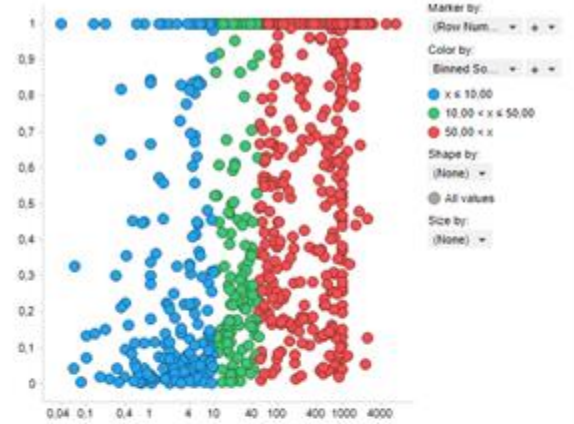
Fast PK (Cl) model

Count

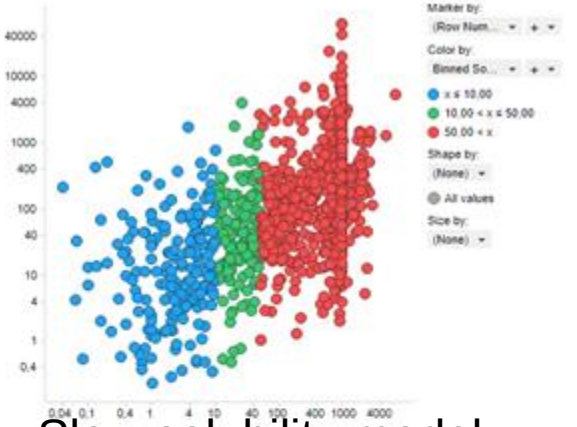


Cl model 2– used in the future!
Better classification <10

Predicted Sol

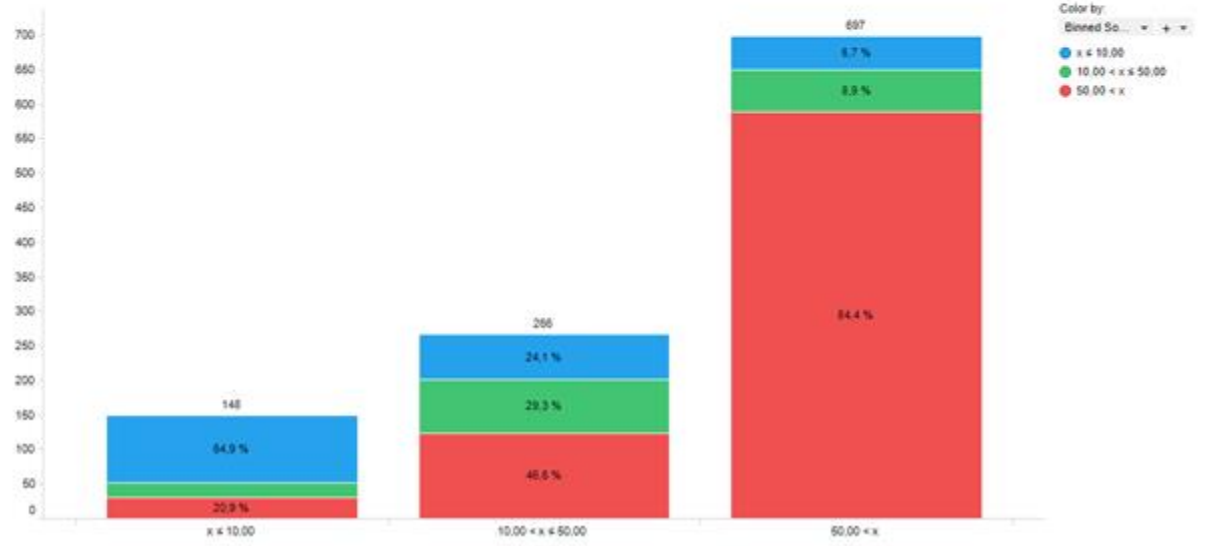


Fast solubility model



Slow solubility model

Count



Slow solubility model binning – better classification <50uM

Outlook & Challenges

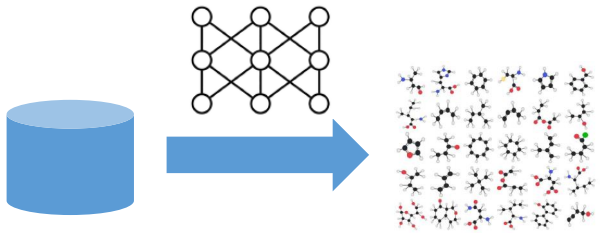
- QSAR models
  - Affinity model ~20% correctly classified
  - The fewer models the better?

- Synthesis is slow – every cluster is novel (all compounds could be made!) [29 in total, 6 different and novel clusters, 3-15 weeks]
  - Reaction prediciton based on available reagents
  - Automatic synthesis

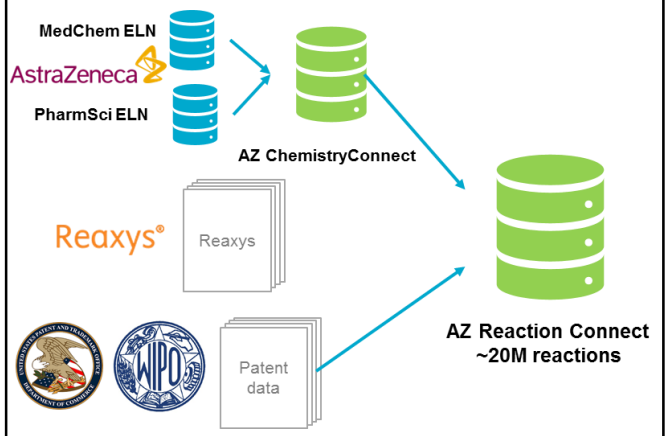
- Need to run significant more cycles with smaller changes (applicability domain!)

AI Guided Drug Discovery Platform - What is Required?

