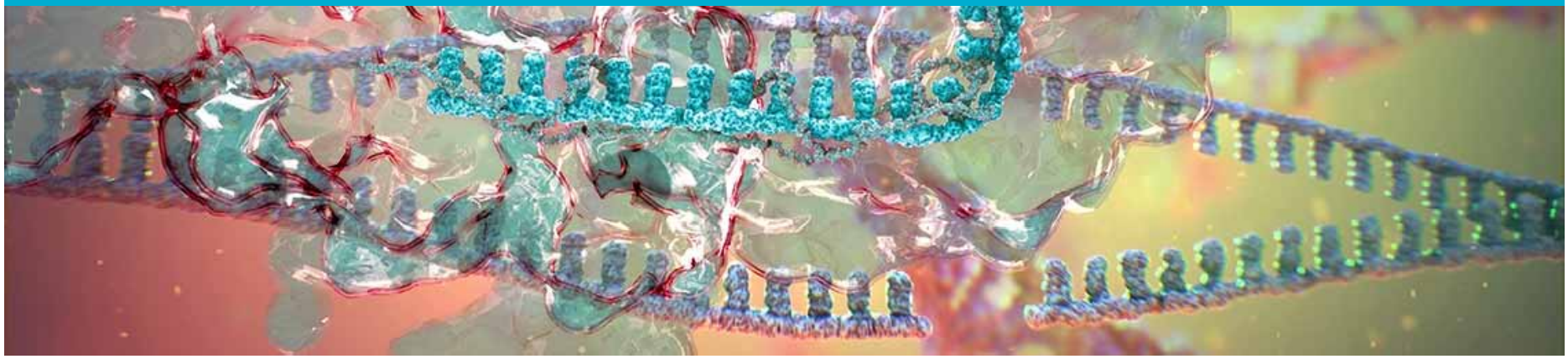
AstraZeneca
IMED Biotech Unit

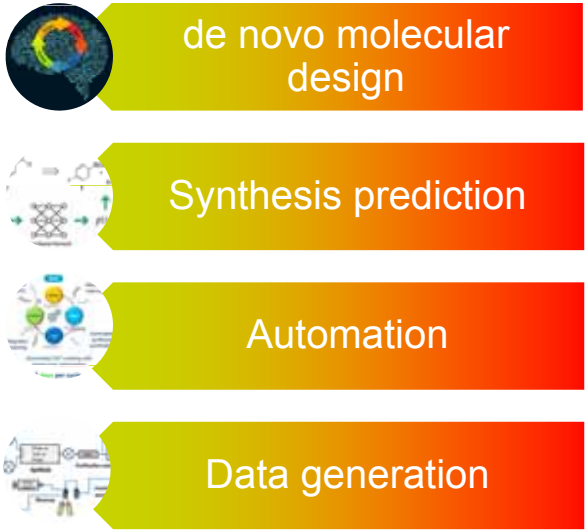
Molecular *de novo* Design through Deep Learning

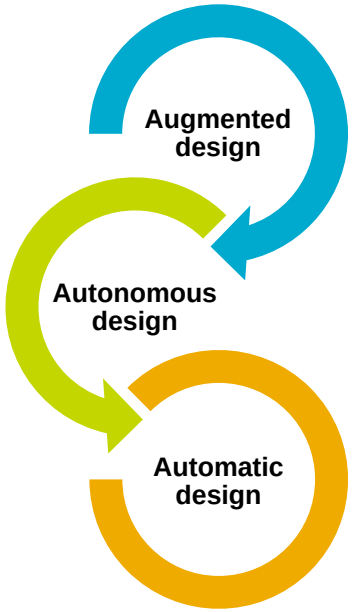
Ola Engkvist, Hit Discovery, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden

RSC Artificial Intelligence in Chemistry June 15 2018

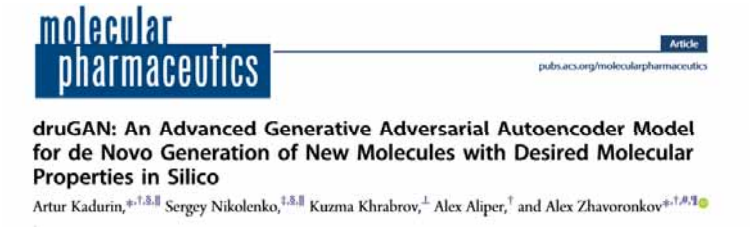


What is different?





