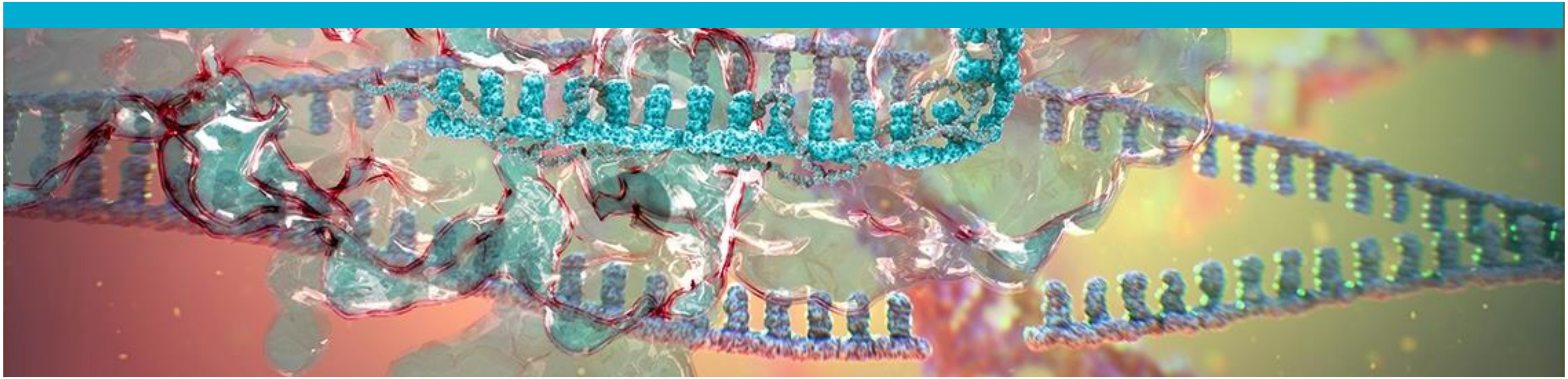
AstraZeneca
IMED Biotech Unit

Machine Learning and AI for Drug Design

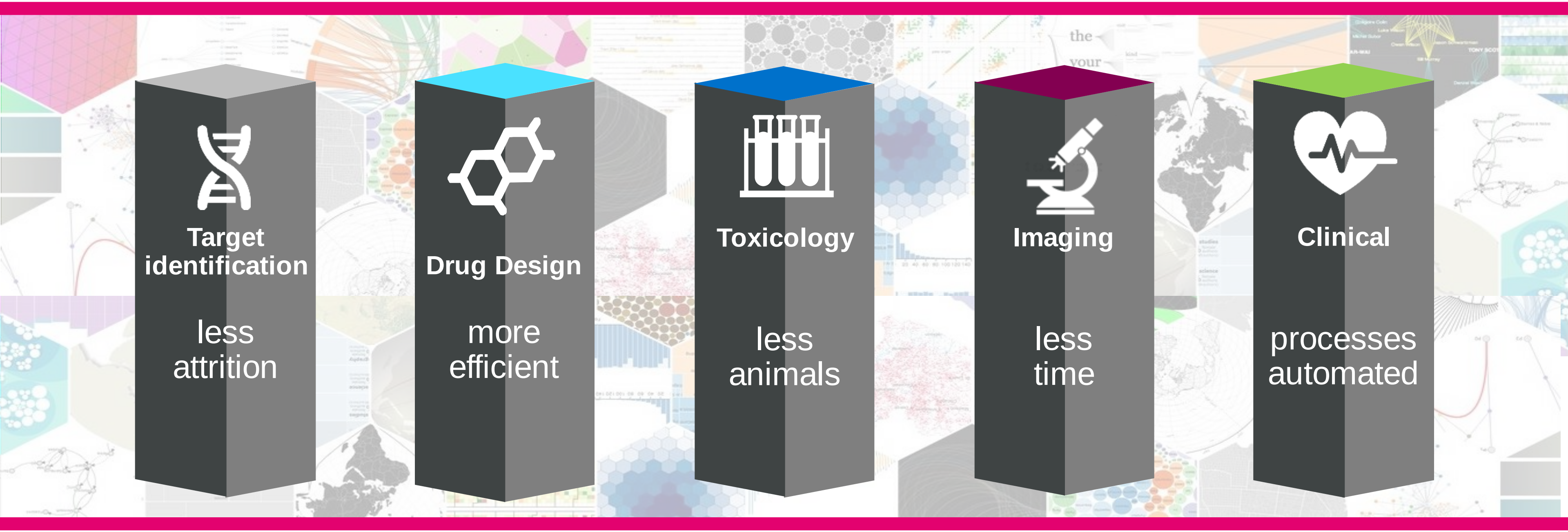
Ola Engkvist, Molecular AI, Discovery Sciences, R&D, Gothenburg, Sweden

Academy of Pharmaceutical Sciences Virtual Seminar

January 20 2022



Where can AI impact drug discovery and development



Target identification

less attrition

Drug Design

more efficient

Toxicology

less animals

Imaging

less time

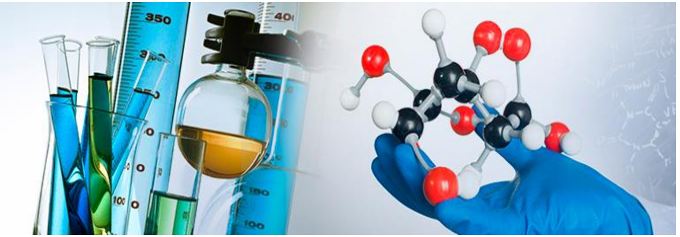
Clinical

processes automated

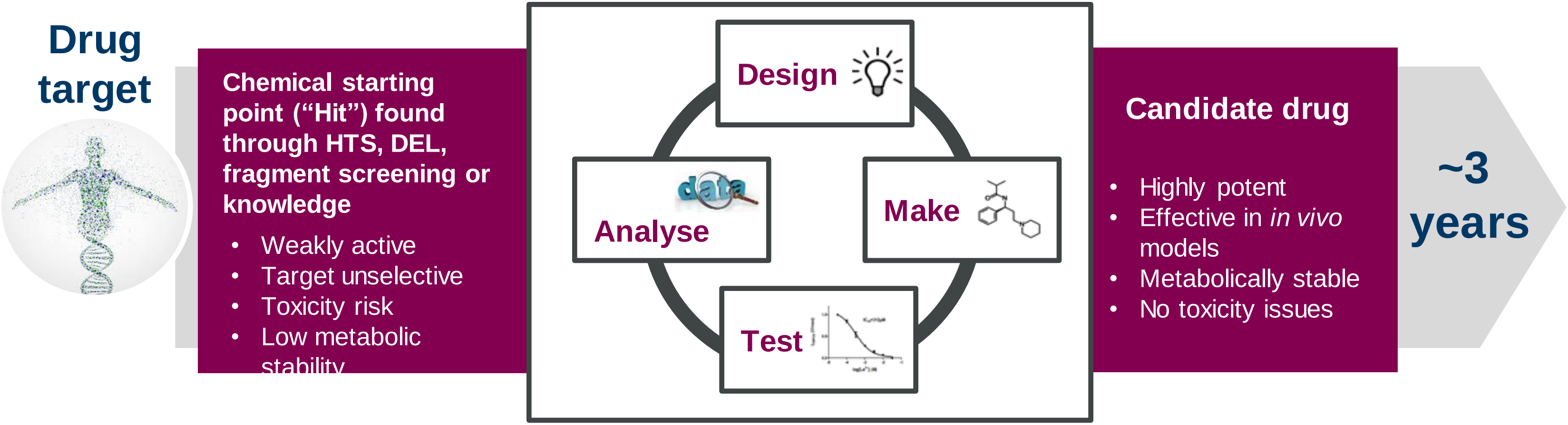
Drug Design

Which compound to make next?

How to make the compound?



The Design Make Test Analyze cycle in Drug Design



Drug target

Chemical starting point (“Hit”) found through HTS, DEL, fragment screening or knowledge

- Weakly active
- Target unselective
- Toxicity risk
- Low metabolic stability

Candidate drug

- Highly potent
- Effective in *in vivo* models
- Metabolically stable
- No toxicity issues

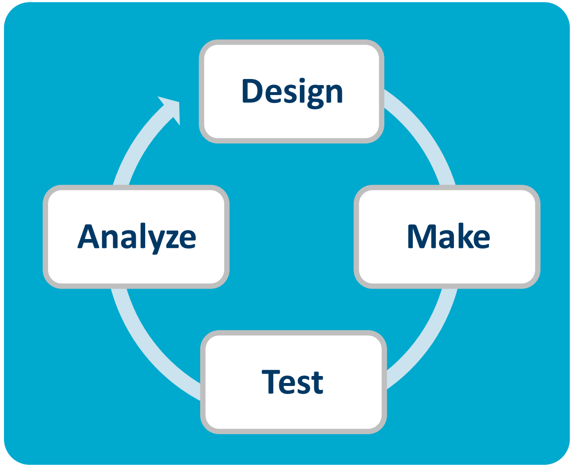
~3 years

Multiple of DMTA cycles

AI based drug design

How can we reduce the time to deliver a clinical candidate?

