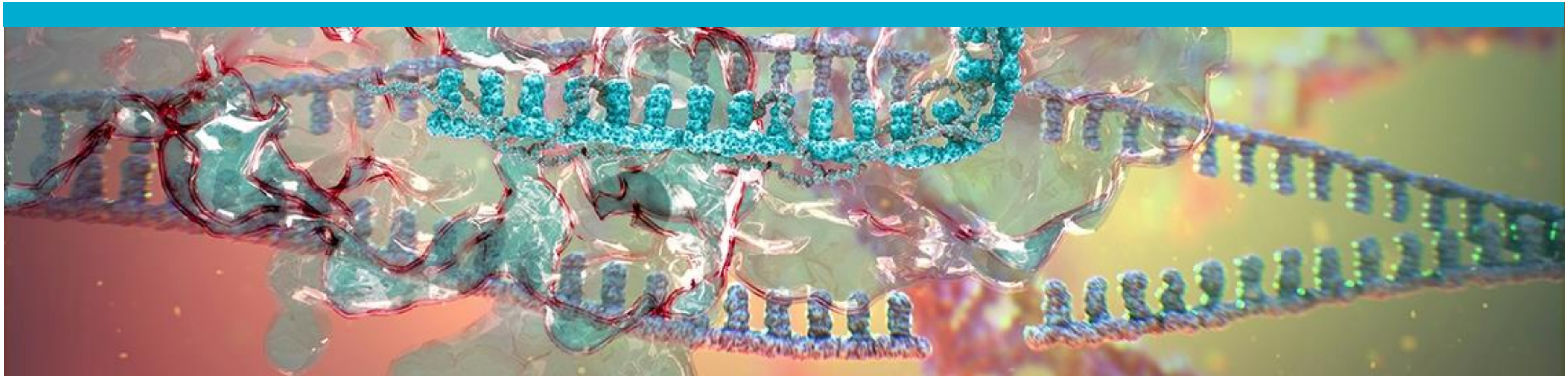
AstraZeneca
IMED Biotech Unit

AI for drug design an industrial perspective

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

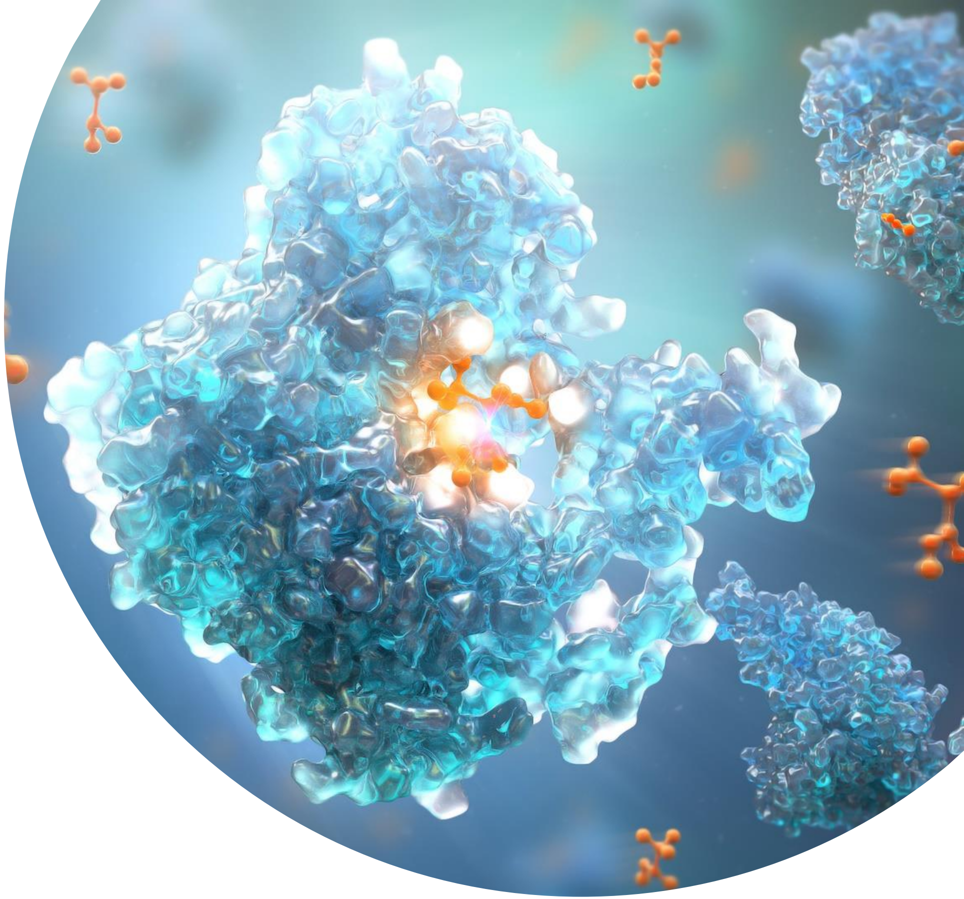
Chemoinformatics Strasbourg Summer School 2022

July 1 2022

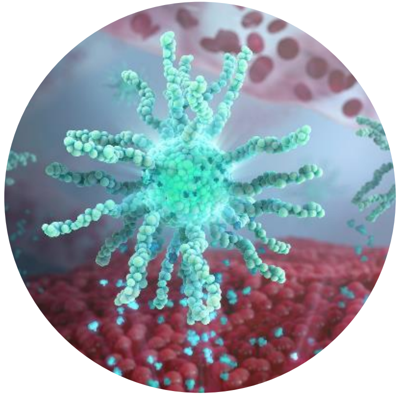


We push the boundaries of science to deliver life-changing medicines

Inspired by our **values** and **what science can do**, we are **focused** on **accelerating the delivery of life-changing medicines** that create enduring **value for patients and society.**



2021 global dimensions



$37.4bn
Total Revenue (incl. COVID-19 vaccine)

+38%
Total Revenue growth (23% excl. COVID-19 vaccine)

$9.7bn
invested in our science

13
medicines with annual sales of more than $1 billion

$13bn
Oncology Product Sales

$8bn
Cardiovascular, Renal & Metabolism Product Sales

$6bn
Respiratory & Immunology Product Sales

$3bn
Rare Disease Product Sales (from 21/7/2021)

22
Regulatory approvals and authorisations in major markets

2.5bn
COVID-19 vaccine doses supplied to more than 180 countries together with our partners

110
successful markets launches

161
projects in clinical phase of development

83,100
employees
(Dec. 2021)

87%
of employees believe strongly in our future direction and key priorities (Nov. 2021)

59%
Reduction in Scope 1 and 2 greenhouse gas emissions since 2015

31m
people reached through our Access to Healthcare programmes

NOTE: All growth rates at Constant Exchange Rates
Source: 2021 Annual Report

Global, science-led, patient-focused biopharmaceutical company

Science and innovation-led

Therapy areas of focus: Oncology; Cardiovascular, Renal & Metabolism; Respiratory & Inflammation; Rare Disease

Diversified portfolio with broad coverage across primary care, specialty care and rare diseases