De novo molecular generation with deep learning has developed very rapidly

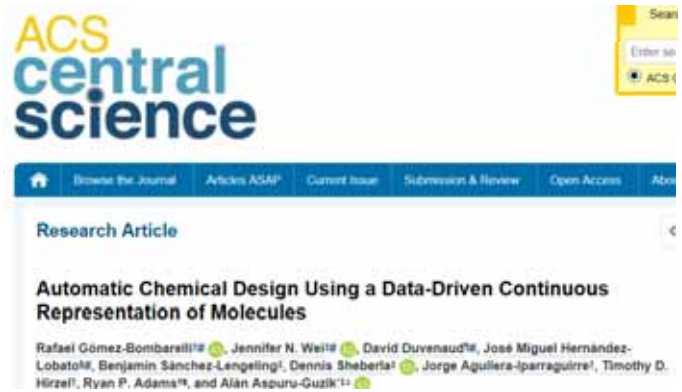

molecular pharmaceutics

Article

pubs.acs.org/molecularpharmaceutics

druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico

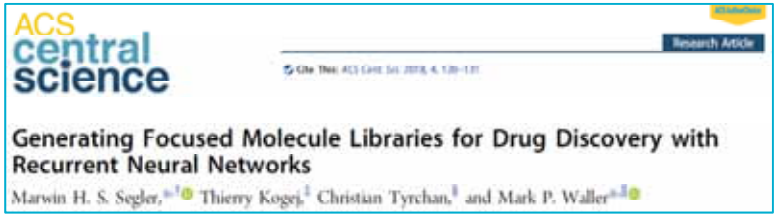
Artur Kadurin, Sergey Nikolenko, Kuzma Khrabrov, Alex Aliper, and Alex Zhavoronkov


ACS central science

Research Article

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik


ACS central science

Research Article

Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks

Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller


RESEARCH

Molecular De-Novo Design through Deep Reinforcement Learning

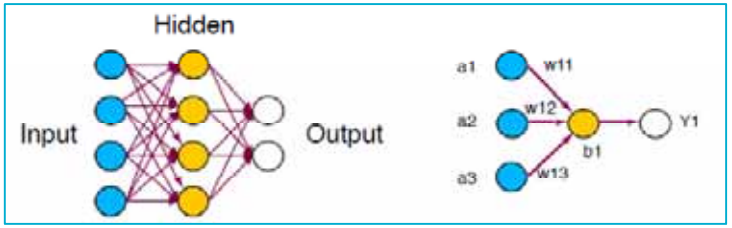
Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen

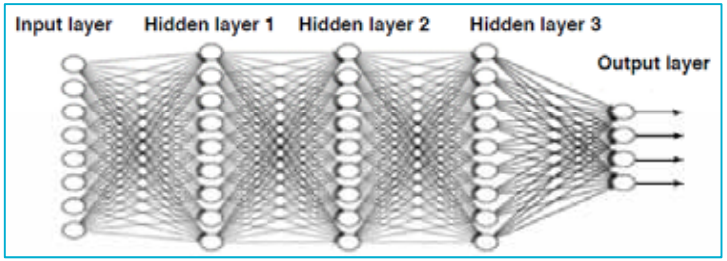

The rise of deep learning in drug discovery

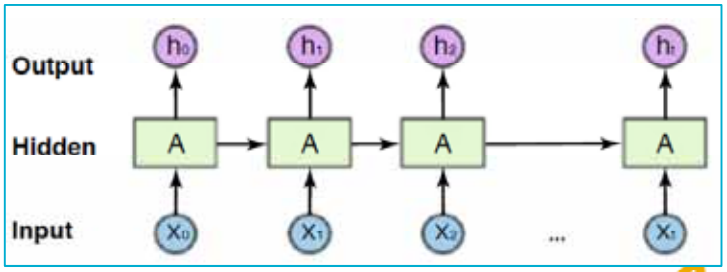
Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona and Thomas Blaschke