Select the most efficient synthetic route



Make information rich compounds in each cycle

Increase speed

Maximize learning

Why now?

Why would this presentation have been science fiction 5 years ago?

- **Increased computational power**

  Never underestimate an exponential law

- **Advances in neural network algorithms**

  New algorithms in other fields that can be adapted to our needs i.e. Image recognition, Natural language processing, Playing Go

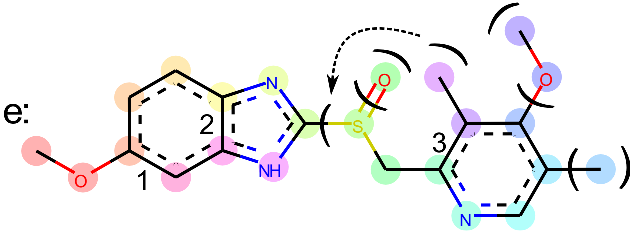
- **Open-source software**

  Python, RDKit, scikit-learn, PyTorch, Tensorflow

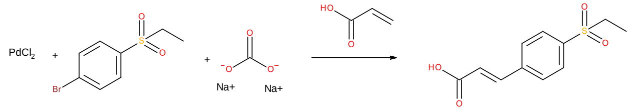
How can we take advantage of the progress in Natural Language Processing?

Molecules can be described with the language SMILES

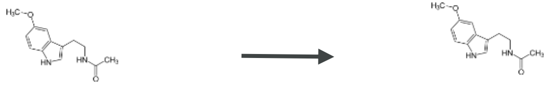
Molecule:



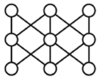
SMILES: COc1ccc2nc(S(=O)Cc3ncc(C)c(OC)c3C)[nH]c2c1



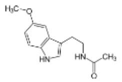
Language Translation ⟷ Synthesis prediction



Language Translation ⟷ Molecular optimization

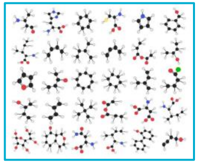


CC(=O)NCCc1c[nH]c2ccc(OC)cc12

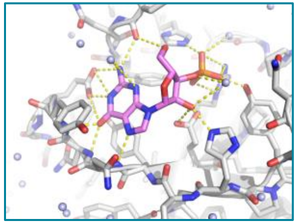


Text generation ⟷ Chemical space exploration

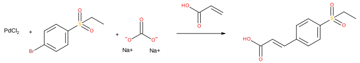
What can we do now with AI that is different?



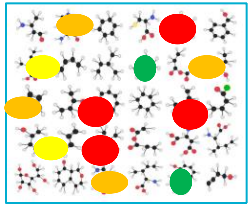
✓ AI generated ideas from the whole relevant chemical space to find novel active molecules



✓ Better prediction of synthetic routes through new algorithms

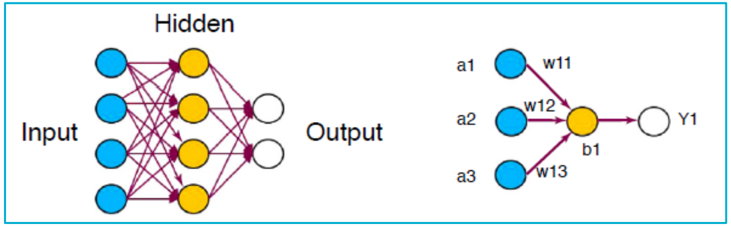


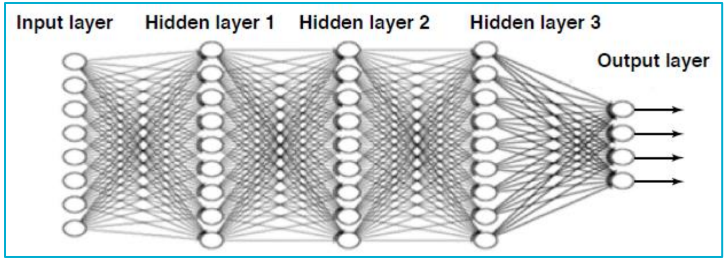
✓ Novel and more flexible ways of predicting molecular properties

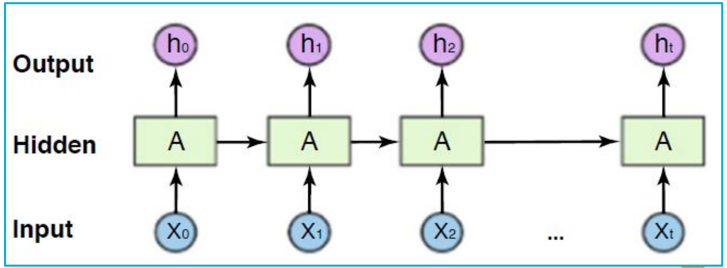


Neural Networks & Deep Learning

- **Neural Networks known for decades**
  - Inputs, Hidden Layers, Outputs
  - Single layer NNs have been used in QSAR modelling for years
- **Recent Applications use more complex networks such as**
  - Multi-layer Feed-Forward NNs
  - Convolutional NNs
    - biological image processing
  - Auto-encoder NNs
  - Recurrent NNs
    - Trained using Maximum Likelihood Estimation to maximize the likelihood of next character







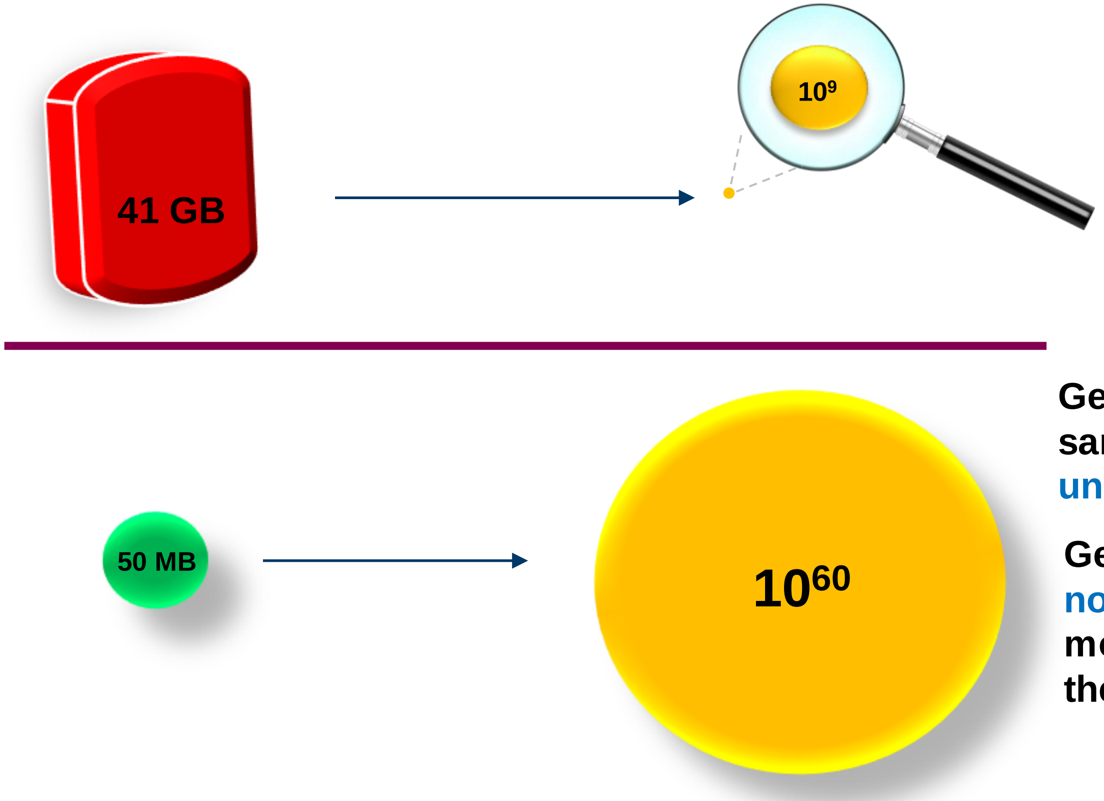
Generative Model vs Enumeration for molecular discovery