Commitment to people and society

Global strength, with balanced presence across regions

Our Strategic Priorities

Deliver growth and therapy area leadership

Accelerate innovative science

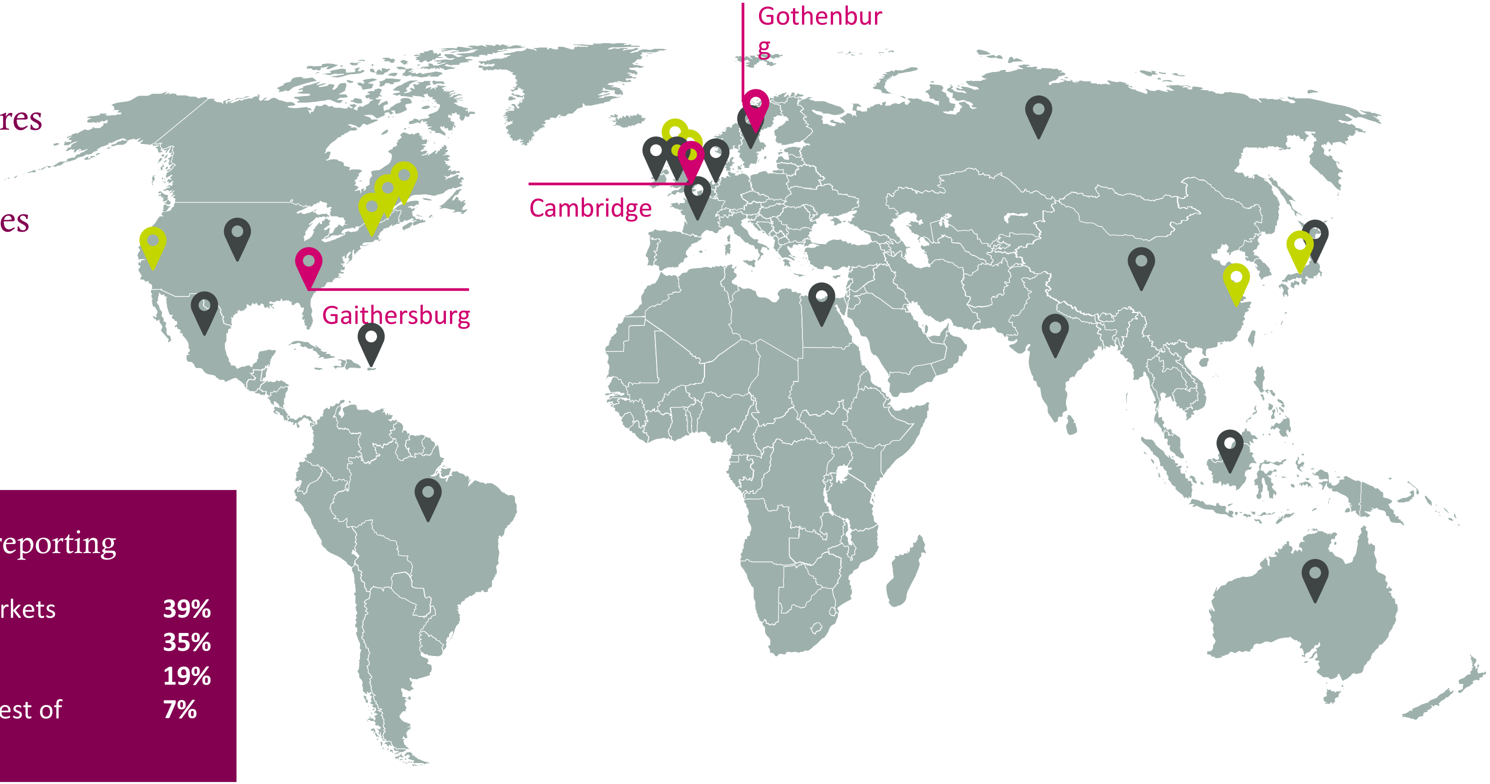
Be a great place to work

Global reach and presence

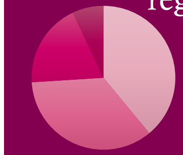
3 global R&D centres

8 other R&D centres and offices

28 manufacturing sites
in 16 countries



Employees by reporting region



Emerging Markets	39%
Europe	35%
US	19%
Established Rest of World	7%

AstraZeneca Gothenburg

AstraZeneca in Gothenburg

Gothenburg has a unique culture of **collaboration and open innovation**, supporting all our main therapy areas throughout the entire life cycle of our medicines.

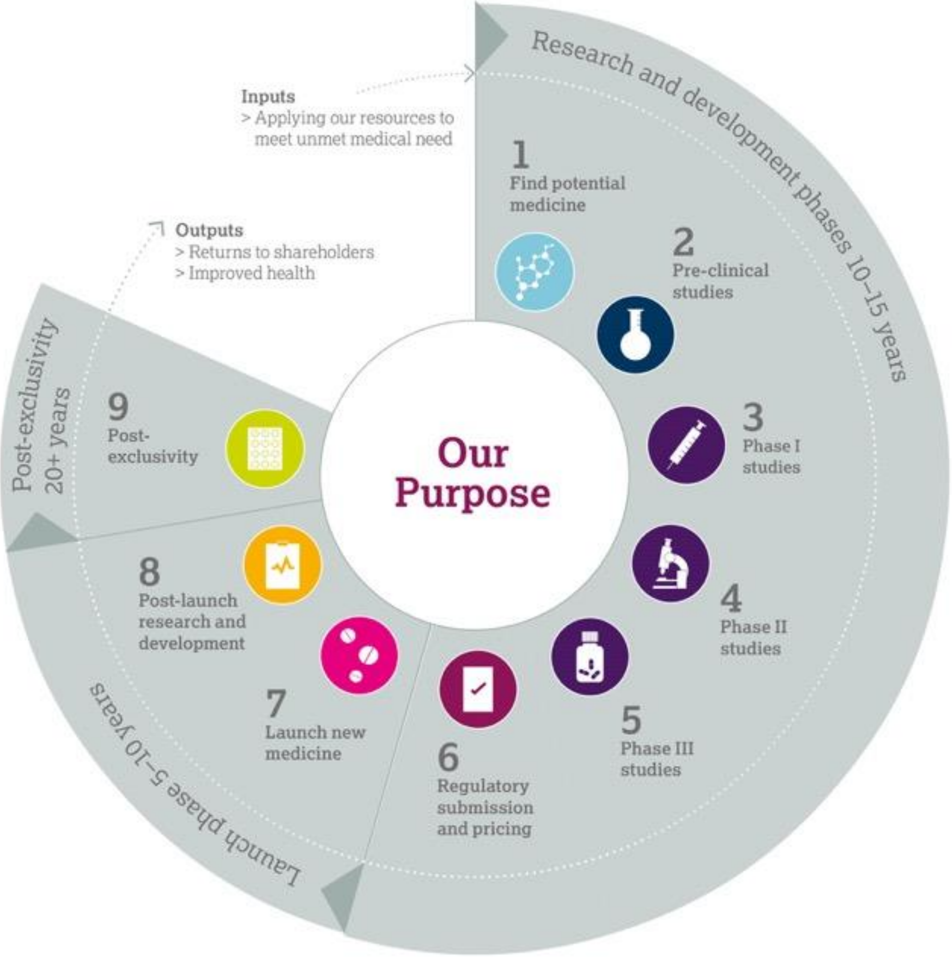
Supporting the entire life cycle of medicines: drug discovery, advanced drug delivery, medical device development, manufacturing for clinical trials.

~**2,800 employees** of **70+ nationalities,** including 600 PhD researchers and 30 professors. 57.9 percent of our line managers in Gothenburg are women.

Reduced the site's carbon dioxide emissions by 99% since 2015 with a clear vision to further reduce our carbon footprint.

From idea to patient

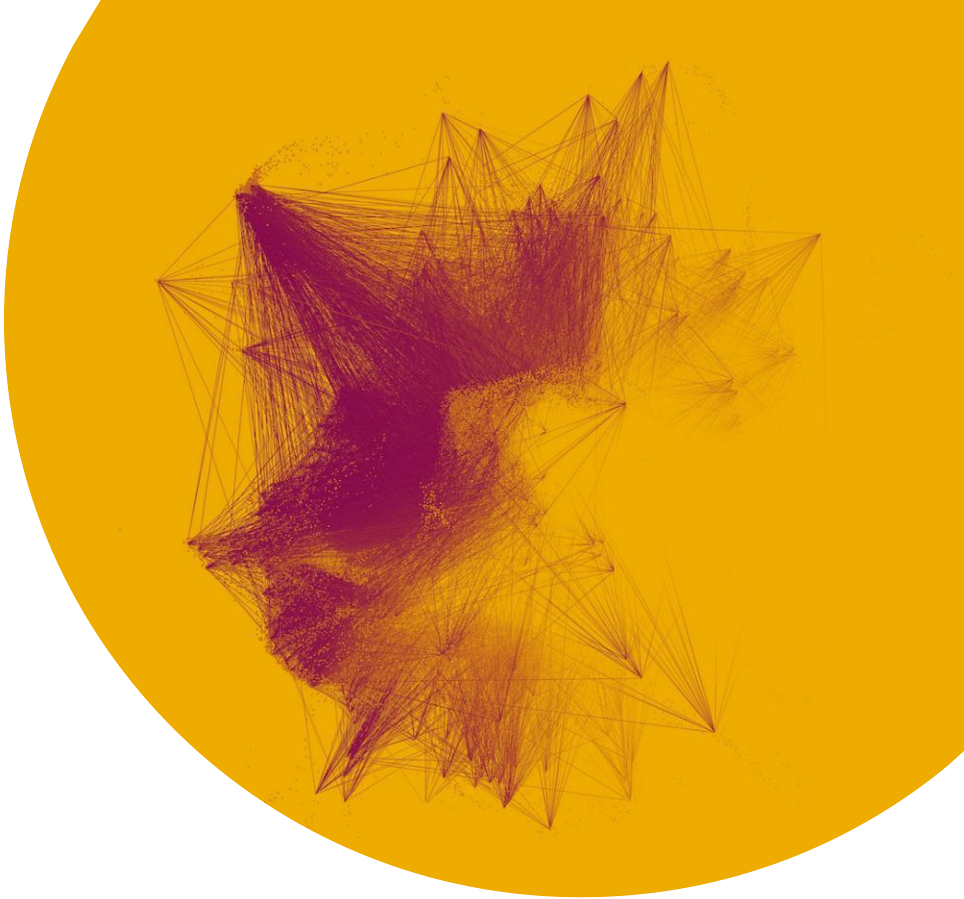
One of the site’s unique features is that we have most of the resources and expertise needed to support the entire lifecycle of a medicine. That’s everything from idea generation, via clinical development, to pilot scale manufacturing and distribution, and global commercialisation and product maintenance.