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



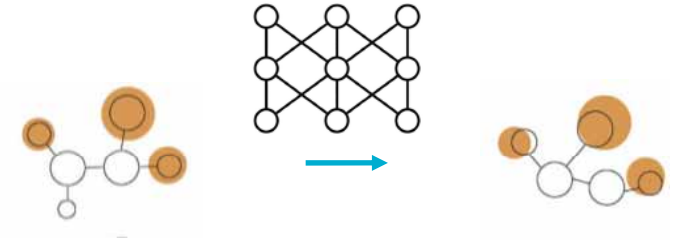




Why? Generation of Novel 10^{60} Chemical Space

10^{10}-10^{12}

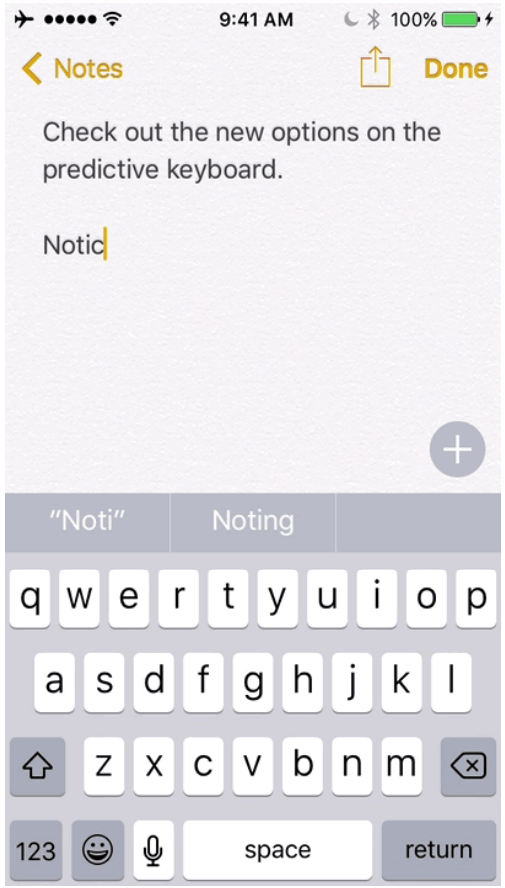
10^{60}



Where´s the impact?

- Use for de novo Molecular Design
  - Scaffold Hopping
  - Novelty
  - Virtual Screening
  - Library Design

Recurrent Neural Network & Natural language generation

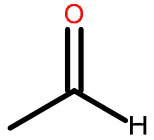


Natural language generation and molecular structure generation

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

The → grass → is → ?

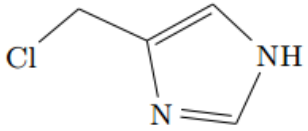
C → C → = → ?



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

Recurrent Neural Networks

- Keeps track of context using a *cell*
- Is often trained using *Maximum Likelihood Estimation*
- Maximize probability of next token, e.g. “CCO”

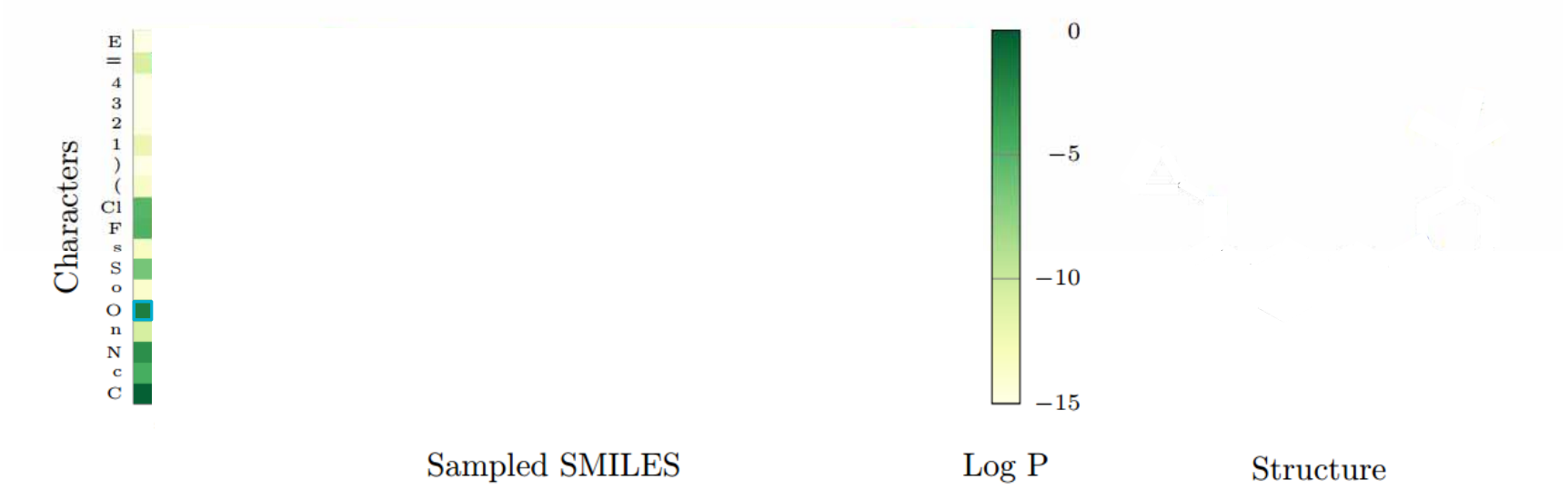
Recurrent Neural Networks

- When trained, can be used to generate new sequences
- Sample from probability distribution at every step
- Use sampled character as next input