Physical Storage Size

Size of Molecular Space

Traditional Enumeration

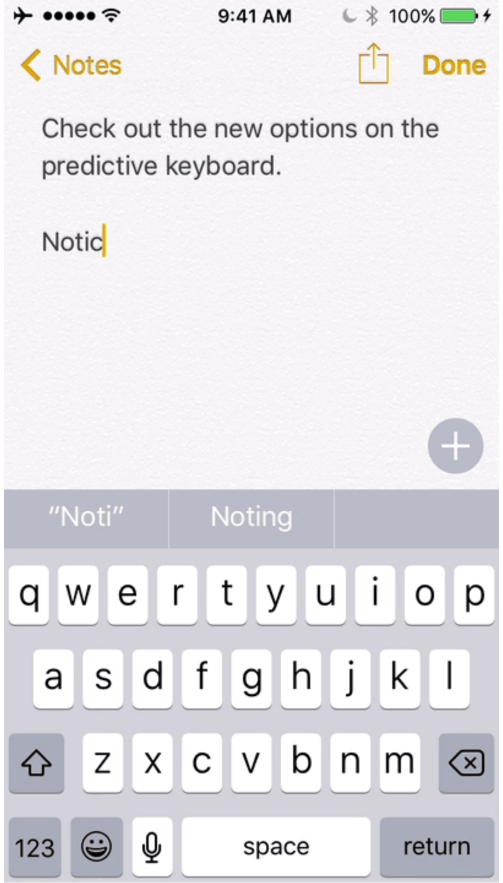
Generative Model



Generative models can sample practically unlimited chemical space

Generative models do not contain any explicit molecules but generate them probabilistically

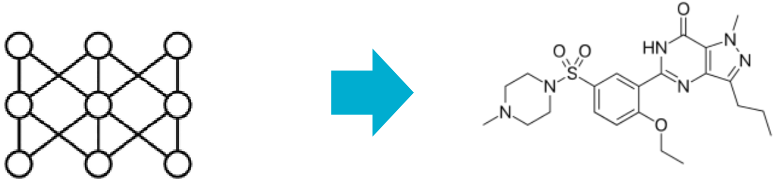
Recurrent Neural Network & Natural language generation



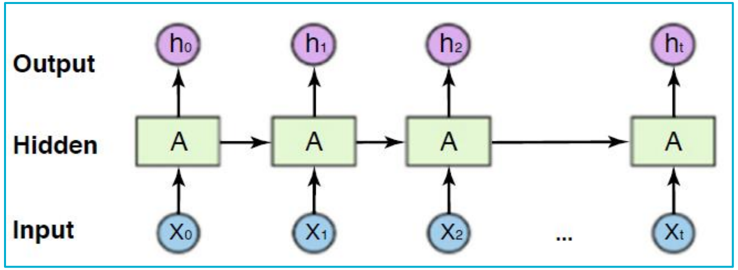
Check out the new options on the predictive keyboard.

Notic

Two different ways how can AI help finding the next molecule to make?



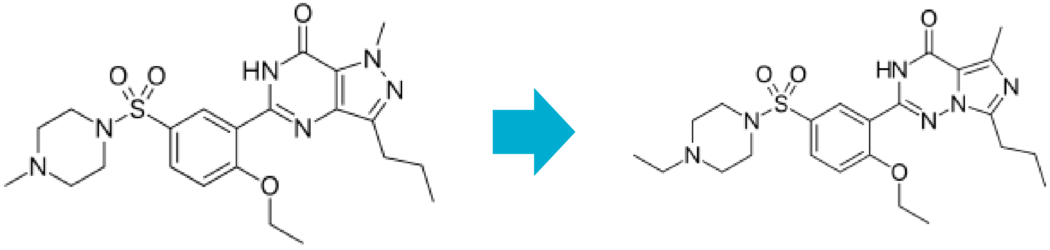
Recurrent Neural Networks

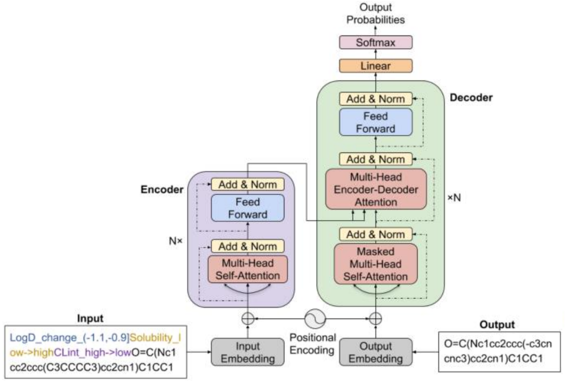


Hit Finding & scaffold hopping
Sample the whole chemical space

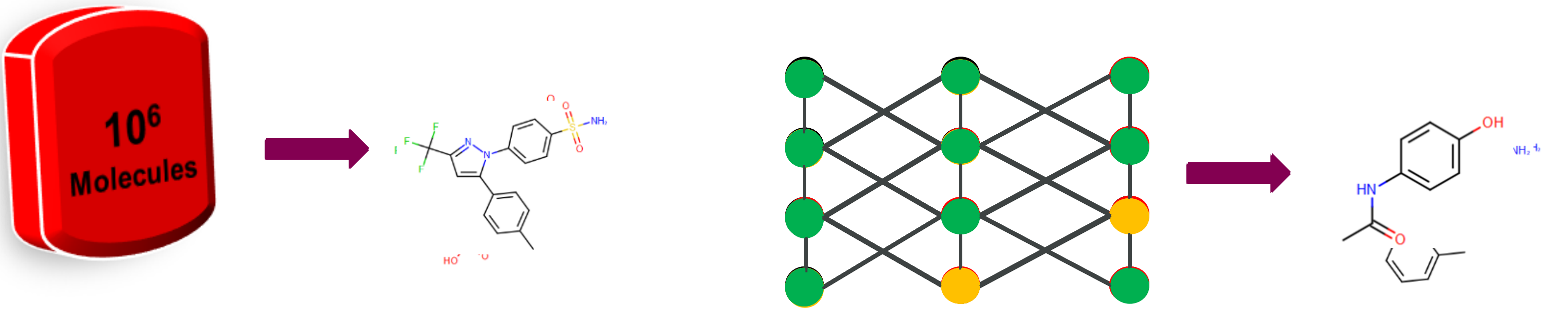
Transformer

Molecular Optimisation
Sample a focused chemical space



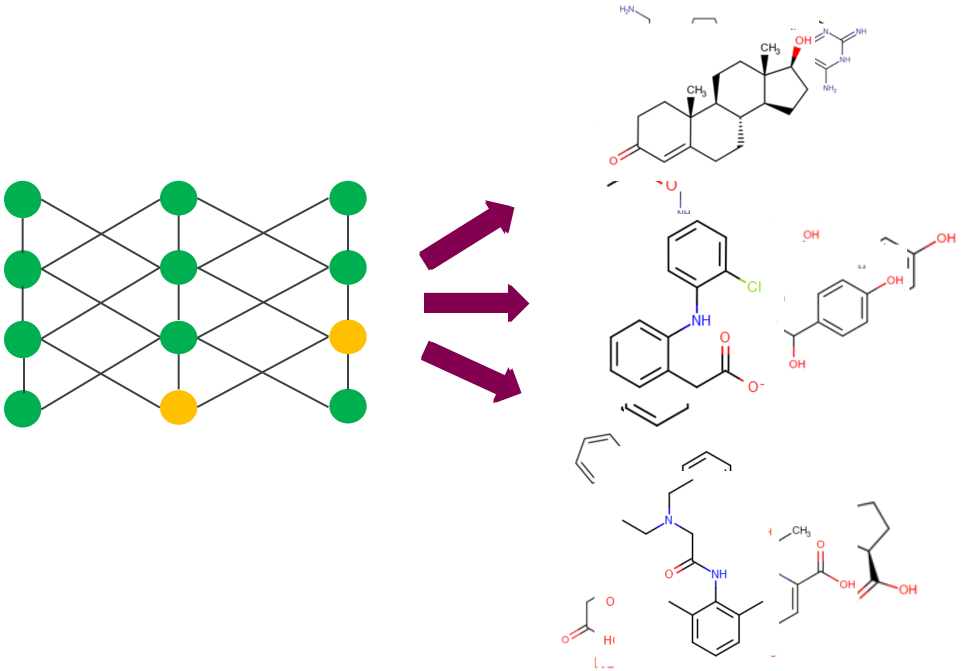

Output Probabilities
Softmax
Linear
Decoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Encoder-Decoder Attention
×N
Add & Norm
Masked Multi-Head Self-Attention
Encoder
Add & Norm
Feed Forward
N×
Add & Norm
Multi-Head Self-Attention
Input
LogD_change_(-1.1,-0.9]Solubility_low->highCLint_high->lowO=C(Nc1cc2ccc(C3CCCC3)cc2cn1)C1CC1
Input Embedding
Positional Encoding
Output Embedding
Output
O=C(Nc1cc2ccc(-c3cncnc3)cc2cn1)C1CC1