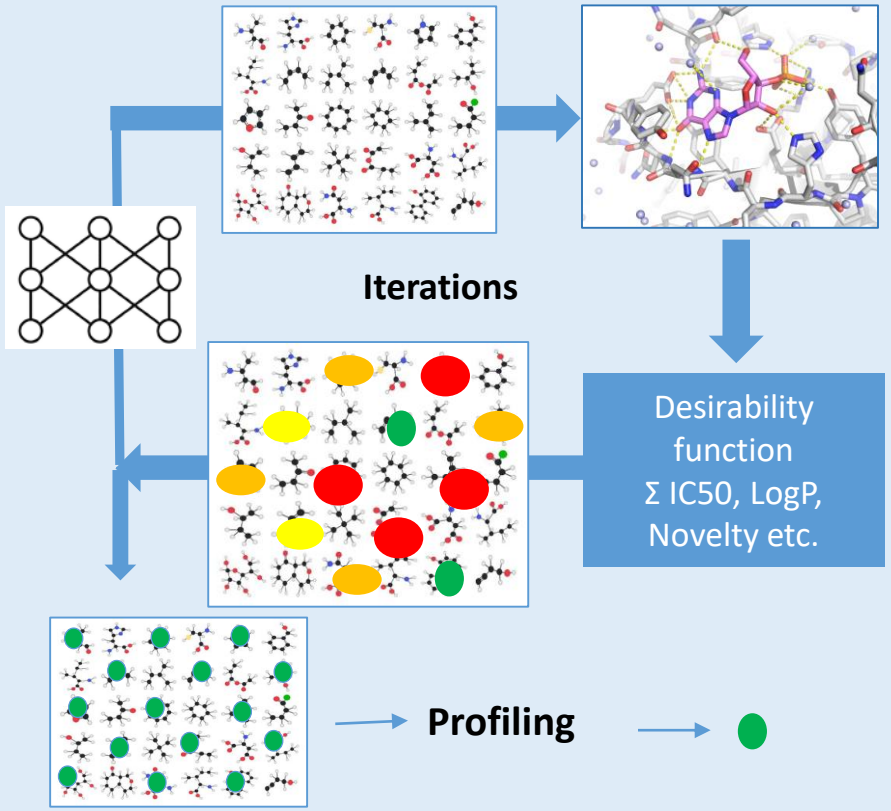
Generation of Novel Chemical Space



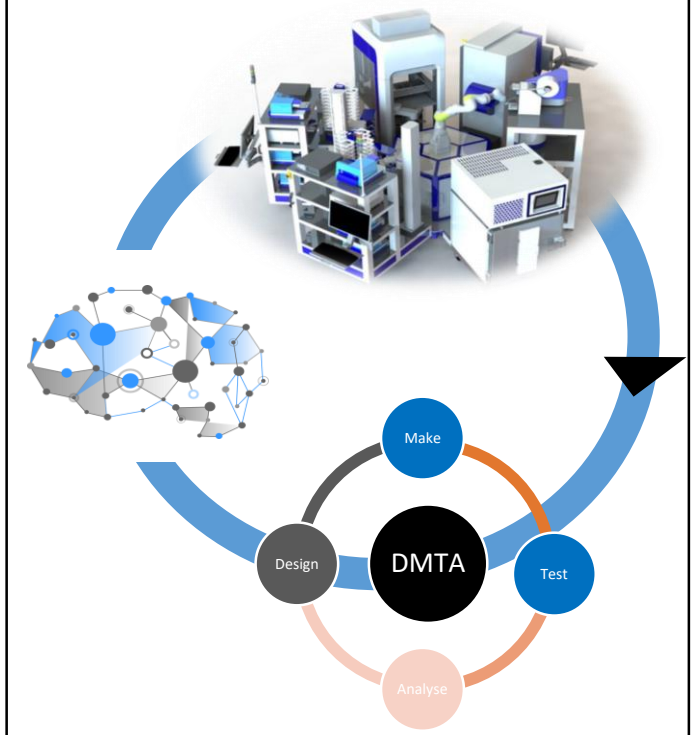
Reaction & Synthesis Prediction



AI Design Platform