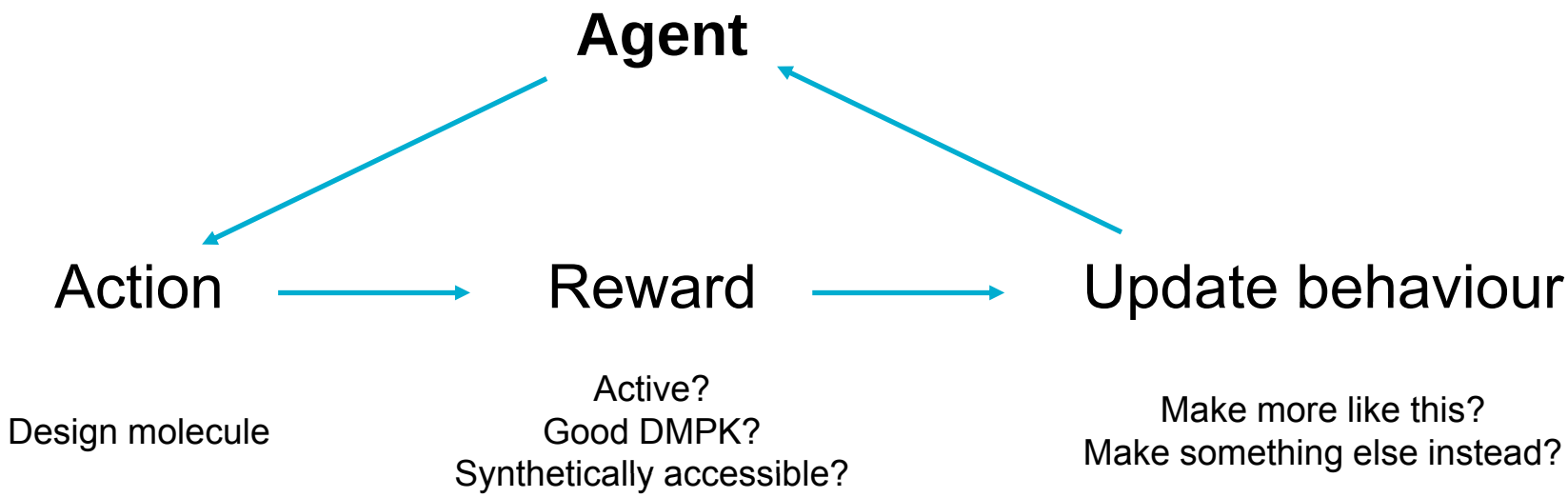
The prior RNN

- Constrain structures to 10 to 50 heavy atoms
- Elements H, B, C, N, O, F, P, S, Cl, Br, I
- 1.5 million SMILES from ChEMBL
- Canonicalized using RDKit