Training an RNN to generate novel molecules



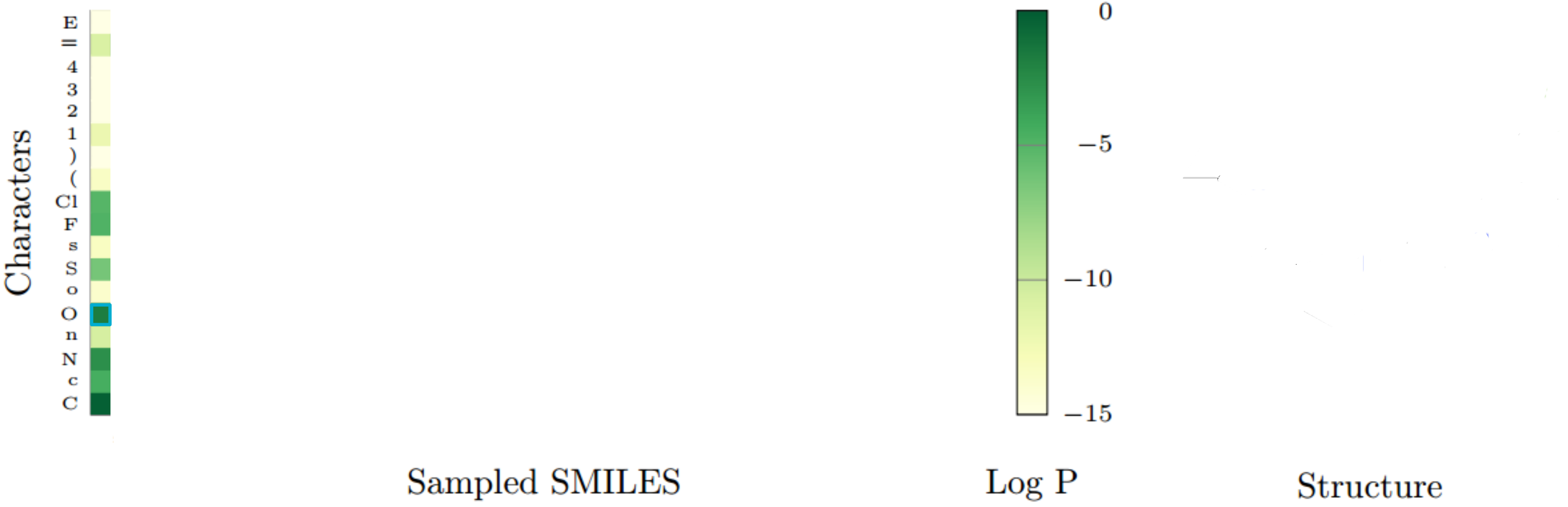
The network learns the rules of chemistry,
not the training examples

The trained RNN can now generate drug-like molecules



The network can generate up to 10^{60} Molecules

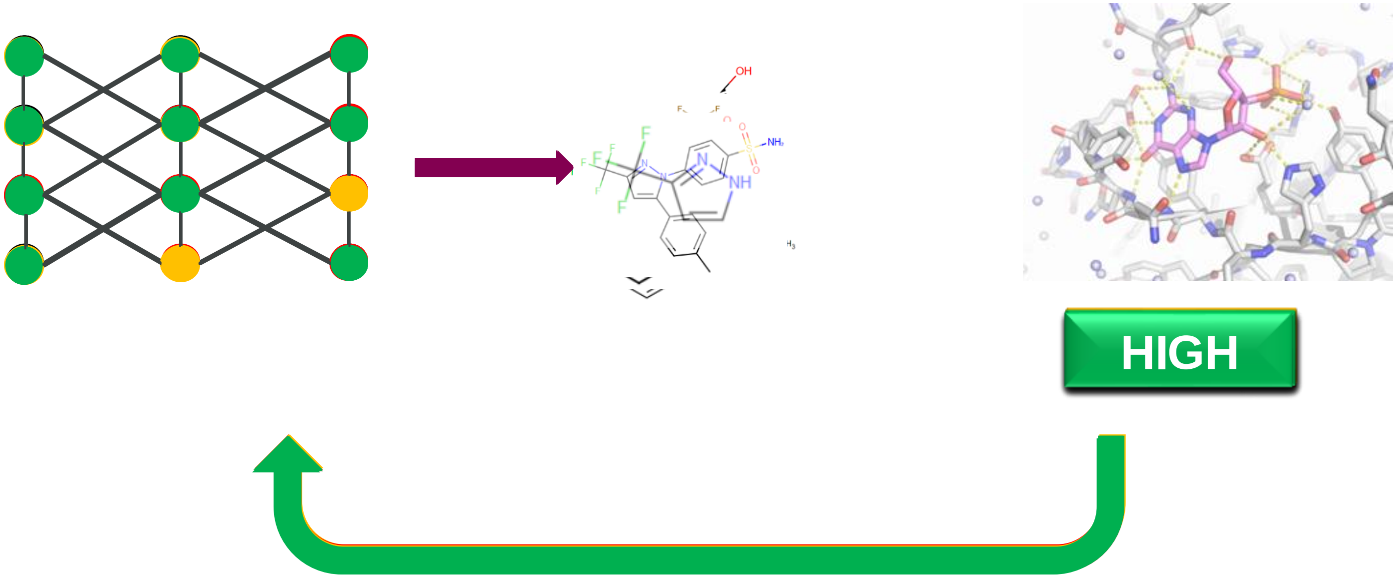
The generative process



Using the trained RNN to find high scoring molecules for a project through Reinforcement Learning

Generate

Score

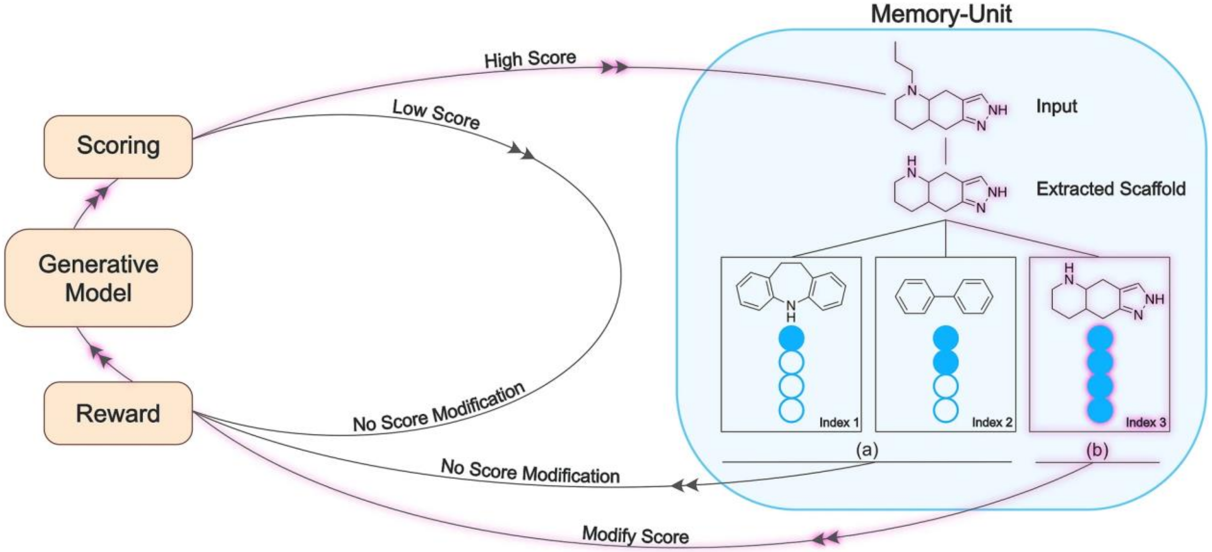


HIGH

Learn

To think about when using reinforcement learning

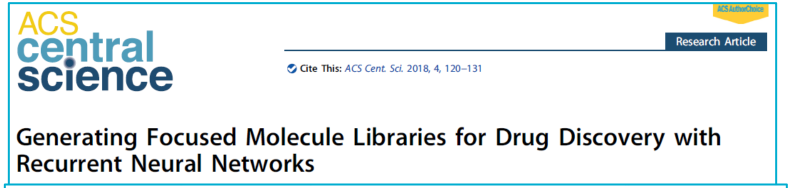
- RL will exploit loopholes in the scoring function
- RL will exploit the first minima it finds