Inputs
> Applying our resources to meet unmet medical need
Outputs
> Returns to shareholders
> Improved health
Research and development phases 10–15 years
1 Find potential medicine
2 Pre-clinical studies
3 Phase I studies
4 Phase II studies
5 Phase III studies
6 Regulatory submission and pricing
7 Launch new medicine
8 Post-launch research and development
9 Post-exclusivity
Launch phase 5–10 years
Post-exclusivity 20+ years
Our Purpose

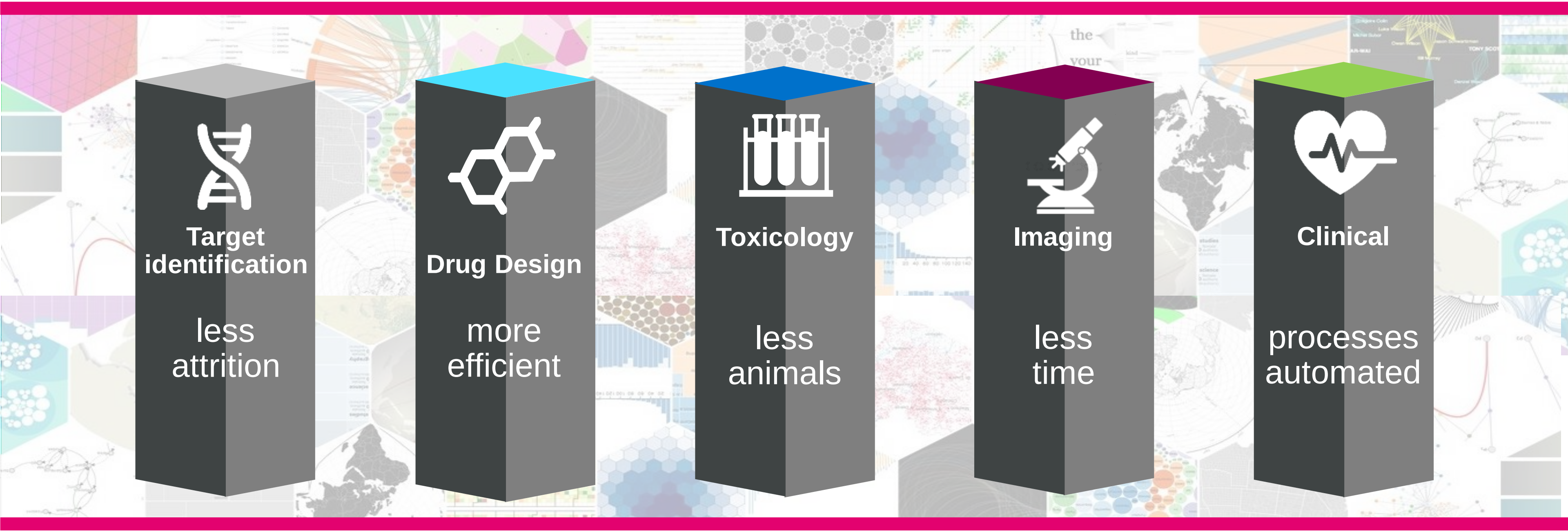
Data science & AI: transforming drug discovery and development

AI and machine learning are transforming the way we discover and develop new medicines.

We aim to increase the probability of success and reduce drug discovery and development timelines by applying advanced AI and machine learning across R&D.



Where can AI impact drug discovery and development



Target identification

less attrition

Drug Design

more efficient

Toxicology

less animals

Imaging

less time

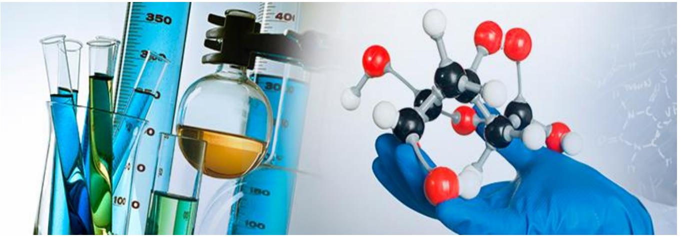
Clinical

processes automated

Drug Design

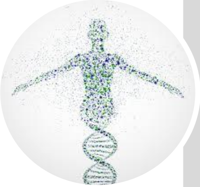
Which compound to make next?

How to make the compound?



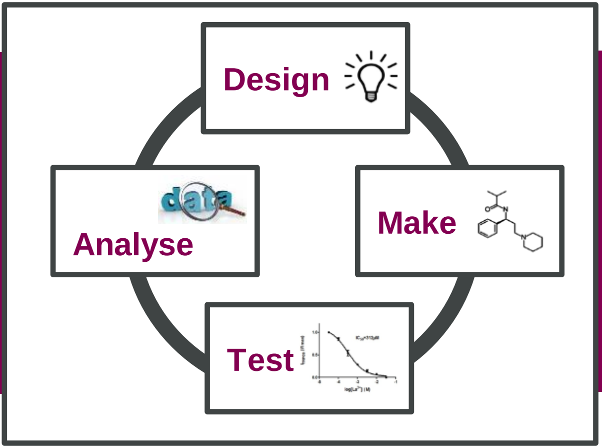
The Design Make Test Analyze cycle in Drug Design

Drug target



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

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



Design

Make

Test

Analyse

Candidate drug

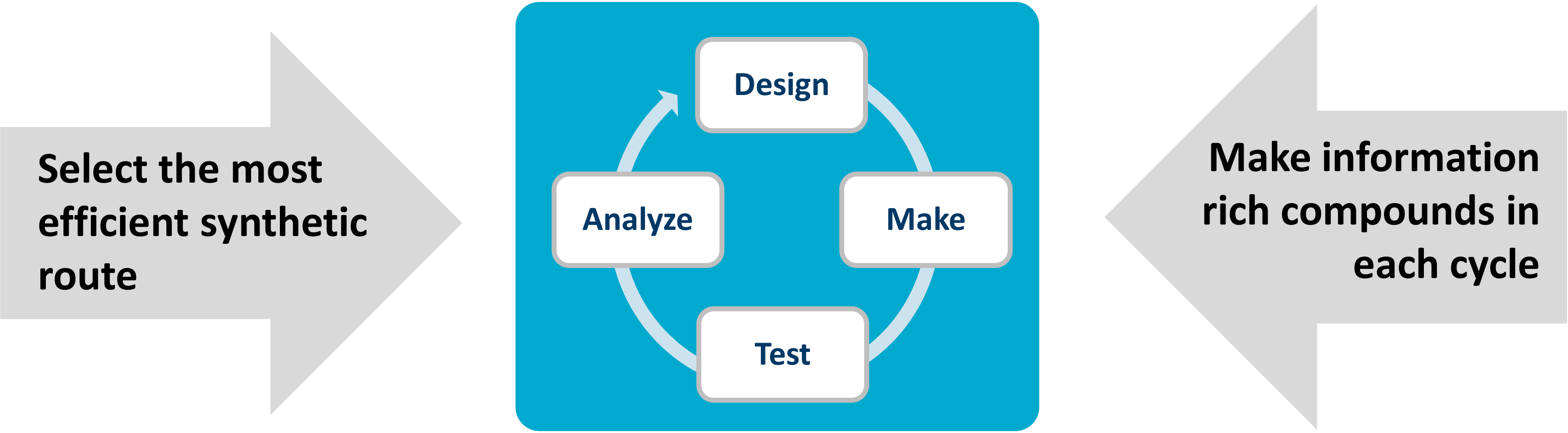
- Highly potent
- Effective in *in vivo* models
- Metabolically stable
- No toxicity issues

~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

Design

Make

Test

Analyze

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

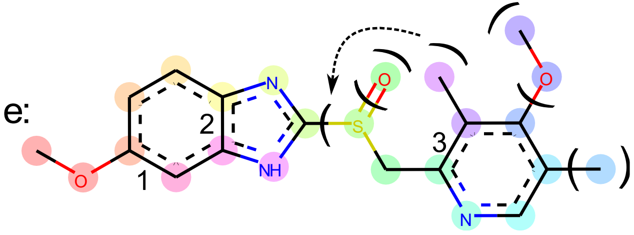
- Open-source software

  Python, RDKit, scikit-learn, PyTorch, Tensorflow

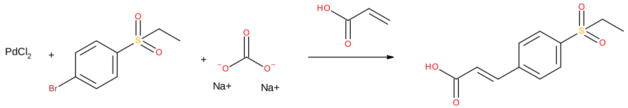
How can we take advantage of the progress in Natural Language Processing?

Molecules can be described with the language SMILES

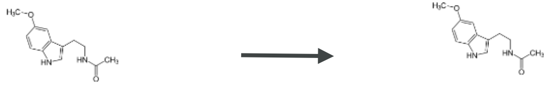
Molecule:



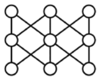
SMILES: COc1ccc2nc(S(=O)Cc3ncc(C)c(OC)c3C)[nH]c2c1



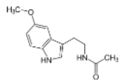
Language Translation ⟷ Synthesis prediction



Language Translation ⟷ Molecular optimization

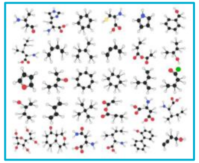


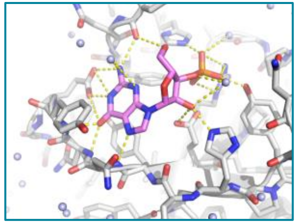
CC(=O)NCCc1c[nH]c2ccc(OC)cc12

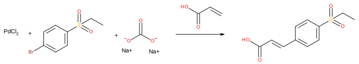


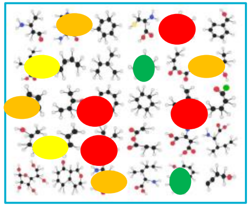
Text generation ⟷ Chemical space exploration

Where are we with AI based drug design?