The generative process

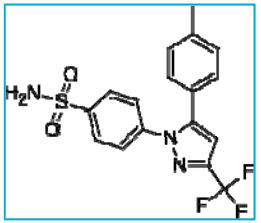


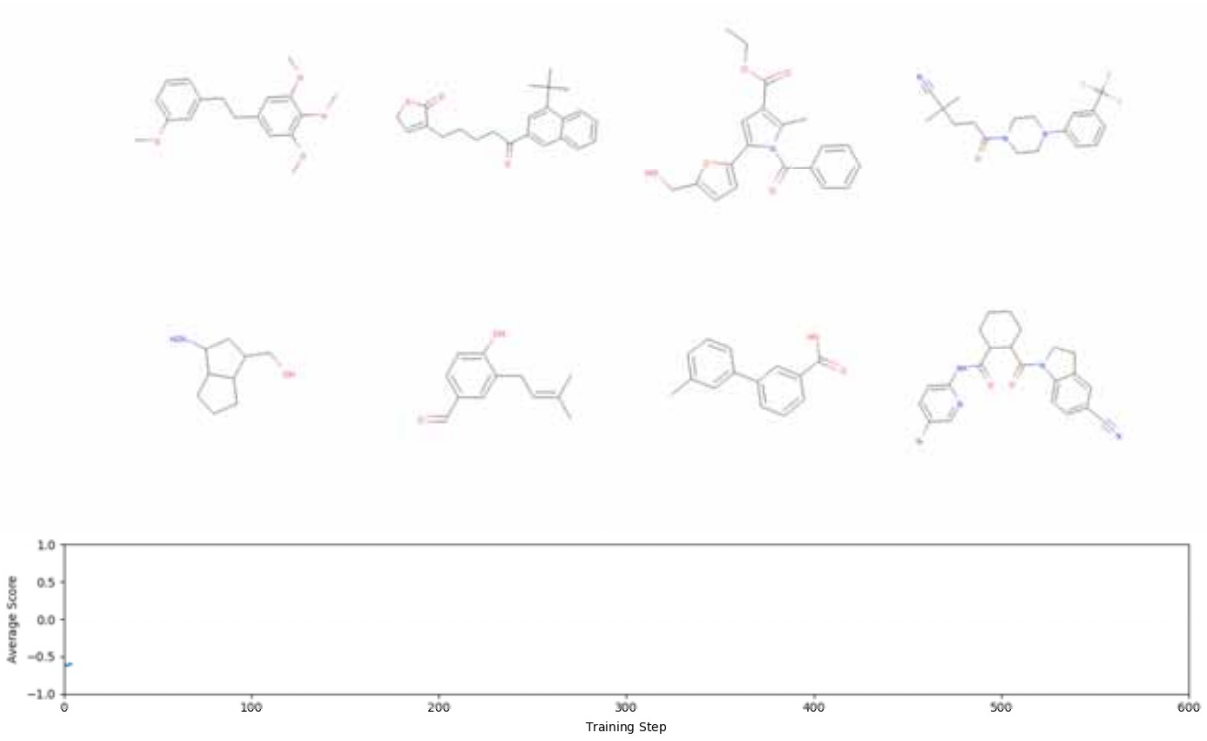
Reinforcement learning



Learning from doing

AI live: Create Structures Similar to Celecoxib



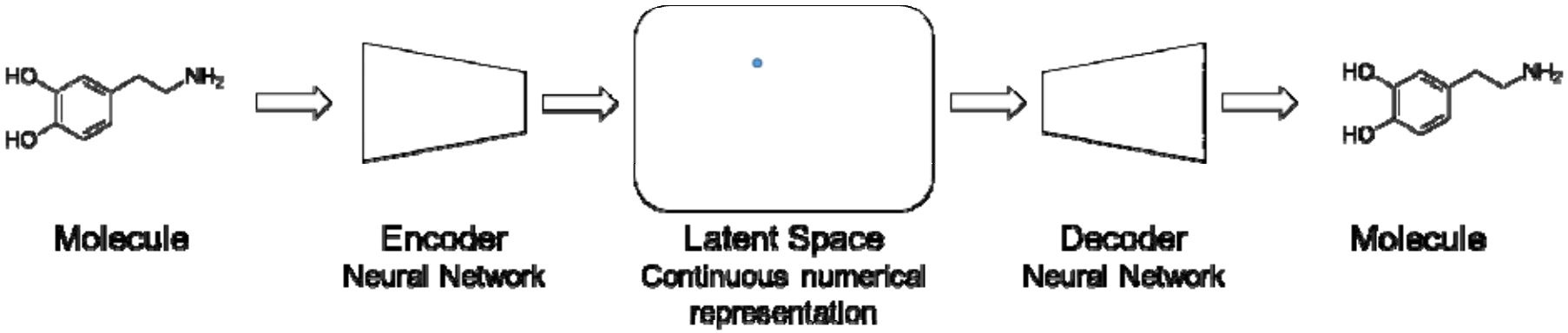


- **Key Message**
  - RNN generates structures similar to Celecoxib
  - Rapid sampling!
  - Average score describes how many learning steps are required to reach similar compounds