Scaffold penalty to assure diverse scaffolds are identified

Blaschke et al Journal of Cheminformatics 2020

Science Molecular AI @AZ



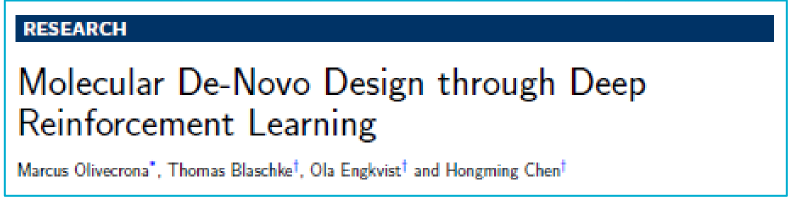
ACS central science

ACS AuthorChoice

Research Article

Cite This: *ACS Cent. Sci.* 2018, 4, 120−131

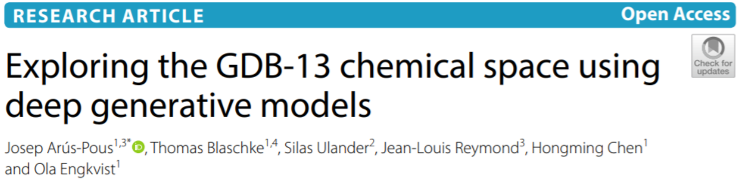
Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks



RESEARCH

Molecular De-Novo Design through Deep Reinforcement Learning

Marcus Olivecrona*, Thomas Blaschke†, Ola Engkvist† and Hongming Chen†

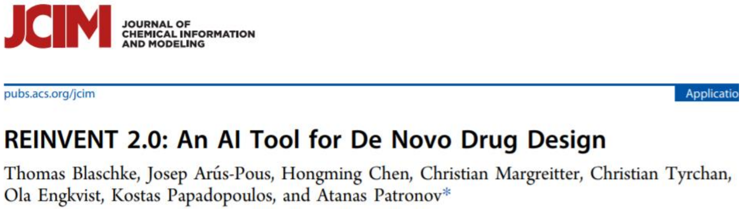


RESEARCH ARTICLE Open Access

Check for updates

Exploring the GDB-13 chemical space using deep generative models

Josep Arús-Pous[1,3*], Thomas Blaschke[1,4], Silas Ulander[2], Jean-Louis Reymond[3], Hongming Chen[1] and Ola Engkvist[1]

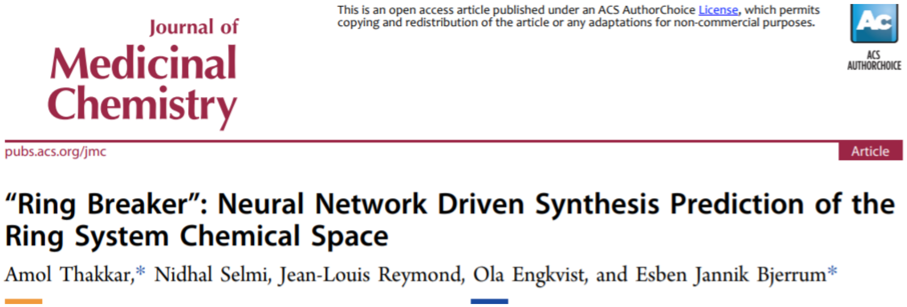


JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Applicatio

REINVENT 2.0: An AI Tool for De Novo Drug Design

Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov*



Journal of Medicinal Chemistry

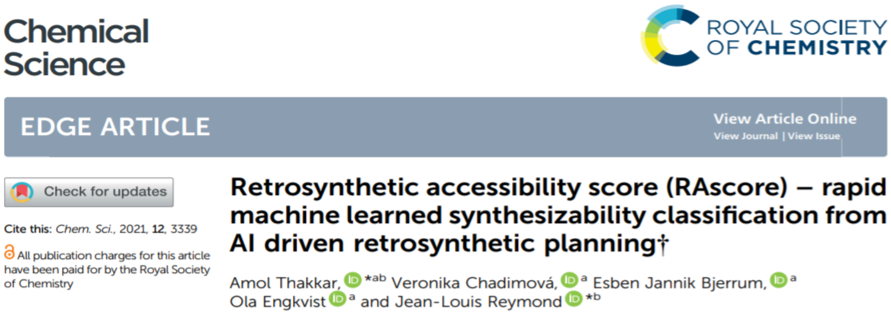
This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

ACS AUTHORCHOICE

pubs.acs.org/jmc Article

"Ring Breaker": Neural Network Driven Synthesis Prediction of the Ring System Chemical Space

Amol Thakkar,* Nidhal Selmi, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum*



Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

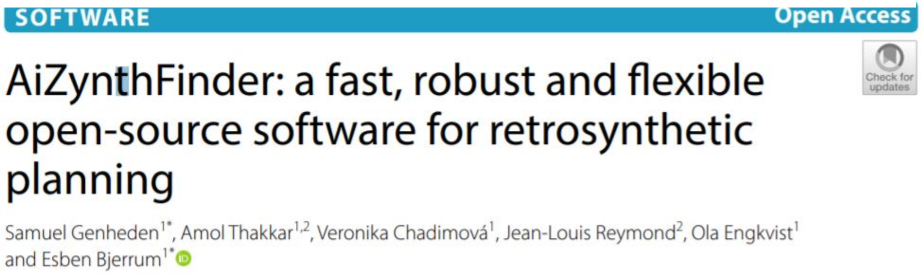
Check for updates

Cite this: *Chem. Sci.*, 2021, 12, 3339

All publication charges for this article have been paid for by the Royal Society of Chemistry

Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning†

Amol Thakkar,[*ab] Veronika Chadimová,[a] Esben Jannik Bjerrum,[a] Ola Engkvist[a] and Jean-Louis Reymond[*b]



SOFTWARE Open Access

Check for updates

AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning

Samuel Genheden[1*], Amol Thakkar[1,2], Veronika Chadimová[1], Jean-Louis Reymond[2], Ola Engkvist[1] and Esben Bjerrum[1*]

Open Source: https://github.com/MolecularAI

The MELLODDY objectives

On average, bringing one drug to market costs €1.9 billion and 13 years[1].

The virtualization of parts of drug discovery by machine learning is a promising approach to improve efficiencies.

MELLODDY aims to show predictive benefits of modelling across tasks, data types and partners at the largest achievable scale.

[1] DiMasi JA et al., 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. Journal of Health Economics 47, 20-33.

In three yearly runs, the increasingly sophisticated platform will learn from:

- > 10 million annotated small molecules
- > 1 billion assay biological activity labels
- Multiple high-complexity phenotypes at high throughput

- Multiple high-complexity phenotypes at high throughput

Privacy preservation of data and federated models is paramount.

Machine Learning Ledger
Orchestration for Drug Discovery

MELLODDY

powered by aws

PHARMA PARTNERS

AMGEN

astellas

AstraZeneca

BAYER

Boehringer Ingelheim

gsk

janssen
PHARMACEUTICAL COMPANIES
OF Johnson & Johnson

MERCK

NOVARTIS

SERVIER

PUBLIC PARTNERS

MŰEGYETEM 1782

IKTOS

Kubermatic

KU LEUVEN

NVIDIA

OWKIN

Substra Foundation

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement N° 831472. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA

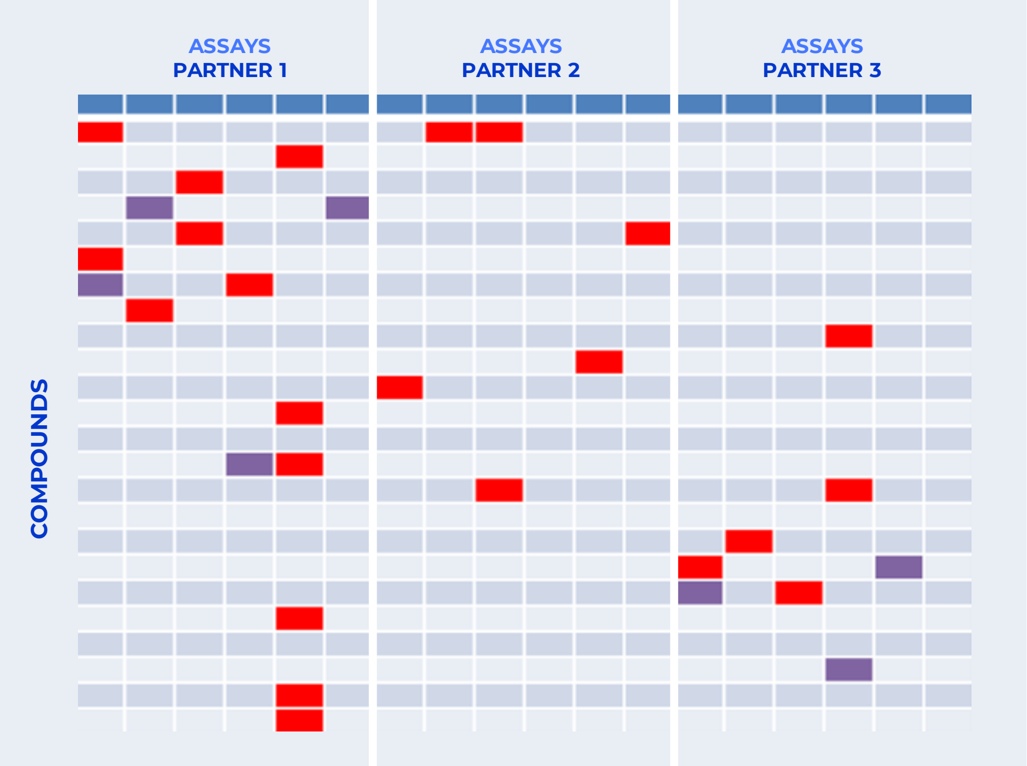
imi innovative medicines initiative

efpia

MULTI-TASK LEARNING
ACROSS PHARMA PARTNERS

Compound and activity data and assay-specific models remain under their owner's control