- ✓ Deep learning based molecular de novo generation
  - ✓ Algorithms to navigate the whole relevant chemical space now exists
  - ✓ Scoring the generated molecules the main bottleneck

- ✓ Synthetic route prediction
  - ✓ Powerful new algorithms have been developed
  - ✓ Further progress needs better data

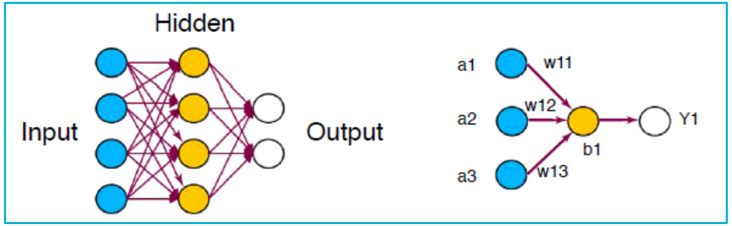
- ✓ Molecular property prediction
  - ✓ Novel more flexible deep learning based methods
  - ✓ No AF2 moment, No progress in prospective competitions (SAMPL, IDG, Cache)

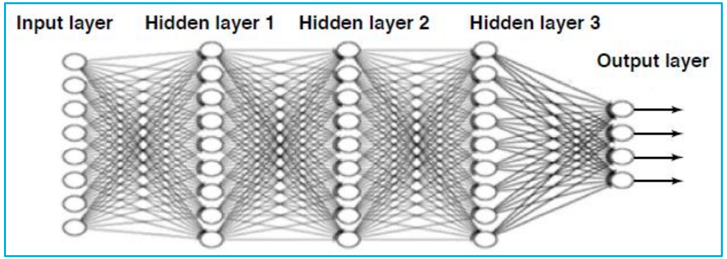
- ✓ 3D protein prediction
  - ✓ Stunning progress with AlphaFold2
  - ✓ Dynamics needs to be included for major impact

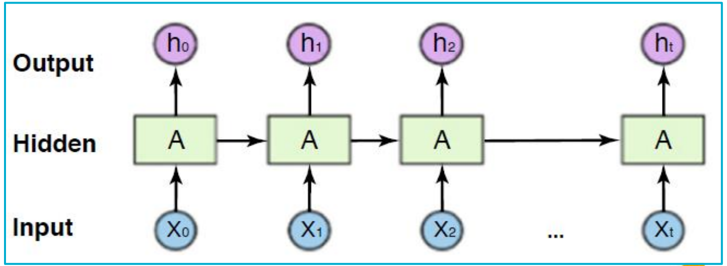
- ✓ It is not only about ML/AI based technology
  - ✓ AI+ vision need to include high-throughput data generation, automation & physics-based modelling
  - ✓ AI first culture
  - ✓ Continiuous training and education of staff

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







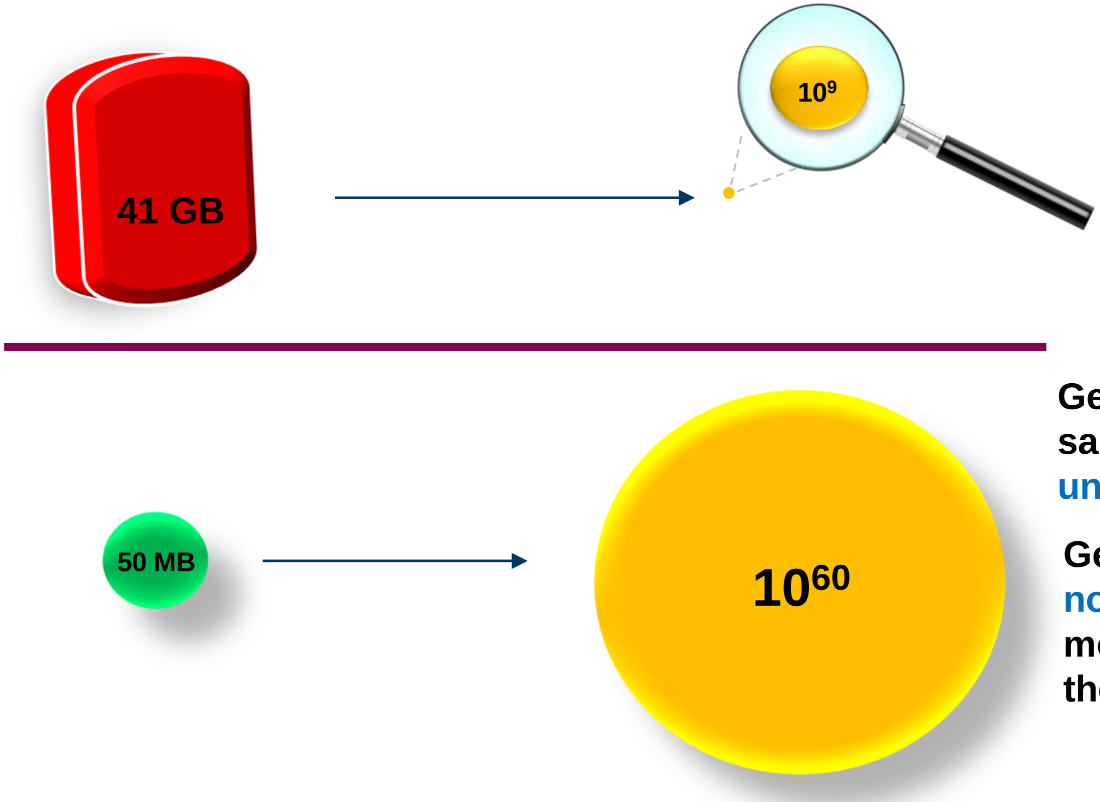
Generative Model vs Enumeration for molecular discovery

Physical Storage Size

Size of Molecular Space

Traditional Enumeration

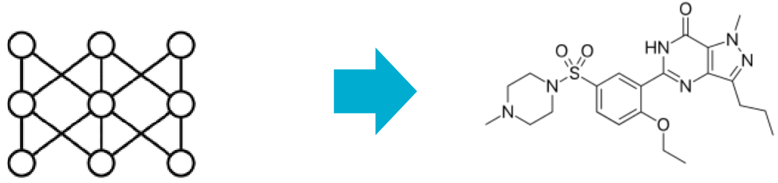
Generative Model



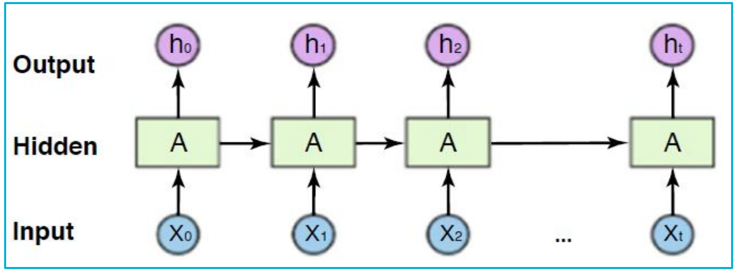
Generative models can sample practically unlimited chemical space

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

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



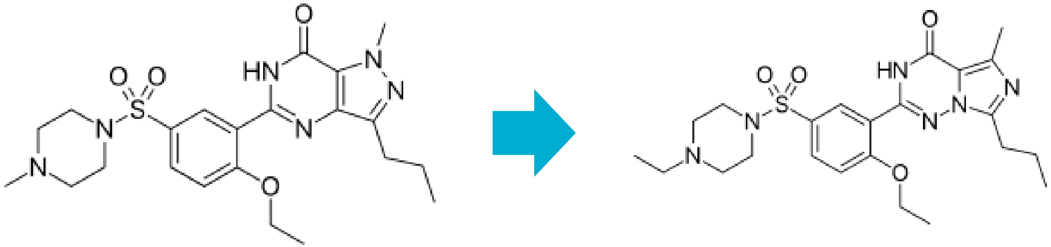
Recurrent Neural Networks

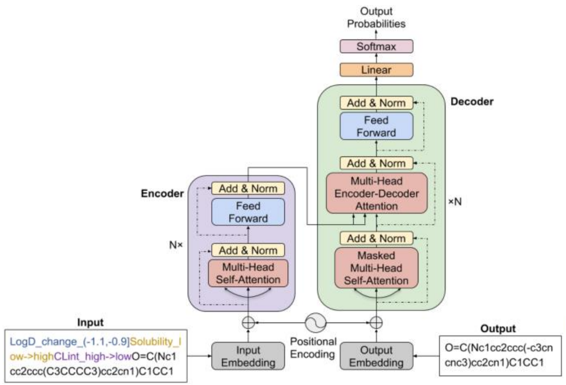


Hit Finding & scaffold hopping
Sample the whole chemical space

Transformer

Molecular Optimisation
Sample a focused chemical space

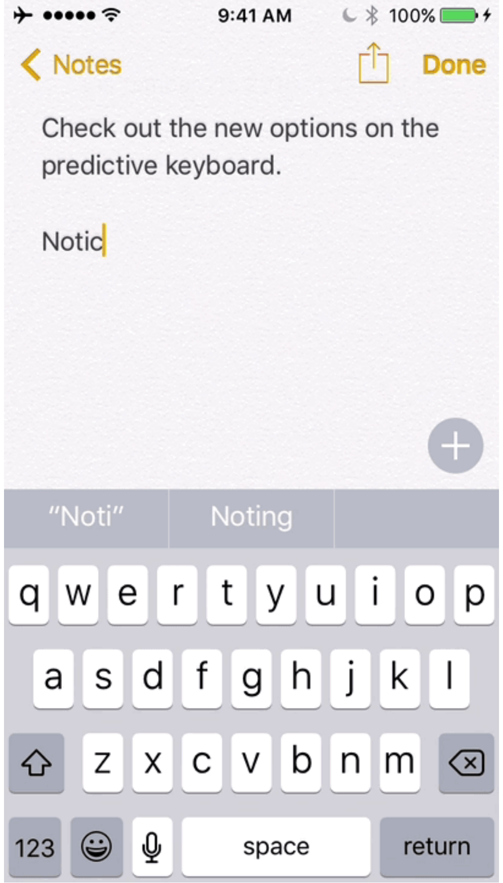




LogD_change_(-1.1,-0.9]Solubility_low->highCLint_high->lowO=C(Nc1cc2ccc(C3CCCC3)cc2cn1)C1CC1

O=C(Nc1cc2ccc(-c3cncnc3)cc2cn1)C1CC1

Recurrent Neural Network & Natural language generation



Check out the new options on the predictive keyboard.