Autoencoder network

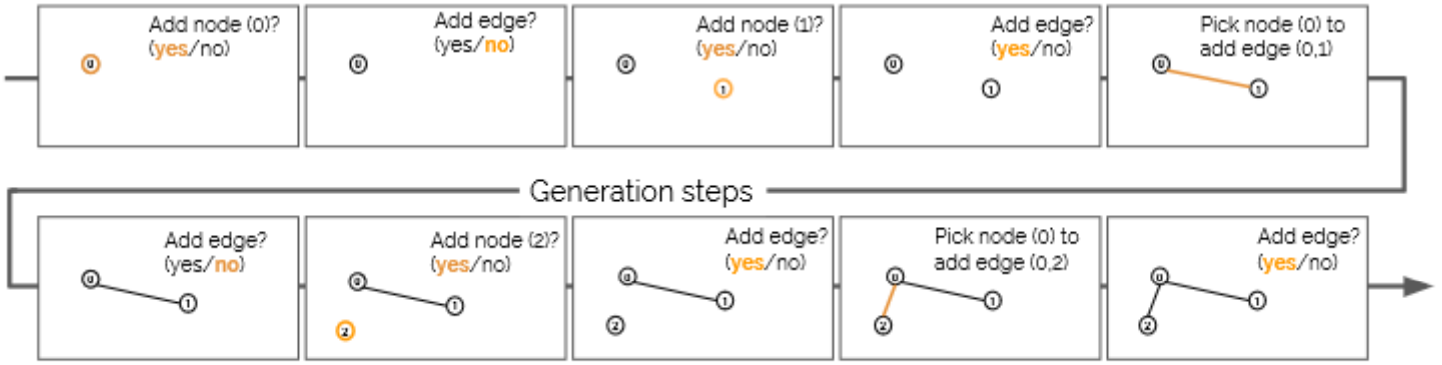
- The autoencoder consists of two separate neuronal networks

  1) Encoder: maps SMILES strings into latent space
  2) Decoder: generates SMILES strings given a point in latent space

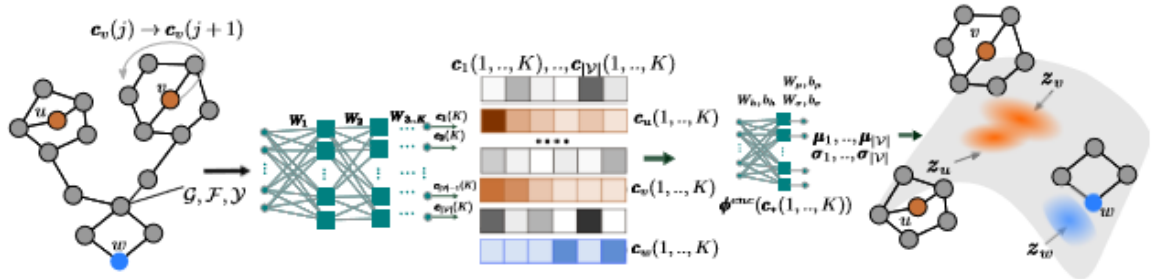


Current research is increasingly focused on graph based methods

Probabilistic graph generation Deepmind arXiv:1803.03324



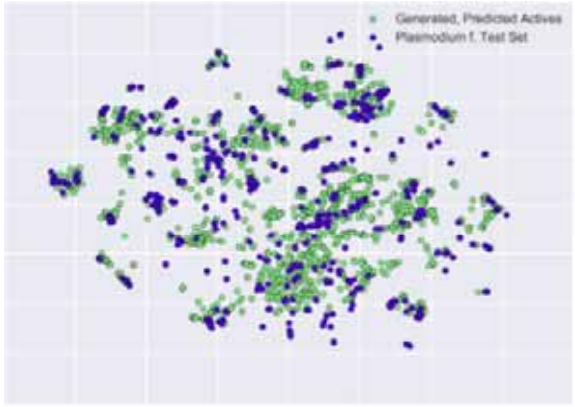
Graph Variational Autoencoder arXiv:1802.05283



Some misconceptions about *de novo* RNN generated molecules

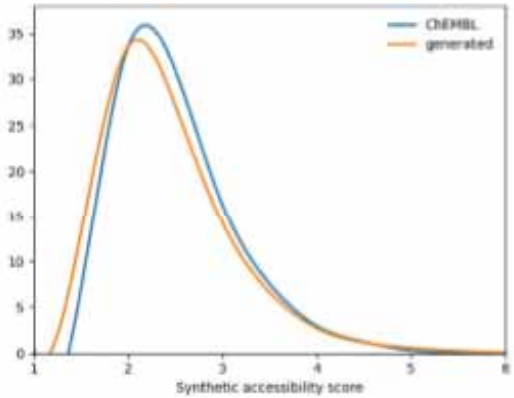
"The molecules are not diverse"
"The molecules are not synthetic feasible"