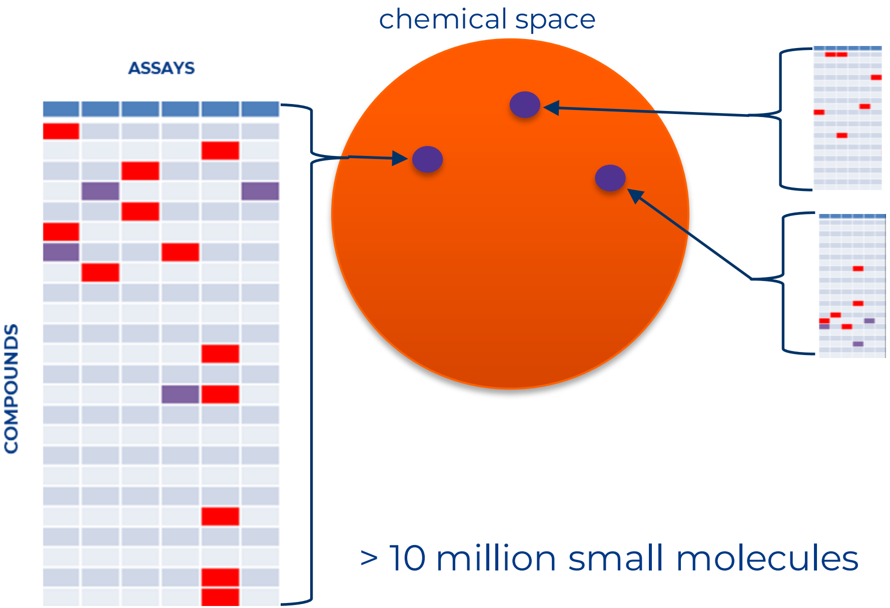
Multi-task approach across partners to improve predictive performance and applicability



AMGEN
astellas
AstraZeneca
BAYER
Boehringer Ingelheim
gsk
janssen
MERCK
NOVARTIS
SERVIER

How to achieve the objective?

Multi-task learning across pharma partners



> 10 million small molecules

> 1 billion biological activity labels

> 100,000 ML tasks

→ improved chemical space coverage

AMGEN astellas

AstraZeneca BAYER

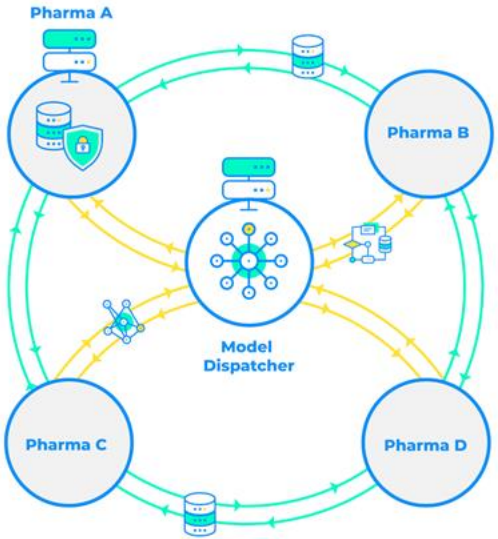
Boehringer Ingelheim gsk

janssen MERCK

NOVARTIS SERVIER

How to achieve the objective?

Multi-task federated learning



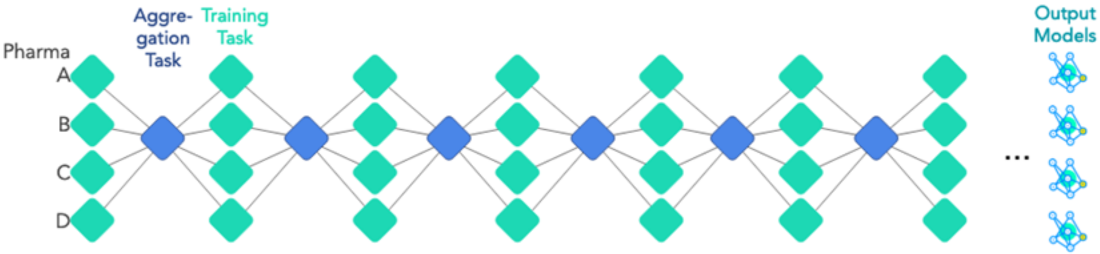
Federated Learning

- Data is not shared between partners. It remains stored in the server of its owner.
- ML model updates travel from one center to another to be trained.

Privacy preserving Multi-task Learning

ML model made of:

- Common trunk shared between partners
- Private heads are not shared between partners.



Compute Plan

- Set of Training, aggregation and evaluation tasks.

2nd federated learning run: success

Evidence of federated model superiority

Year 1: creation of a secure predictive modelling platform, operated at scale

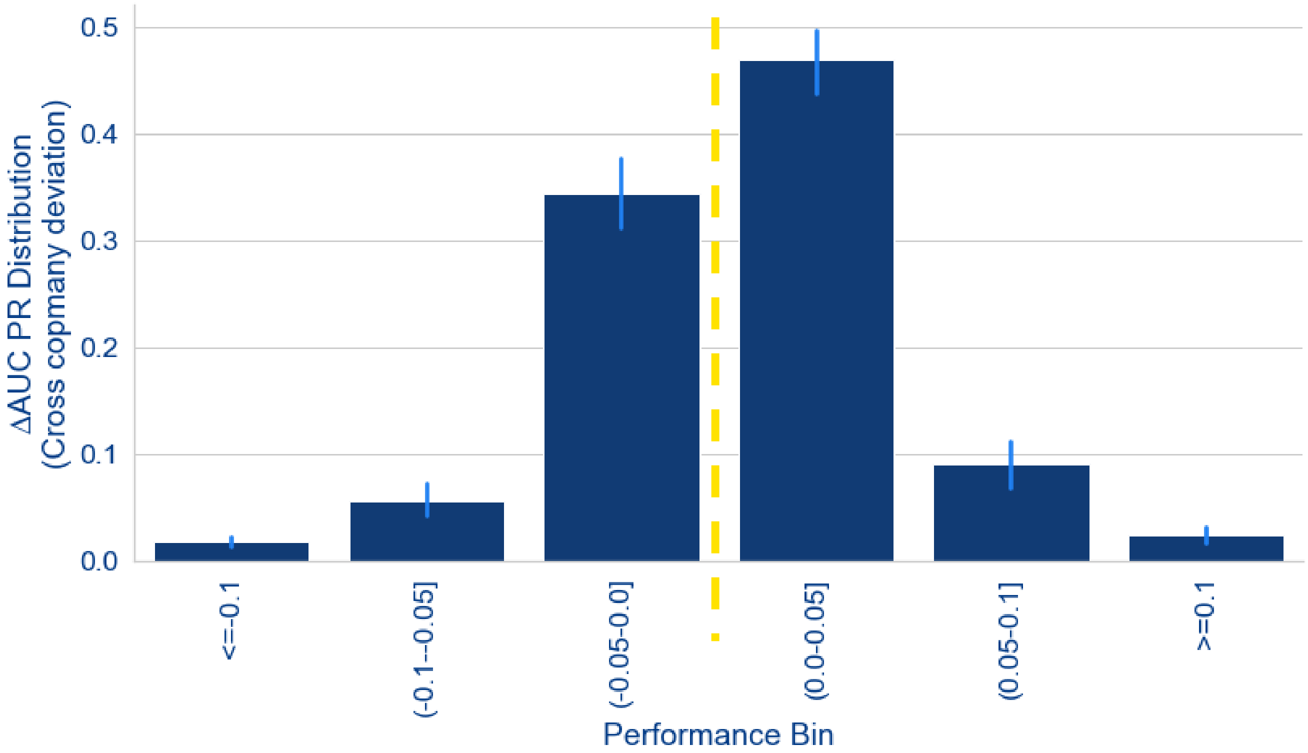
Year 2: study hypothesis that multi-partner modelling yields superior predictive models in drug discovery