Notic

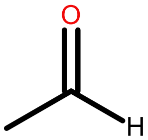
Natural language generation and molecular structure generation

- 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 → = → ?

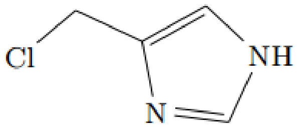


- $P(O \mid =, C, C)$

Tokenization of SMILES

- Tokenize combinations of characters like “Cl” or “[nH]”
- Represent the characters as one-hot vectors

Graph:

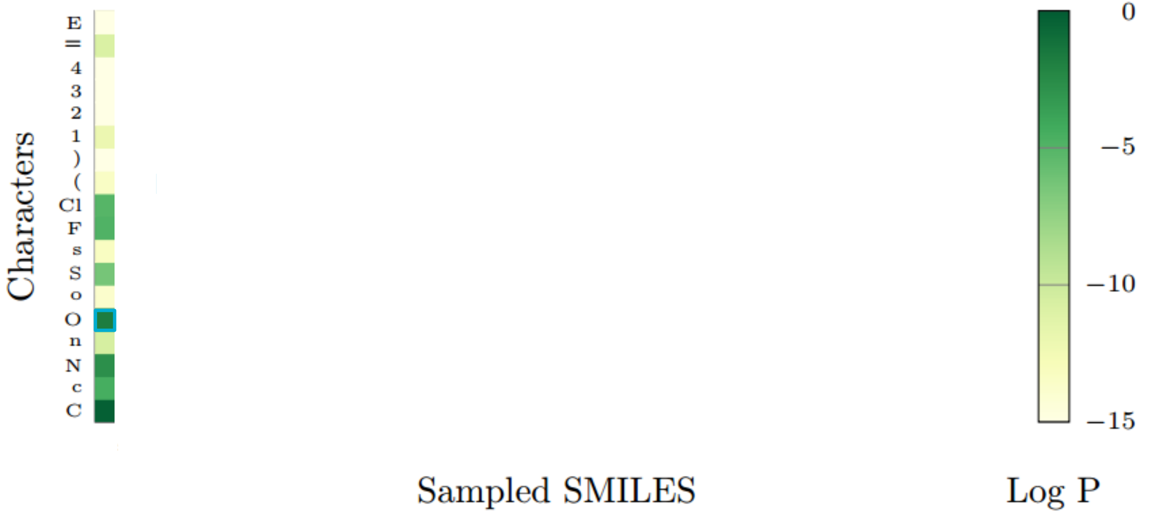


SMILES: ClCc1c[nH]cn1

One-hot encoding:

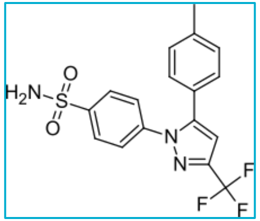
	Cl	C	c	1	c	nH	c	n	1
C	0	1	0	0	0	0	0	0	0
c	0	0	1	0	1	0	1	0	0
n	0	0	0	0	0	0	0	1	0
1	0	0	0	1	0	0	0	0	1
nH	0	0	0	0	0	1	0	0	0
Cl	1	0	0	0	0	0	0	0	0

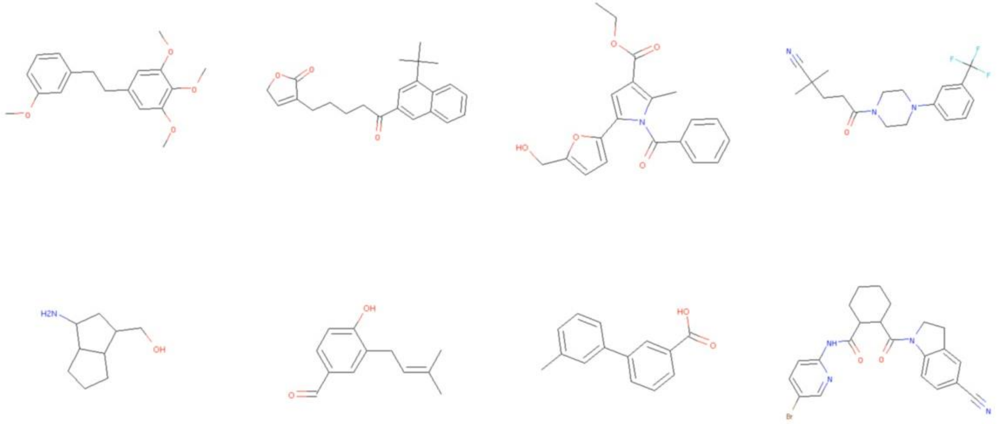
The generative process

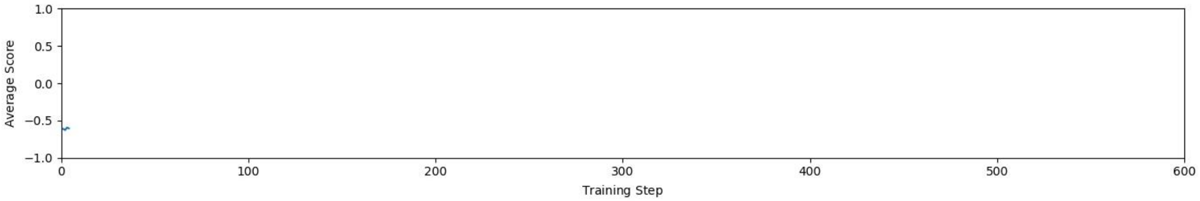


Structure

AI live: Create Structures Similar to Celecoxib



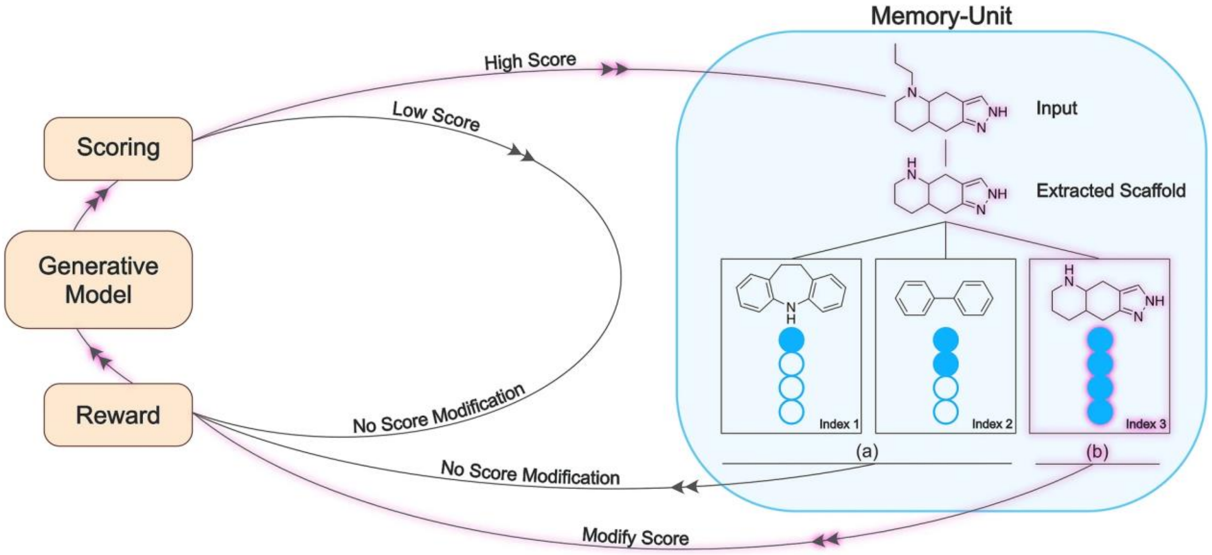




- **Key Message**
  - RNN generates structures similar to Celecoxib
  - Rapid sampling!
  - Average score describes how many learning steps are required to reach similar compounds