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



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

Diversity



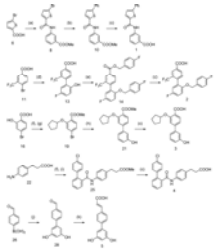
Ertl et al arXiv:1712.07449

Synthetic feasibility

How to evaluate algorithms for deep learning generated molecules

% generated molecules are syntactically correct smiles (95%)
% generated molecules that are novel and not in the prior (90%)
% generated molecules that are unique (90%)
“Druglikeness” and diversity of generated molecules

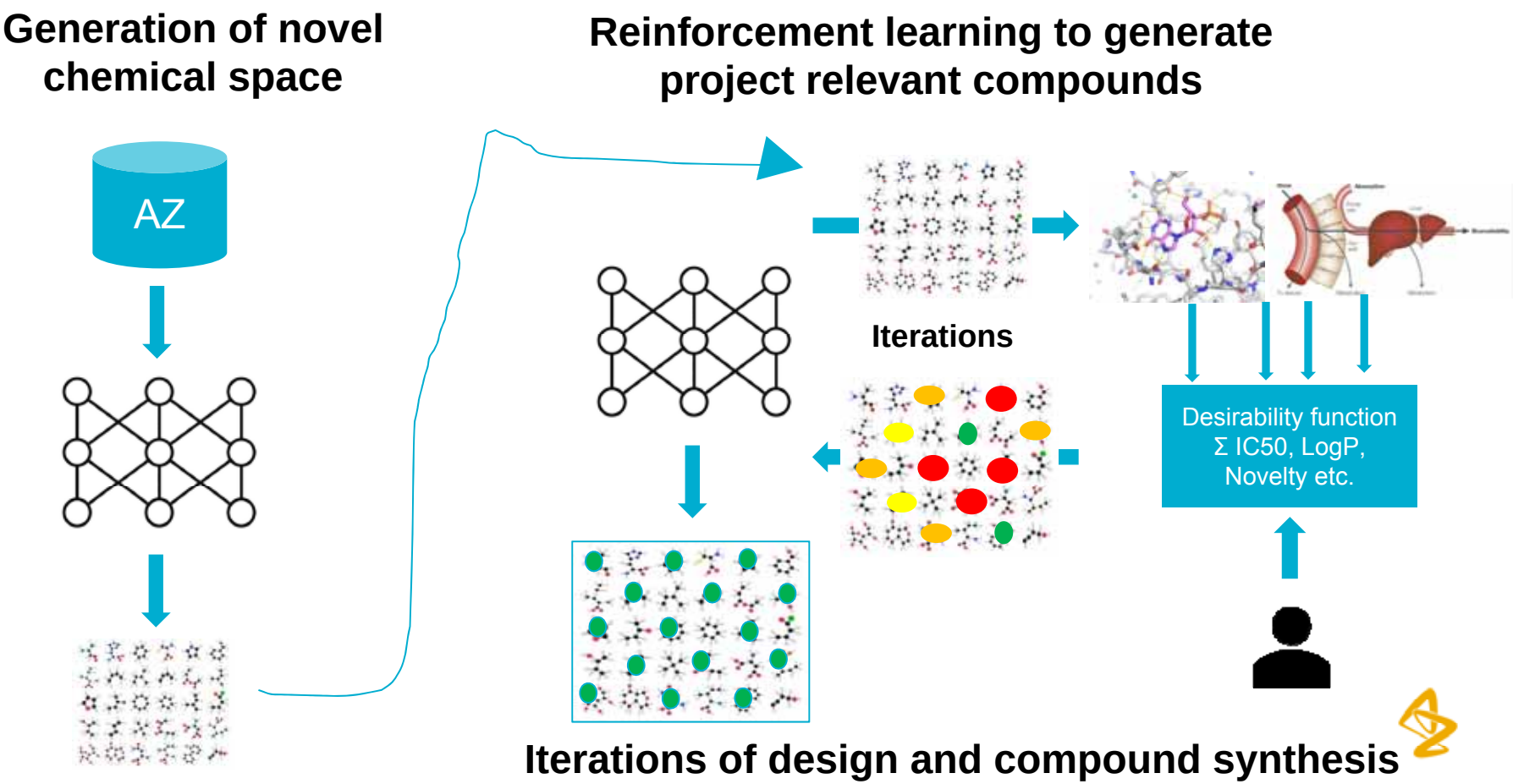
Deviation in overall distribution in properties between prior and generated molecules (Klambauer, arXiv:1803.09518)