- Early benefits of modelling across tasks, data types and partners
- Strong support for the working hypothesis of superior prediction quality and/or applicability domain of the common predictive drug discovery model to the single-partner modelling effort
- Open-source codebase & pending scientific publications and conferences

Year 3: improve predictive performance

2nd federated learning run: success

Evidence of federated model superiority

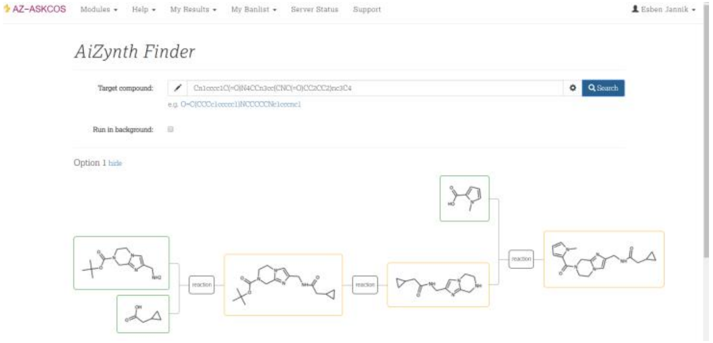


Average 0.60% delta AUC-PR improvement across the board (SD 0.008).

AiZynthFinder

https://github.com/MolecularAI/aizynthfinder

Web-GUI based on MIT MLDPS consortium tools

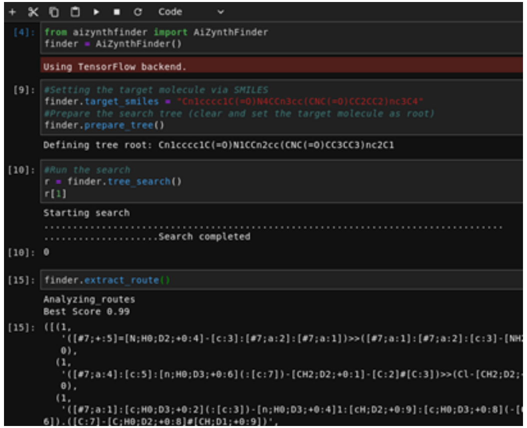


Target compound: Cn1cccc1C(=O)N4CCn3cc(CNC(=O)CC2CC2)nc3C4

e.g. O=C(CCCc1ccccc1)NCCCCCNc1cccnc1

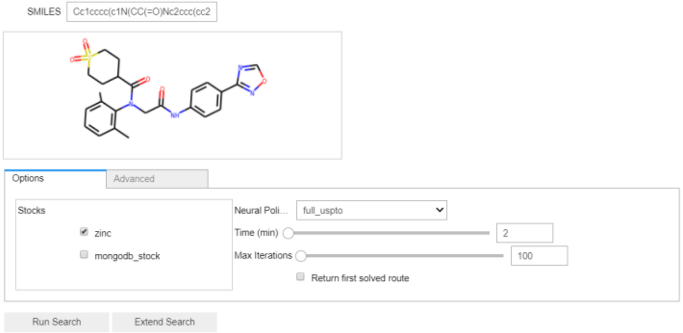
Run in background: □

Scripting access via Python Objects



```
[4]:  from aizynthfinder import AiZynthFinder
      finder = AiZynthFinder()

      Using TensorFlow backend.

[9]:  #Setting the target molecule via SMILES
      finder.target_smiles = "Cn1cccc1C(=O)N4CCn3cc(CNC(=O)CC2CC2)nc3C4"
      #Prepare the search tree (clear and set the target molecule as root)
      finder.prepare_tree()

      Defining tree root: Cn1cccc1C(=O)N1CCn2cc(CNC(=O)CC3CC3)nc2C1

[10]: #Run the search
      r = finder.tree_search()
      r[1]

      Starting search
      ..................................................................................
      ......................Search completed

[10]: 0

[15]: finder.extract_route()

      Analyzing_routes
      Best Score 0.99

[15]: ([(1,
         '([#7;+:5]=[N;H0;D2;+0:4]-[c:3]:[#7;a:2]:[#7;a:1])>>([#7;a:1]:[#7;a:2]:[c:3]-[NH
         0),
        (1,
         '([#7;a:4]:[c:5]:[n;H0;D3;+0:6](:[c:7])-[CH2;D2;+0:1]-[C:2]#[C:3])>>(Cl-[CH2;D2;
         0),
        (1,
         '([#7;a:1]:[c;H0;D3;+0:2](:[c:3])-[n;H0;D3;+0:4]1:[cH;D2;+0:9]:[c;H0;D3;+0:8](-[
      6]).([C:7]-[C;H0;D2;+0:8]#[CH;D1;+0:9])',
```

Jupyter based GUI



SMILES Cc1cccc(c1N(CC(=O)Nc2ccc(cc2

Stocks
☑ zinc
□ mongodb_stock

Neural Poli... full_uspto

Time (min) 2

Max Iterations 100

□ Return first solved route

Retrosynthesis

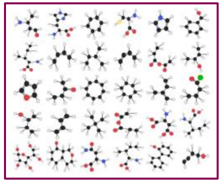
@retrosynthchan

Twitter bot that conducts retrosynthetic analysis

Genheden et al. AiZynthFinder: A Fast Robust and Flexible Open-Source Software for Retrosynthetic Planning. **2020**. https://doi.org/10.26434/chemrxiv.12465371.v1 .

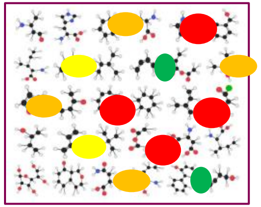
Artificial Intelligence Guided Drug Design Platform (REINVENT)

Molecular Generation



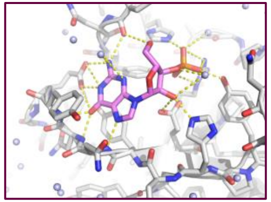
- Recurrent neural networks
- Reinforcement learning
- Scaffold penalty
- Library design

Scoring



- Synthetic accessibility score
- QSAR models (ADME, off-target)
- 3D shape
- Docking

Postprocessing



- free energy perturbation binding affinity
- Synthetic route assesment
- Clustering

Core is Based on Open Source Software

Commercial Plugins when appropriate for scoring

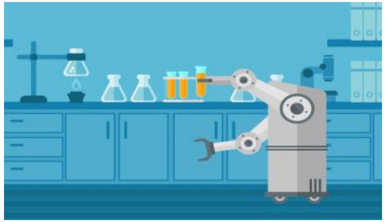
AI+ vision for drug design

AI can't transform drug design alone

High-Throughput Data Generation

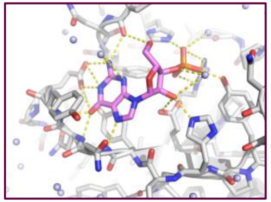
- The most important determinant of the usefulness of a model is the size and quality of the data set for training
- High-Throughput Experimentation for generating chemical reaction data
- Cell-paint & transcriptomics to create molecular signatures
- DNA Encoded Library models to score molecules

Automatize Make & Test



- Autonomous optimization of compounds is needed to radically cut timelines for clinical candidate delivery
- Multistep reactions with intermediate purification on automation platform
- Automatic testing after synthesis & purification
- Autonomous decision making under uncertainty which compounds to make
- Human-in-the-loop modelling

Combine AI with physics

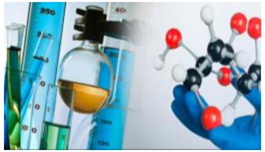


- More accurate models for difficult to predict properties can be created through combining physics and AI
- Relative binding free energy perturbation binding affinity in molecular optimization
- Absolute binding free energy perturbation to estimate binding energies in hit finding and for scaffold hopping
- Estimation of thermodynamic solubility
- Combine ML/MD to identify cryptic pockets

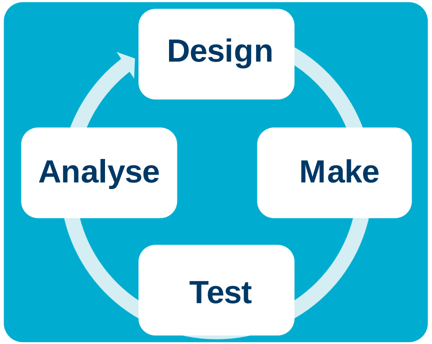
Integration of AI and automation

AI

What to make next ?