Fully Automated DMTA Cycle



Reaction informatics: Research Questions

Three Main Questions

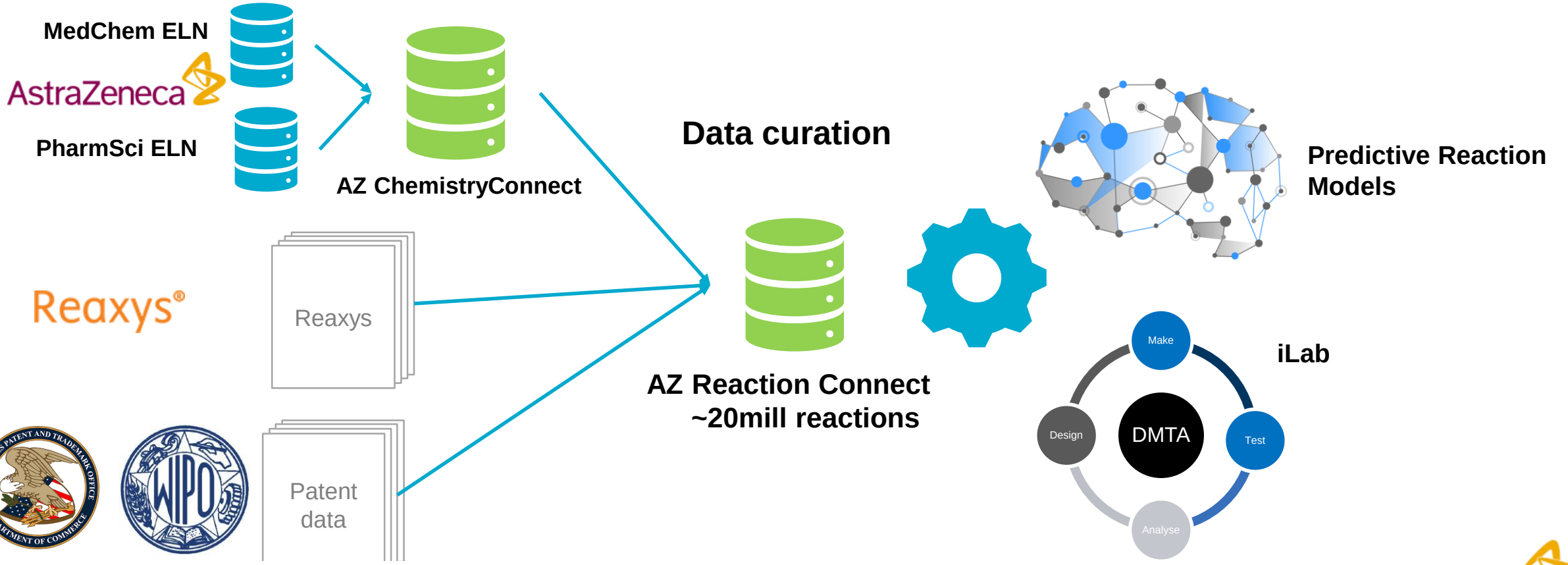
1. How feasible is a given reaction?