To think about when using reinforcement learning

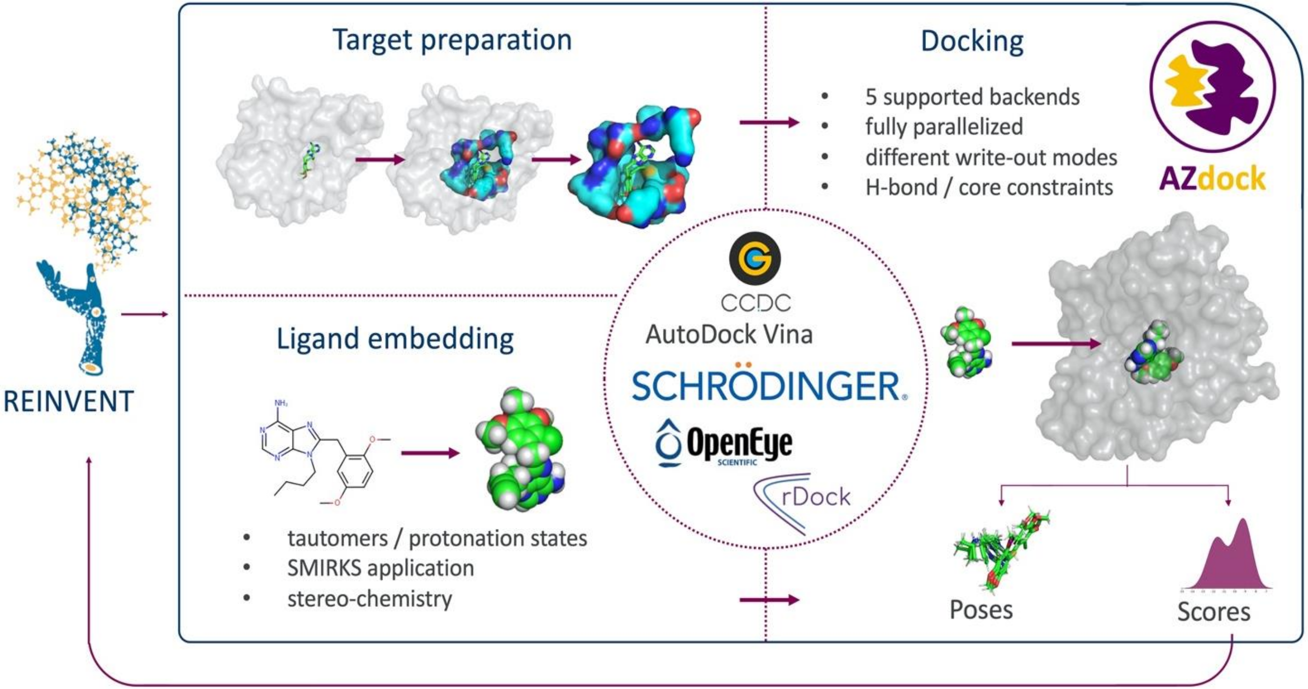
- RL will exploit loopholes in the scoring function
- RL will exploit the first minima it finds



Scaffold penalty to assure diverse scaffolds are identified

Blaschke et al Journal of Cheminformatics 2020

Docking is essential to score molecules



Docking

- 5 supported backends
- fully parallelized
- different write-out modes
- H-bond / core constraints

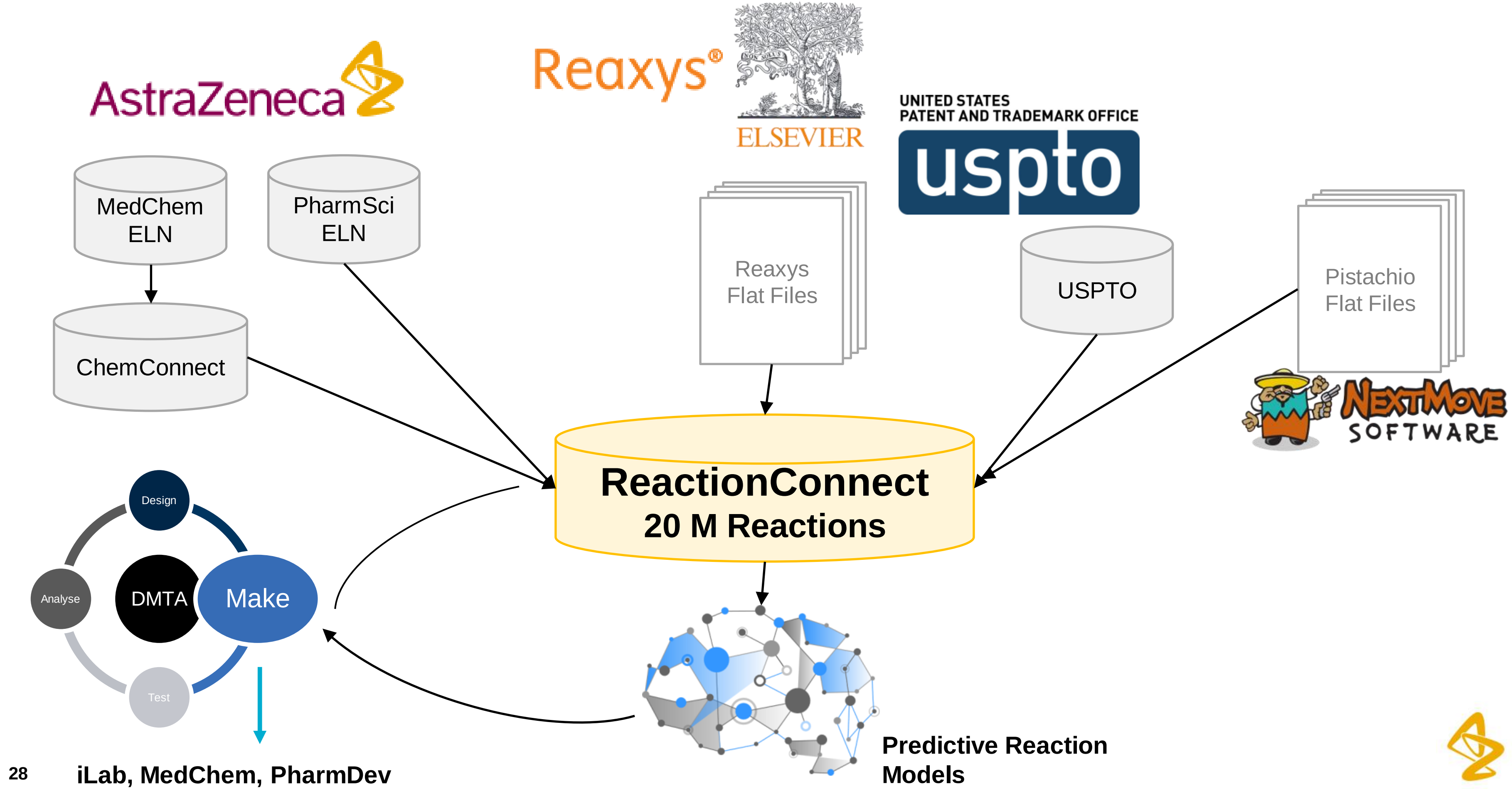
Ligand embedding

- tautomers / protonation states
- SMIRKS application
- stereo-chemistry

DockStream: A Docking Wrapper to Enhance De Novo Molecular Design

Guo et al ChemRxiv

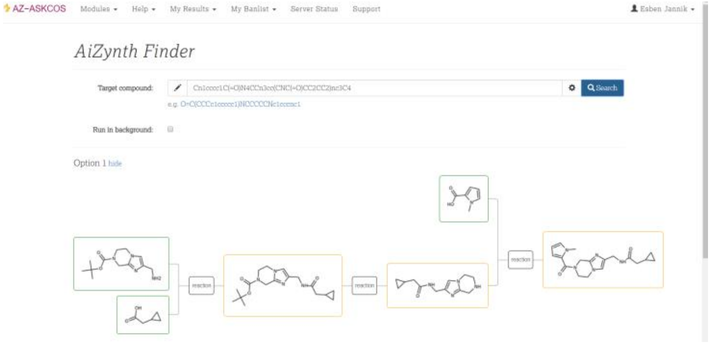
Our knowledgebase: ReactionConnect



AiZynthFinder

https://github.com/MolecularAI/aizynthfinder

Web-GUI based on MIT MLDPS consortium tools



AiZynth Finder

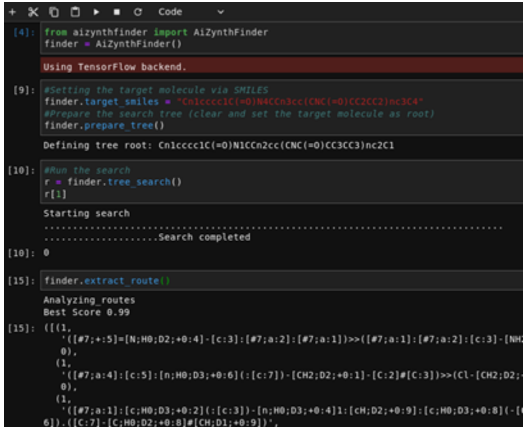
Target compound: Cn1cccc1C(=O)N4CCn3cc(CNC(=O)CC2CC2)nc3C4

e.g. O=C(CCCc1ccccc1)NCCCCCNc1ccnc1

Run in background: □

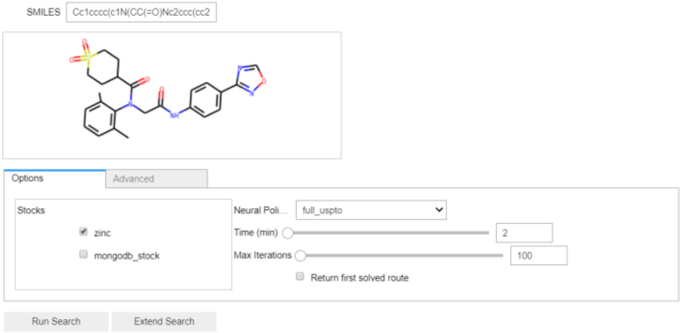
Option 1 hide

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

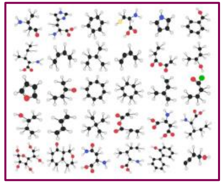
Genheden et al. AiZynthFinder: A Fast Robust and Flexible Open-Source Software for Retrosynthetic Planning. **2020**. https://doi.org/10.26434/chemrxiv.12465371.v1 .