How to make it ?



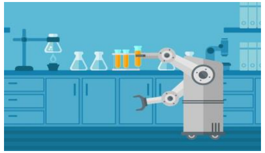
Design

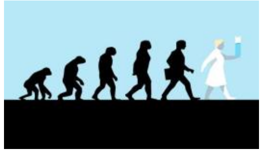
Make

Test

Analyse

How to Automate Synthesis ?

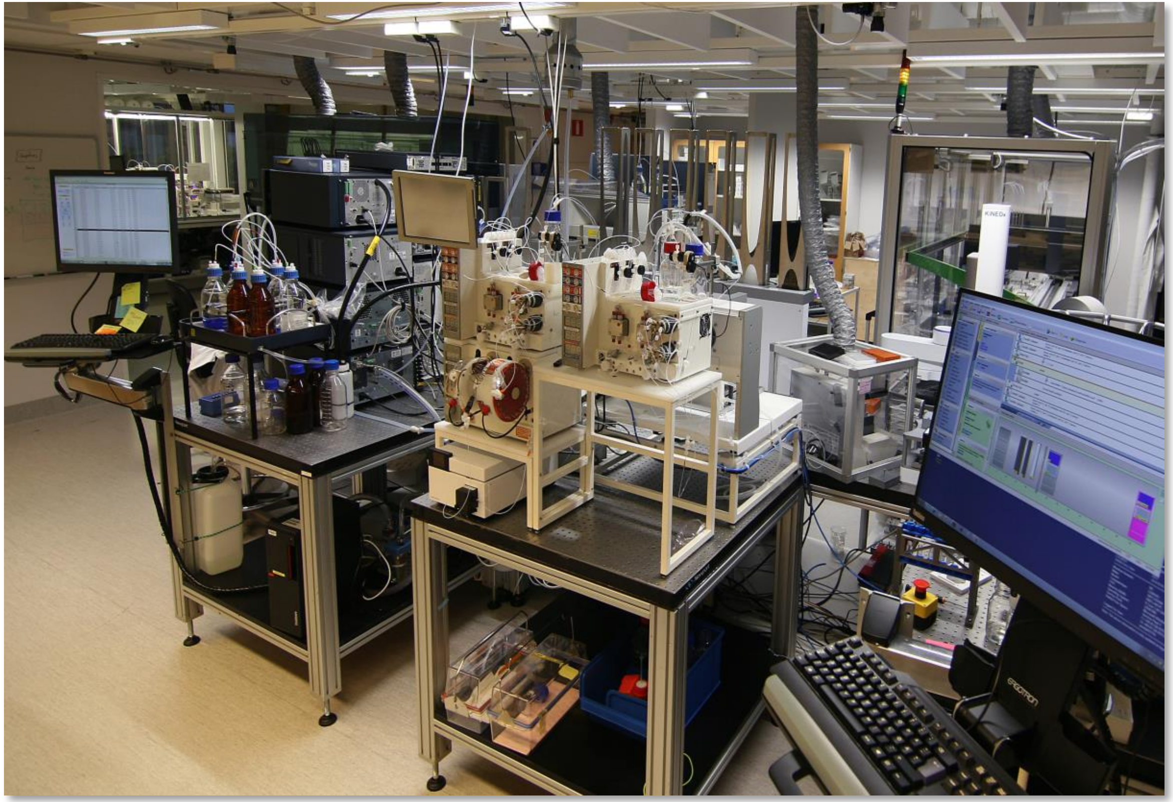




How to increase synthesis throughput

iLAB

Automated Synthesis Platform @AstraZeneca



What about AlphaFold2?

- Terrific achievement!
  - Winning a prospective competition with margin based on public data!
  - Big Science (People, Compute)
  - Public release will encourage further development & innovation
  - Looking forward to the next generation of models (capturing protein dynamics, RNA structures)

- Impact on drug design
  - Facilitate solving x-ray and Cryo-EM structures
  - Lack of protein dynamics have limited the use so far

What does success look like?

- Metrics like time saving are the results of success not the success itself

- Trust in the AI designed molecules in the same way as for instance x-ray crystal structures are trusted
  - Trust in the predictions for individual molecules
  - Trust that the AI generated molecules are the best molecules taking the project most efficiently to a clinical candidate

What are the challenges for AI driven drug design?

- Scaling ML/AI solutions for drug design to a whole drug discovery project portfolio including projects with low data volume
  - (pre-trained) molecular transformers
  - Privacy-preserving ML

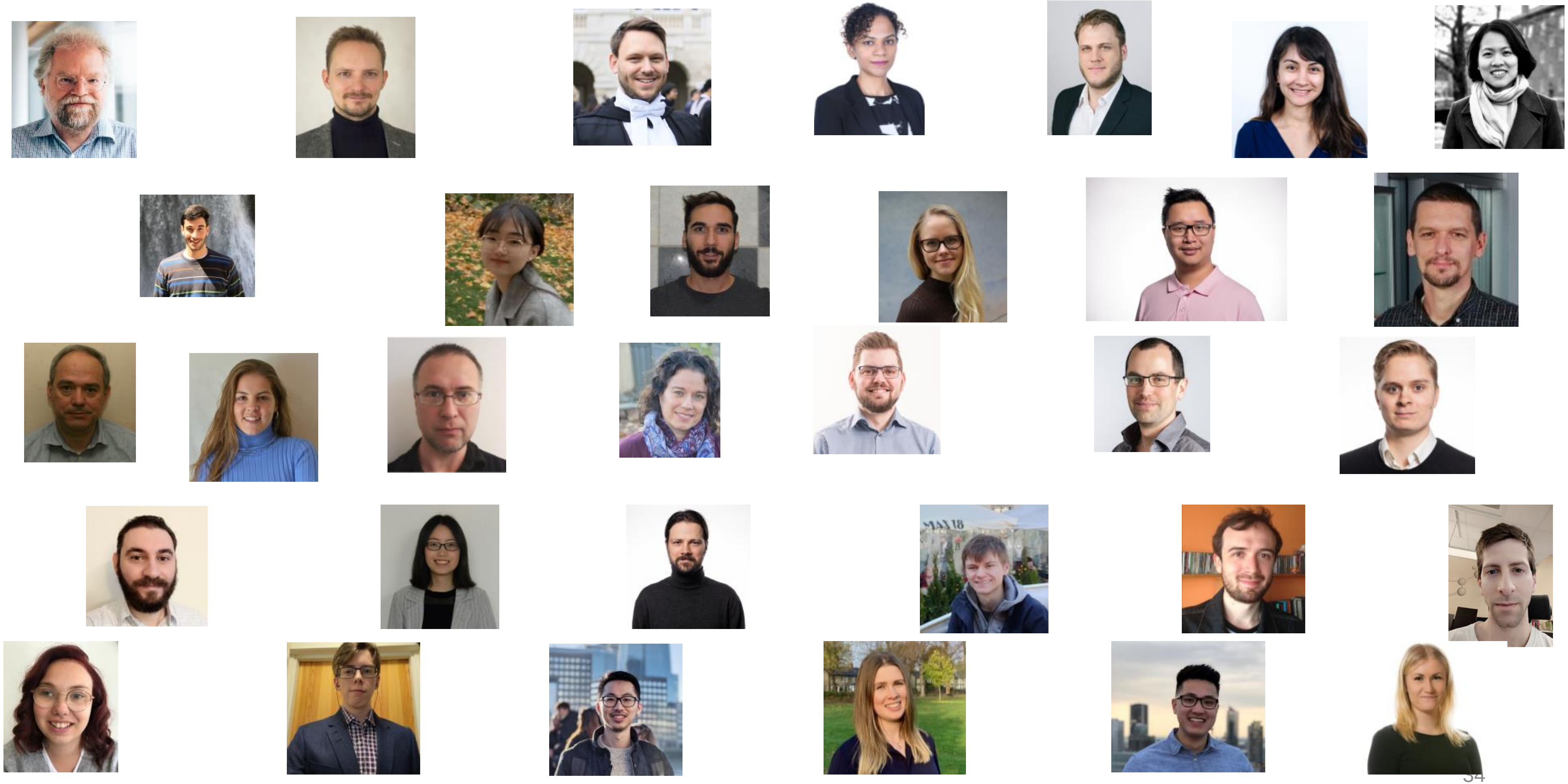
- Physics based modelling
  - Binding affinity and solubility predictions are major bottlenecks

- “Cambrian revolution” of new AI methods makes it difficult to assess progress

- Flexibility of chemistry automation

- Educational, cultural & logistical challenges besides scientific

Molecular AI



Confidentiality Notice