Identify prospectively active molecules
(Schneider et al Mol. Inf. 2017, 37, 1700153)

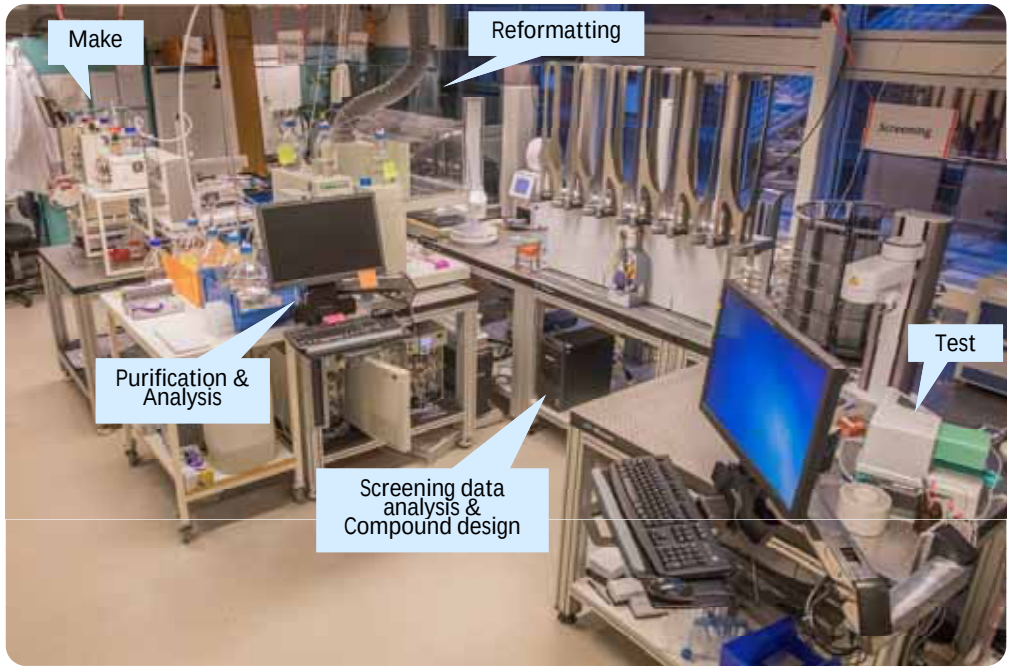
OpenAI Gym (Segler et al ICLR 2018)

Components of the AZ de novo Molecular Augmented Design Platform



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



The Future Journey together with chemistry automation

2018

Augmented Design

- AI as idea generator
- AI generates and scores molecules
- All decisions by chemist based on AI
- Synthesis route design by chemist

2020

Autonomous Design

- AI as designer
- Chemist defines, monitors and updates the desirability function

2025

Automated Design

- AI as designer
- AI monitors and updates desirability function when needed


Formalize Chemical Intuition


Learn from diverse data


Extend to new modalities


Explore New Deep Learning architectures

"Tools on tap"

Conclusions

- Molecular de novo generation with deep learning methods is one of the most exiting developments in cheminformatics during the last years
- Preprints and open source have facilitated very quick progress
- Properties of the generated molecules follows the underlying prior including diversity and synthetic accessibility
- On our way to solving how to search the chemical space
- Scoring generated molecules will be a focus area

Acknowledgements

Discovery Sciences CompChem ML/AI
Thierry Kogej
Hongming Chen
Noe Sturm (Postdoc)
Philipp Buerger (Postdoc)
Thomas Blaschke (PhD student)
Josep Arus Pous (PhD student)
Michael Withnall (PhD student)
Oliver Laufkötter (PhD student)
Laurent David (PhD student)
Marcus Olivecrona (AZ GradProgram)
Dhanushka Weerakoon (AZ GradProgram)

Discovery Sciences
Garry Pairaudeau
Clive Green
Lars Carlsson

Respiratory disease area
Christian Tyrchan
Werngard Czechtizky

Academic Collaborators
Marwin Segler (Munster)
Juergen Bajorath (Bonn)
Jean-Louis Reymond (Bern)
Andreas Bender (Cambridge)
Sepp Hochreiter (Linz)
Gunther Klambauer (Linz)
Sami Kaski (Helsinki)

Confidentiality Notice