2. Given a set of starting materials, what are the most feasible and likely reaction pathways (forward synthesis)?

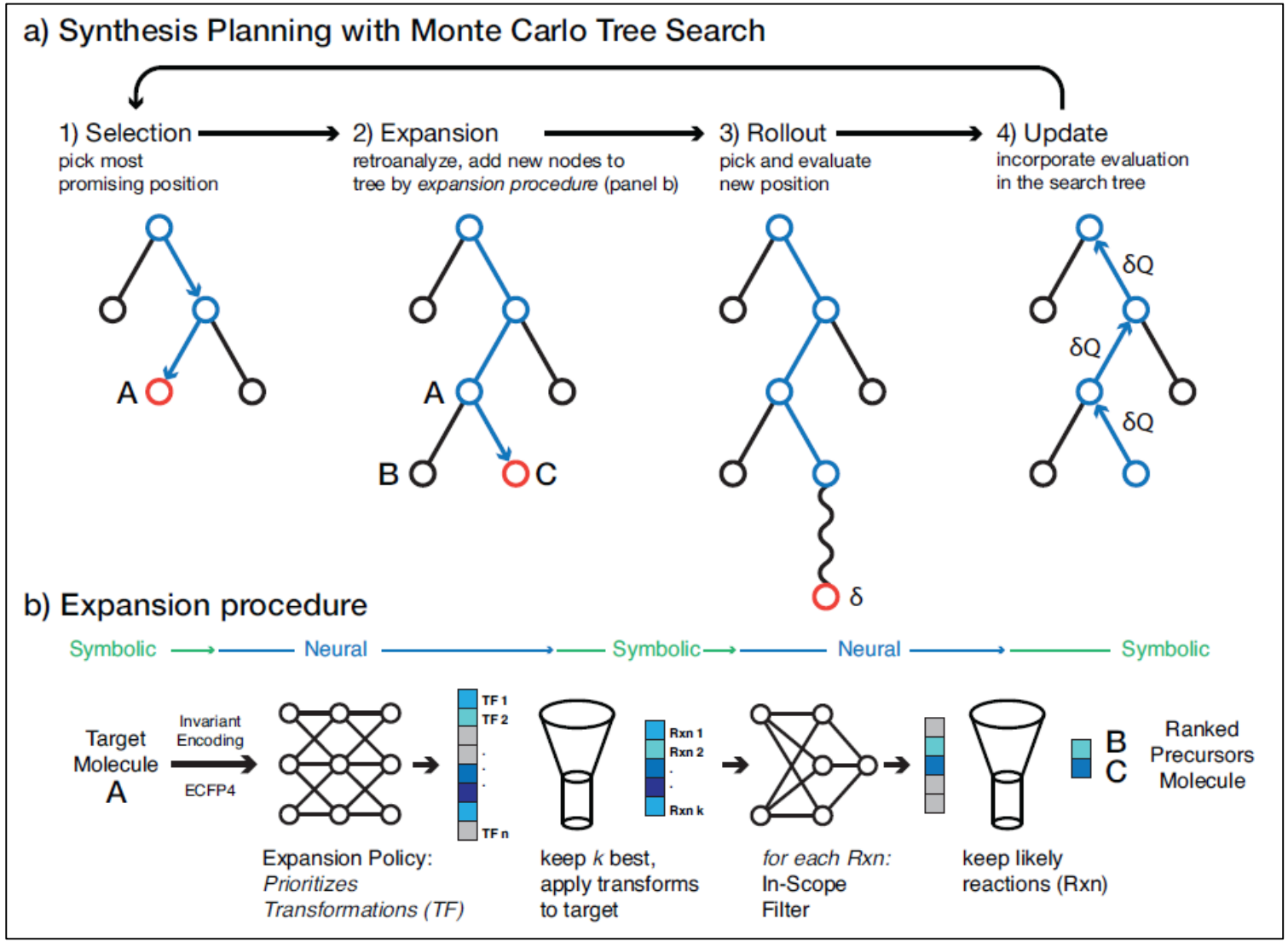
3. What are the most likely reaction pathways identified by retrosynthetic analysis resulting in a successful synthesis of a target molecule (retrosynthesis)?