So which lessons have we learned?

- Needs in discovery chemistry and process chemistry are very different
- Extracting and integrating reaction data is hard work
- Challenging to assess the utility of different tools
- Advanced building block look-up
- Impact on synthetic routes have mainly been from specialized tools like the Ringbreaker

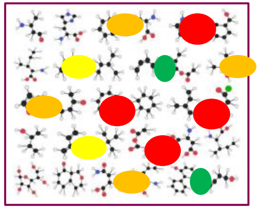
Artificial Intelligence Guided Drug Design Platform (REINVENT)

Molecular Generation



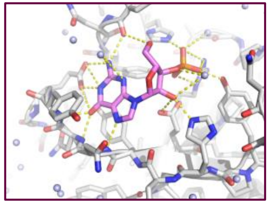
- Recurrent neural networks
- Reinforcement learning
- Scaffold penalty
- Library design

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

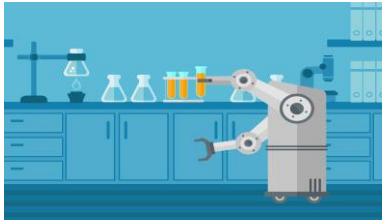
AI+ vision for drug design

AI can't transform drug design alone

High-Throughput Data Generation

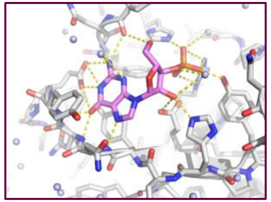
- The most important determinant of the usefulness of a model is the size and quality of the data set for training
- High-Throughput Experimentation for generating chemical reaction data
- Cell-paint & transcriptomics to create molecular signatures
- DNA Encoded Library models to score molecules

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

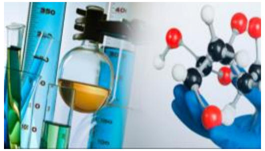


- More accurate models for difficult to predict properties can be created through combining physics and AI
- Relative binding free energy perturbation binding affinity in molecular optimization
- Absolute binding free energy perturbation to estimate binding energies in hit finding and for scaffold hopping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

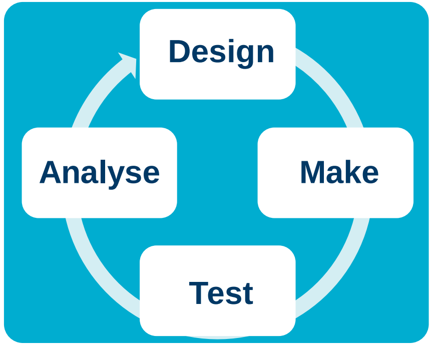
Integration of AI and automation

AI

What to make next ?



How to make it ?



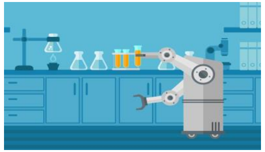
Design

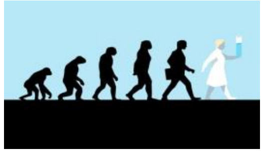
Make

Test

Analyse

How to Automate Synthesis ?

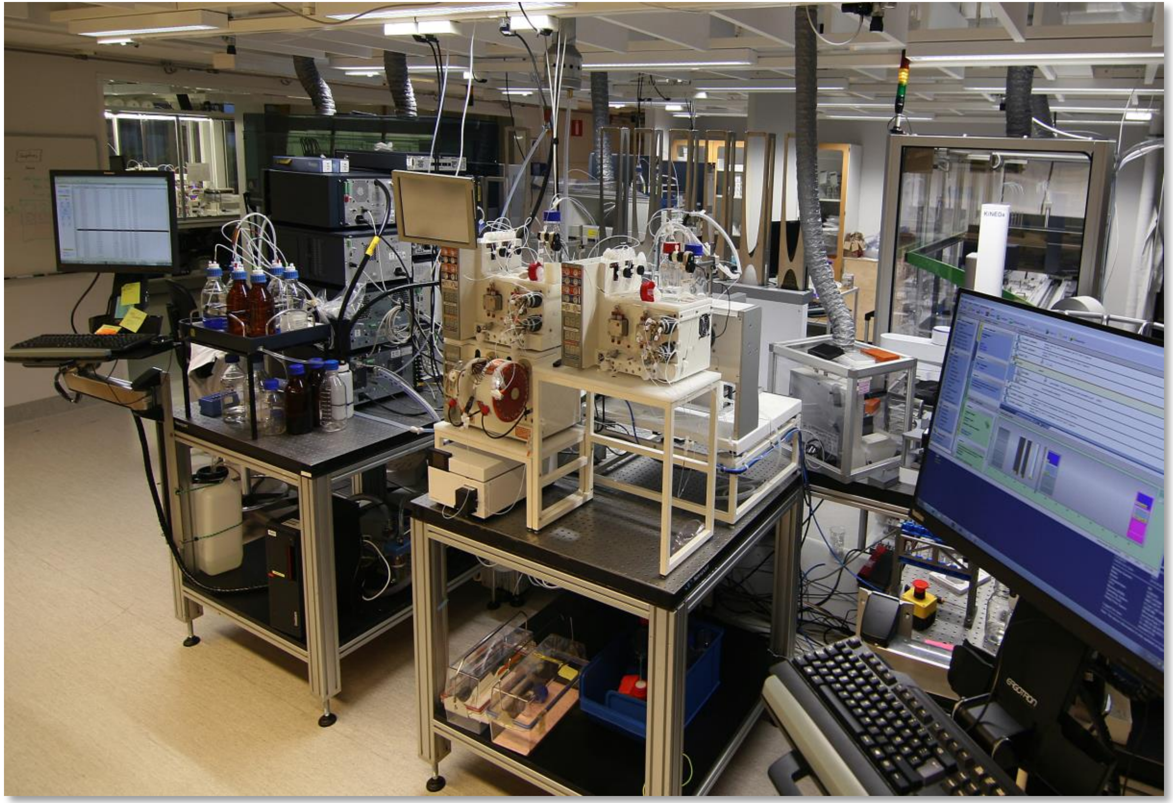


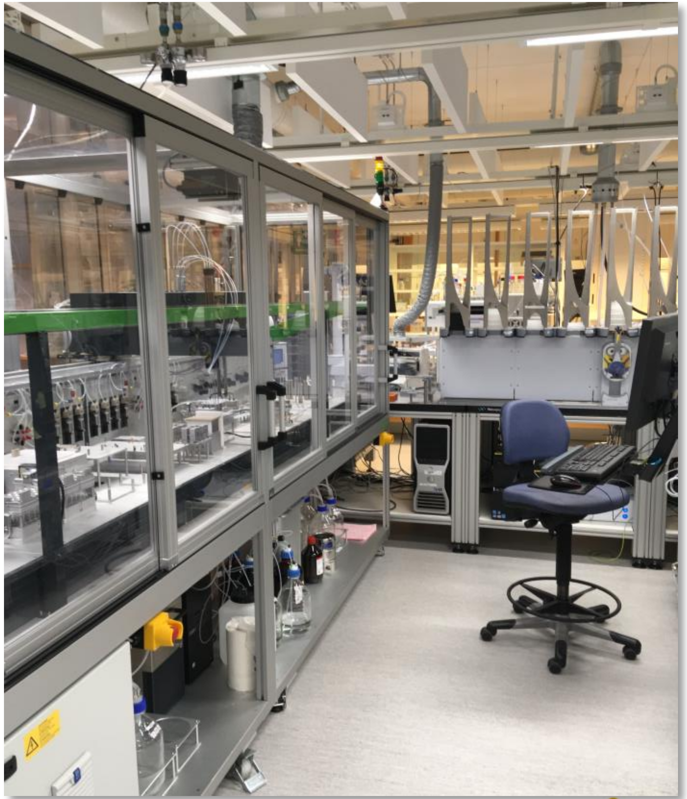


How to increase synthesis throughput

iLAB

Automated Synthesis Platform @AstraZeneca





Keep a balanced view!