Deep learning at AstraZeneca: Reaction informatics

- First steps, building:
  - World-class Reaction Knowledge Base
  - On our work (past collaboration with M. Segler)
  - Support RI, LSF PostDocs



State-of-the-art AI solution

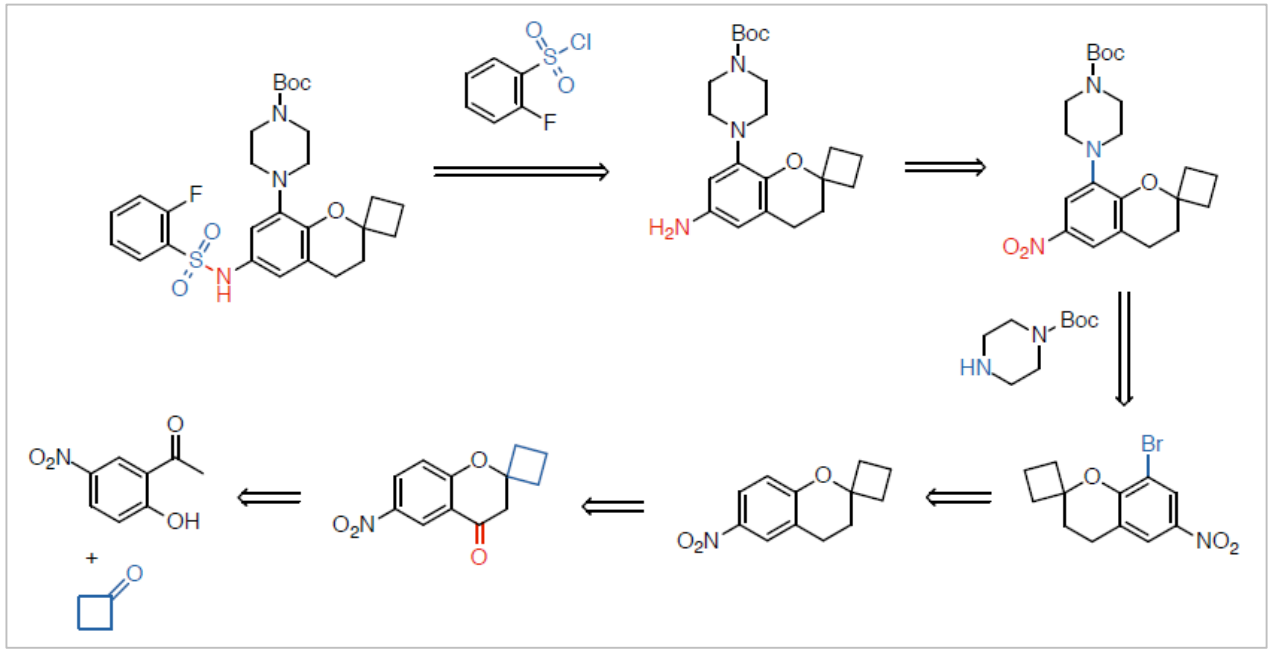


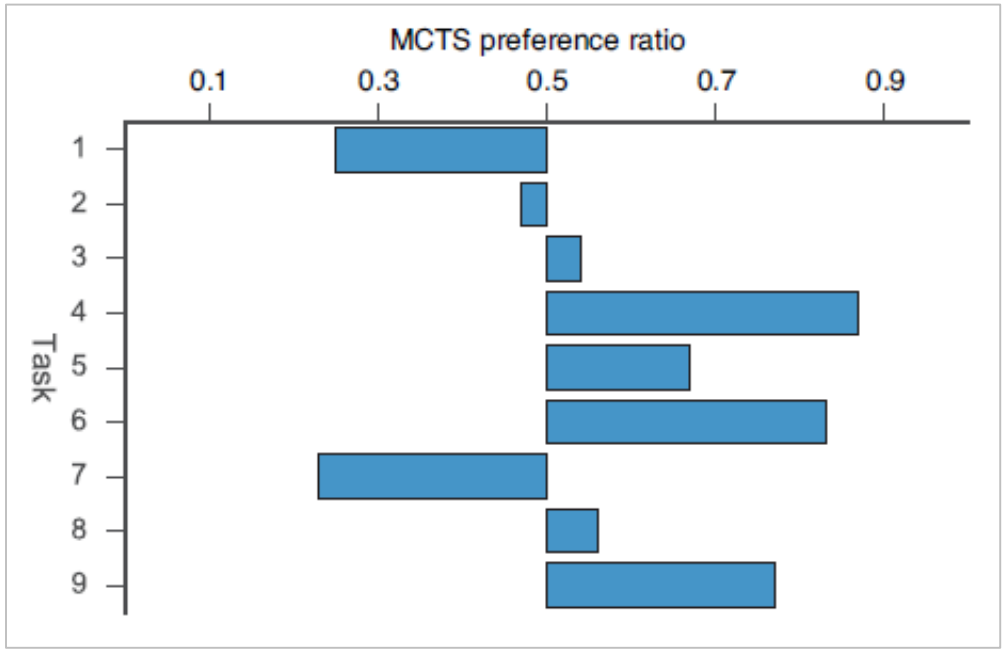
Key features

- Guided into promising directions by proposing a **restricted number** of possible, automatically extracted transformations
- **Predicts** whether the corresponding **reactions** are actually feasible
- Networks were trained on **12.4 million reactions** from the Reaxys database

Learning to plan chemical synthesis; Marwin H. S. Segler, Mike Preuss and Mark P. Waller; arXiv, 2017 arXiv:1708.04202

State-of-the-art in chemical synthesis planning





- Model finds a 6-step synthesis route to the intermediate drug candidate autonomously in **5.4 seconds**
- The route is **identical** to that originally published in 2015 – the published route was not part of the test set

- Chemists **did not** significantly prefer literature routes over routes found by the model
- Unsolved?: **natural product** synthesis and stereochemistry

Learning to plan chemical synthesis; Marwin H. S. Segler, Mike Preuss and Mark P. Waller; arXiv, 2017 arXiv:1708.04202

ELN Data

Reaction Normalisation & Validation

- Covers only ELN data (in-house reactions)!
- Reactions need to be annotated and cleaned
- Reaction normalisation and validation
  - Inconsistent component-molecule definitions
    - CC_ID = null (~210,000 / ~5.3%) and/or MOLECULE_ID = 0 (~700,000 / ~17.5%)
- Old 3rd party software for reaction type classification with limited license
- Reduce naming inconsistencies
  - Registered as ”sodium” but should be ”sodium triacetoxyhydroborate”
  - Same entitiy registered as different “molecules” in ChemConnect (sodium;hydroxide, sodium hydride;hydroxide, hydroxysodium)
- Different molecules are linked to the same representation
  - CC_ID = 200003711 is registered with 75 different molecules (~ 3,700 representations)
- Problems with extracting reactions

… many more inconsistencies!