- Progress have been made!
  - Molecular Generation
  - Synthesis prediction
  - AlhpaFold2

- Progress will continue!
  - Better hardware
  - High-throughput data generation
  - Novel innovative algorithms
  - Better deep learning based force-fields for affinity estimation etc

- Some predictions will be difficult!
  - In vivo properties from only molecular structures
  - Expensive data generation, Noisy data, Non-smooth surface to learn

Why am I an optimist

- Increased computational power
  - Never underestimate an exponential law
- Increased automation provides large and consistent datasets
  - HTE, DEL, Cell Paint, RNA-seq
- Advances in computational algorithms
  - Merging of physics-based modelling & ML
  - GPT-3, Codex

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

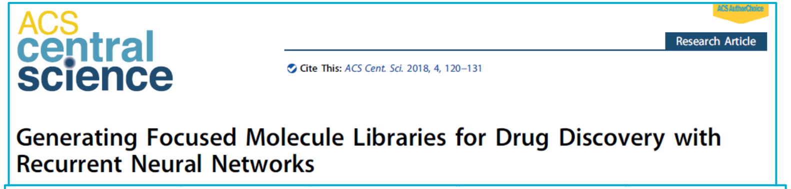
- Physics based modelling
  - Binding affinity and solubility predictions are major bottlenecks

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

- Flexibility of chemistry automation

- Educational, cultural & logistical challenges besides scientific

Science Molecular AI @AZ



ACS central science

ACS AuthorChoice

Research Article

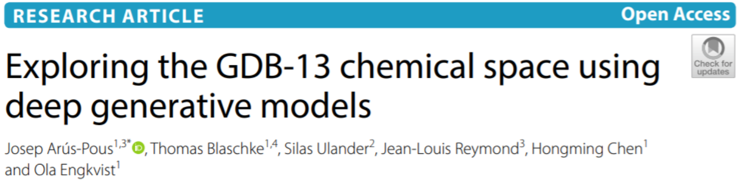
Cite This: *ACS Cent. Sci.* 2018, 4, 120−131

Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks

RESEARCH

Molecular De-Novo Design through Deep Reinforcement Learning

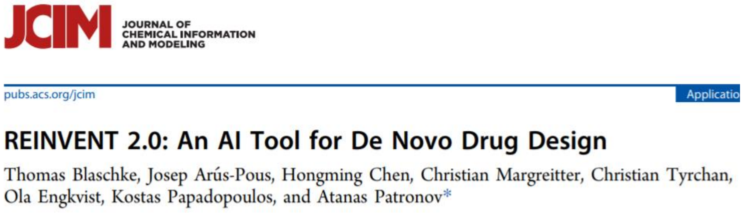
Marcus Olivecrona*, Thomas Blaschke†, Ola Engkvist† and Hongming Chen†



RESEARCH ARTICLE Open Access

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]

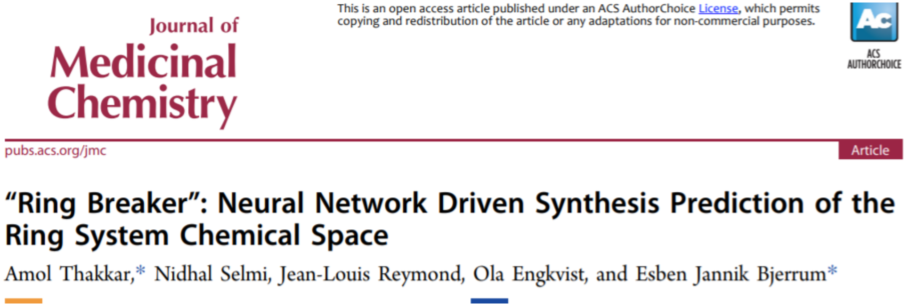


JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

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]

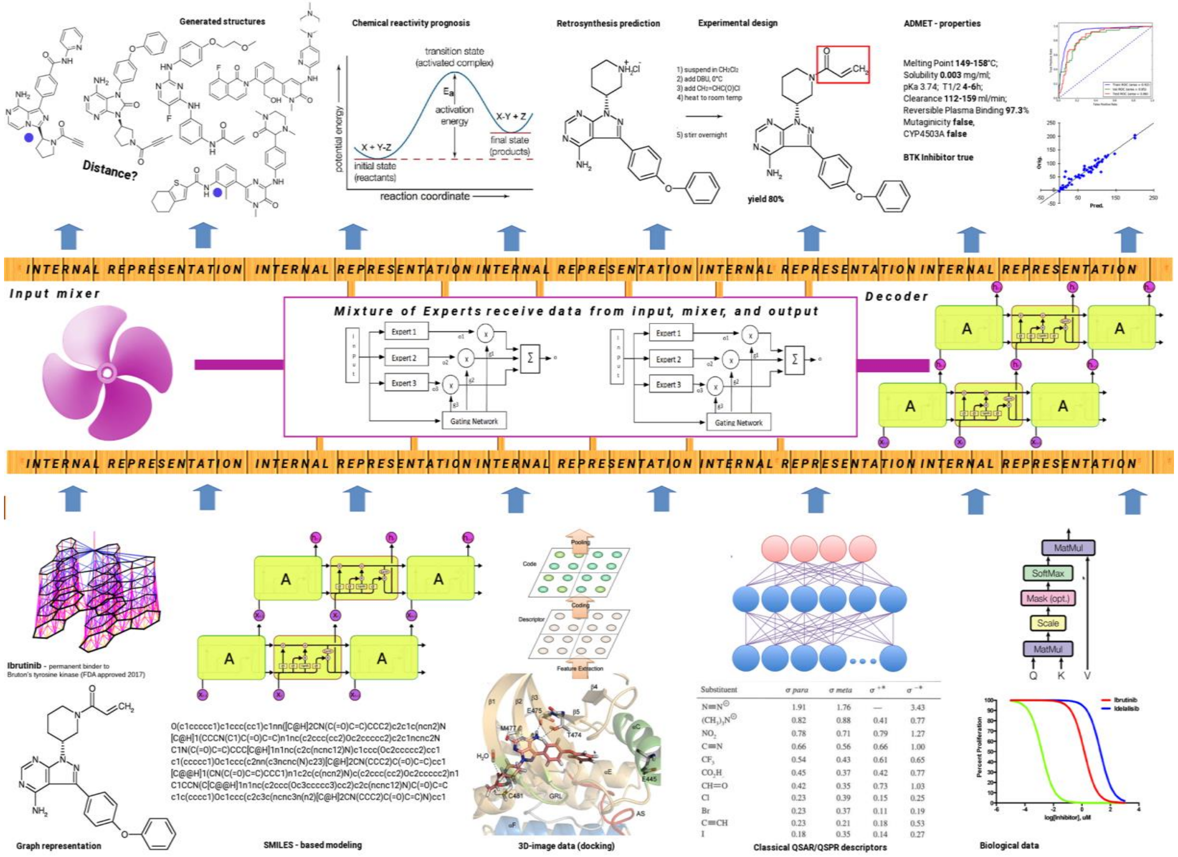
SOFTWARE Open Access

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

Advanced Machine Learning for Innovative Drug Discovery https://ai-dd.eu



1) suspend in CH_2Cl_2
2) add DBU, 0°C
3) add CH_2=CHC(O)Cl
4) heat to room temp
5) stirr overnight

Melting Point **149-158**°C;
Solubility **0.003** mg/ml;
pKa 3.74; T1/2 **4-6**h;
Clearance **112-159** ml/min;
Reversible Plasma Binding **97.3**%
Mutaginicity **false**,
CYP4503A **false**

BTK Inhibitor true

```
O(c1ccccc1)c1ccc(cc1)c1nn([C@H]2CN(C(=O)C=C)CCC2)c2c1c(ncn2)N
[C@H]1(CCCN(C1)C(=O)C=C)n1nc(c2ccc(cc2)Oc2ccccc2)c2c1ncnc2N
C1N(C(=O)C=C)CCC[C@H]1n1nc(c2c(ncnc12)N)c1ccc(Oc2ccccc2)cc1
c1(ccccc1)Oc1ccc(c2nn(c3ncnc(N)c23)[C@H]2CN(CCC2)C(=O)C=C)cc1
[C@@H]1(CN(C(=O)C=C)CCC1)n1c2c(c(ncn2)N)c(c2ccc(cc2)Oc2ccccc2)n1
C1CCN(C[C@@H]1n1nc(c2ccc(Oc3ccccc3)cc2)c2c(ncnc12)N)C(=O)C=C
c1c(cccc1)Oc1ccc(c2c3c(ncnc3n(n2)[C@H]2CN(CCC2)C(=O)C=C)N)cc1
```

Substituent	σ para	σ meta	σ ⁺*	σ ⁻*
N≡N$^{\oplus}$	1.91	1.76	—	3.43
$(CH_3)_3N^{\oplus}$	0.82	0.88	0.41	0.77
NO_2	0.78	0.71	0.79	1.27
C≡N	0.66	0.56	0.66	1.00
CF_3	0.54	0.43	0.61	0.65
CO_2H	0.45	0.37	0.42	0.77
CH=O	0.42	0.35	0.73	1.03
Cl	0.23	0.39	0.15	0.25
Br	0.23	0.37	0.11	0.19
C≡CH	0.23	0.21	0.18	0.53
I	0.18	0.35	0.14	0.27

- ✓ On-line courses
- ✓ Schools in AI & ML

Public lectures by zoom

See announcements at:

https://ai-dd.eu/news

Or follow:

https://twitter.com/**AiddOne**