Overview

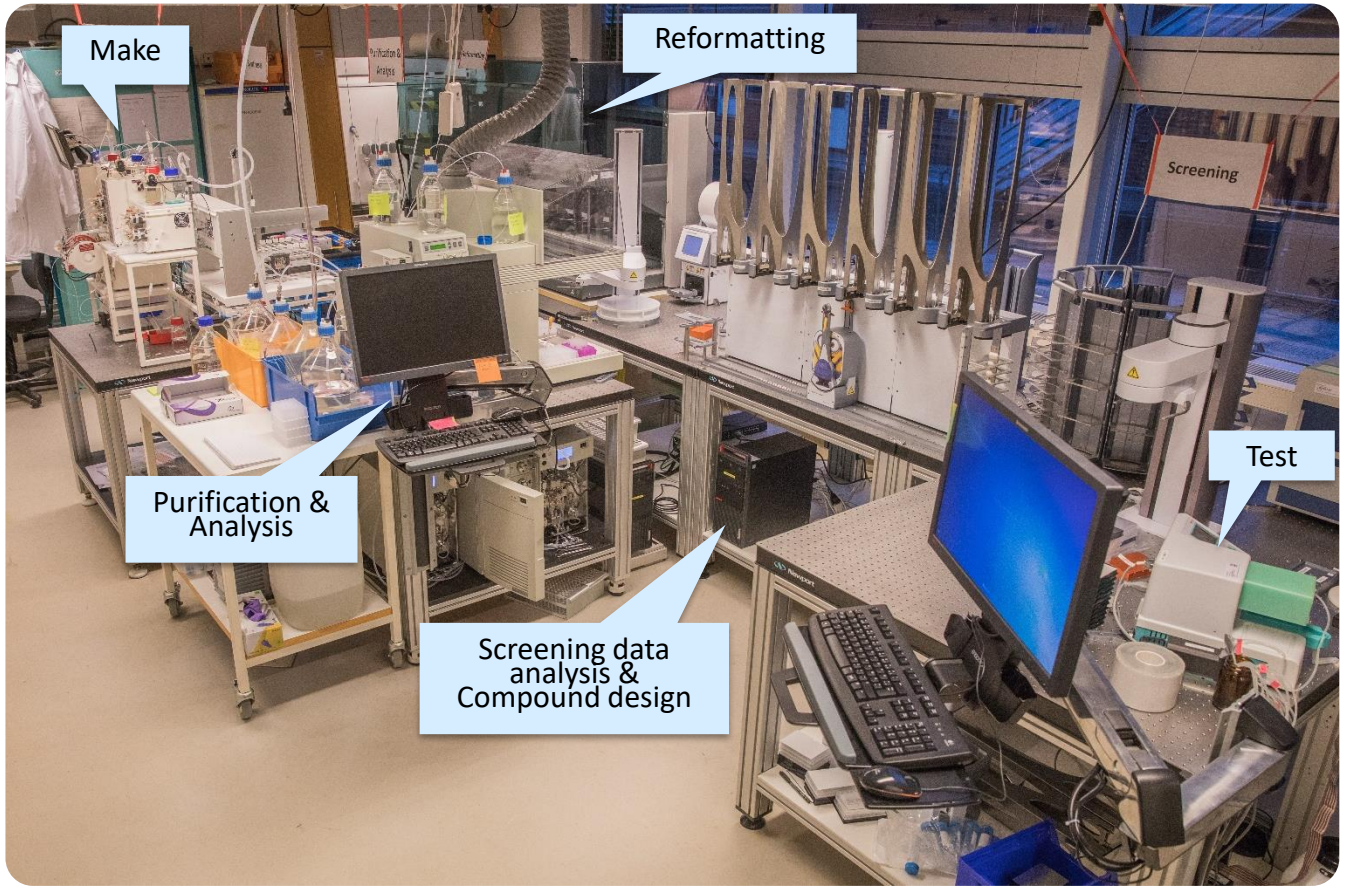
Number of Instances (2017-09-24)

Type	# unique
Reaction type	298
Reaction	658,224
Component	3,994,278
Molecule	663,058
Variation	1,113,932
Stage	1,113,934

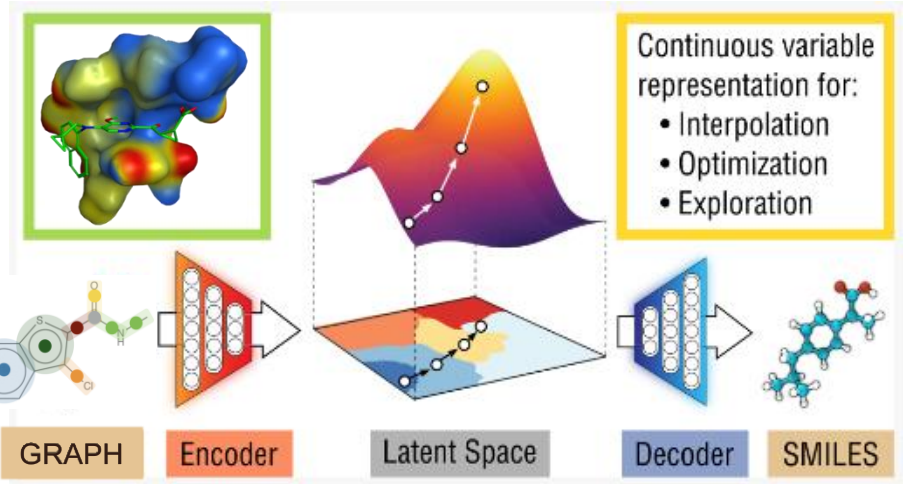
Current state, not updated, ELN+ ***HazELNut*** classification

AZ's first DMTA automation platform

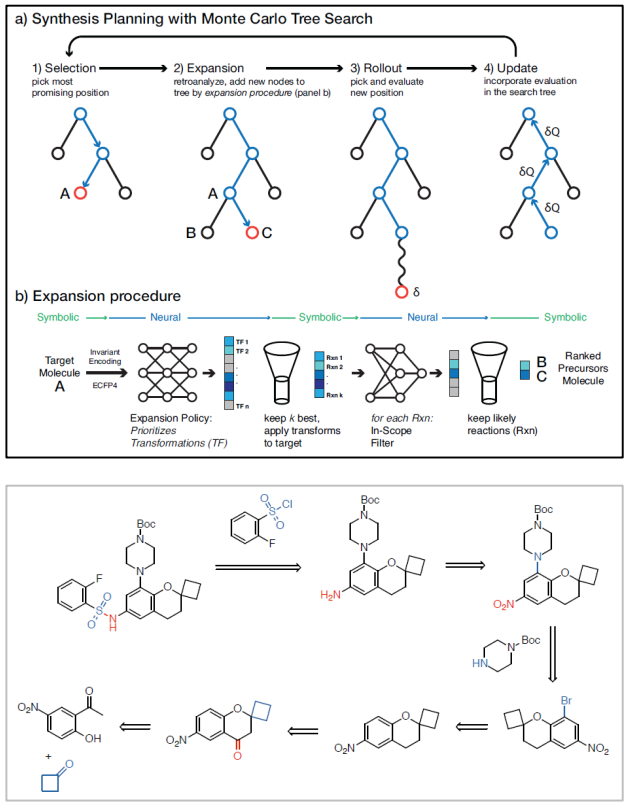
- First protype built during 2017
- All DMTA steps fully integrated
- Suited for 100s of uninterrupted DMTA cycles
- Cycle times of ca. 2h
- Successfully applied in ongoing research project



Deep Learning: Outlook



Chemical space exploitation



Synthesis Prediction

Autonomous design

R. Gómez-Bombarelli et al., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, ACS Central Science 2018 4 (2), 268-276

The core team

Discovery Sciences
Ola Engkvist
Hongming Chen
Thierry Kogej
Clive Green
Gary Pairaudeau

RIA
Ina Terstiege
Igor Shamovsky
Christian Tyrchan
Werngard Czechtizky

PhD/PostDoc
Thomas Blaschke
Josep Arus Pous
Marcus Olivecrona